CENTRUM FÜR INFORMATION UND SPRACHE (CIS)
der Ludwig-Maximilians-Universität München

Fachbereich Sprach- und Literaturwissenschaften

Term-driven E-Commerce

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
an der Ludwig Maximilians-Universität München
im Studiengang Computerlinguistik

eingereicht von:	Gerhard Rolletschek
eingereicht am:	10. November 2006
Betreuer:	Herr Prof. Dr. Franz Guenthner
Zweitgutachter:	Herr Prof. Dr. Klaus Schulz

Datum der mündlichen Prüfung:

5.02.2007

Contents

1. Preface

During a large span of the time I spent on researching, I received a scholarship by Espotting Media Ltd. I would especially like to thank Dylan Fuller and Heike Röttgers of Espotting Media / Miva, not only for feeding and watering me, but also for sharing with me their insights, ideas and constructive suggestions. The extensive discussions with devoted professionals, namely with Ross Leher and Marc Brombert (WAND Inc.), Christoph Röck (Pangora) and Frank Wenz (TVG Verlag) helped to give me most valuable insight into practical issues of the topic.

The work is greatly indebted to my supervisor Professor Franz Guenthner without whose guidance, encouragement, overview and detailed knowledge this thesis would not have been possible. Warm thanks also go to my colleagues at the CIS, especially Yeong Su-Lee and Michaela Geierhos.

2. Introduction

This thesis delves into the textual dimension of E-Commerce, in particular into its vocabulary dimension.

There are good arguments why such a direction of work is important — not only for commercial applications as its title suggests, but also as a test-bed for term studies and thereby for language processing methods in general. While the practical importance of the topic chosen will probably be almost universally accepted, the lexicalist approach followed here might lack grace and elegance to some. The suspicious absence of a rich formula apparatus and dot graphs is the visual consequence of this approach that is decidedly differently put forth than today's prevailing statistical approaches.

This thesis aims (high) at demonstrating the necessity of a slow but sturdy progress involving many hours of manual work. While there might be magic out there, searching for it without knowing before what to look for does not bring us closer to a solution.

Apart from academic research, there is no lack of enterprizes, software and white papers all dedicated to enhancing E-Commerce information applications[1] However, what is still a desiderate is an integrative approach combining all available terminology information and corpora data as well as exploiting algorithms for extracting information out of these accumulated data. Sketches of how such an integrative approach could look like, what theories and methods it could incorporate and in which fields of application it could be beneficiary are presented in this thesis under the title of Term-driven E-Commerce.

The work presented here is based on the conviction that there is no point in ignoring knowledge already available and that when tackling new problems an iterative approach should be considered in first place, before building completed and fully elaborated frameworks which make future amendments — by more humble minds — difficult. In short, this paper follows a lexical stance and prefers empirically based descriptions over the seek of the "one" algorithm that does the trick to solve everything at once. This does not dispense this work from stating its theoretical and methodical foundations, but puts formalisms in the ancillary place from which they seem to have struggled free in some branches of research.

[1] Good starting points the three major web search engines and their labs featuring projects and publications which cover a broad range of E-Commerce information aspects. Visit Google's `http://labs.google.com/papers.html [Nov. 1, 2006]`, Yahoo's `http://research.yahoo.com/publication.shtml [Nov. 1, 2006]` and Microsoft's Search, Retrieval and Knowledge Management research `http://research.microsoft.com/research/detail.aspx?id=7 [Nov. 1, 2006]`, especially its adCenter Lab at `http://adlab.microsoft.com [Nov. 1, 2006]`. In addition, the annual WWW conferences (`www2007.org [Nov. 1, 2006]`, the homepage of the upcoming conference) often featured papers dealing with E-Commerce related topics.

Following a broad consensus, the most important questions for a lexicographer resides in deciding what is to be gathered, what to be excluded and how this information is arranged both in terms of microstructure and macrostructure[2]. All these decisions should not only result from theoretical principles, but justify themselves in practice, in terms of coverage and accuracy.

A word more is needed on the choice of topic. Arguably, there are words with a sweeter sound to it than E-Commerce which first failed high-flung expectations and then revigorated for many in the pestering form of popup ads and bulk mail[3] However, it may be worth to note that if commercially relevant intentions are behind a user's actions (and they often are, as can be empirically observed), the question is not so much between neutral vs. commercial results but rather that between qualitative good vs. bad results. As Aaron Wall put it in his SEO-blog, "paid search listings are often more relevant than spam-filled algorithmic search results"[4]. There are many search terms for which this observation certainly is true[5]. Although from the viewpoint of media policy it is necessary to separate transparently content from advertisement, this does not prevent a clearly demarcated and highly relevant advertisement (though this term is both in its connotations as well as in its etymology misleading) from being just the optimal answer in some cases. One may well ask why keyword-driven online advertisements should do better in this respect than conventional advertising activities, such as static banner ads. The answer to this is twofold: It is based both in the high degree of interaction and in the exact metrics of the ads' impact. Given that advertisement is costly and has to pay off by attracting the target audience, in usual market situations irrelevant ads are not likely to sustain long. This contrasts with other online marketing instruments such as unsolicited bulk mails with little extra costs for a higher number of expositions and therefore a much lower threshold for an acceptable response rate.

The first part of this paper develops a model of term-driven E-Commerce as the methodological answer to the challenge of information driven E-Commerce. This model evolves by viewing E-Commerce from the perspective of virtual market places. This virtual market place is then characterized by its agents and stakeholder and by the interactions taking place between them. Starting from the general picture of total numbers, key players and major trends, the focus then lies on the data accumulation related to each agent. The notion of term-driven E-Commerce also sheds light on new data resources. Prominent among such resources is the inventory of keywords for

[2] Good introductions into the practical issues of lexicography are provided by Dictionaries : the art and craft of lexicography / Sidney I. Landau. New York :. Scribner, 1984, Howard Jackson Lexicography and Teaching and Researching Lexicography by Reinhard Hartmann.

[3] For an inspiring model of technology maturing, see Jackie Fenn's Hype Cycles, at `http://www.gartner.com/pages/story.php.id.8795.s.8.jsp [Nov. 1, 2006].`

[4] His blog can be found at `http://www.seobook.com [Nov. 1, 2006]`

[5] These terms, however, change heavily over short time and there is little point in giving examples that worked only few days in the past. Current battlefields of E-commerce marketing competitions are reported in Part D, Adspaces. See also `http://www.google-watch.org/woodard.html [Nov. 1, 2006]`

which advertisers bid in order to place their ads. Further resources that have especial value for term-driven E-Commerce include query logs from Yellow Pages and shopping sites (see Part A, Term Spaces). In addition to this, large repositories of accumulated Yellow Pages branch headings and product categories are to be examined.

The methods for analyzing E-commerce data and key findings are presented in the second part. Three algorithmic issues will recur over the whole practical part, namely co-occurrence metrics, minimal coverings of sets and left-right bootstrapping. They are used in solving term matching, variation and enrichment challenges. Part B is organized along the traditional modules of linguistics, i.e. into a section on orthographic, on syntactic-morphologic and finally semantic variation. The notion of Sense-Morphemes as building blocks of E-Commerce vocabulary is continuously deepened throughout this chapter.

The last two parts of this thesis are concerned with applications of term handling processes. Part C deals with query recognition and processing. Here, the notion of heads vs. containers and of query types is tested against real-life data of various origins. Results will be both presented in the form of in-depth studies as well as by stating overall figures. Part D presents three case-studies that are pivotal to TE-Commerce: Using and adapting ontologies for TE-Commerce, the world of paid keywords and issues of enhancing Yellow Pages sites through vocabulary enrichment.

Some writing conventions: Trademarks are the property of their respective owners, although they are not specifically marked throughout this thesis. As this paper is on real-world E-Commerce, it abounds in trademarked terms. Some fundamental terms are written with capital initial letters, such as *Web* or *E-Commerce*. This should not imply that they are used in any non-standard or non-intuitive way.

Singulars as names of groups – such as *the user* when referring to all users – are treated as female forms.

Part A.

The framework and methodology of Term-driven E-Commerce

3. Term-driven E-Commerce

Introducing the concept of Term-driven E-Commerce requires the following components which will be discussed step by step below:

- Model a framework for agents and their interactions in the field of TE-Commerce
- Highlight the role of term-related functional enhancements
- Corroborate the hypothesis that E-commerce is currently in a stage of TE-commerce, i.e. not yet a fully textual-driven digital commerce, but based on terms and term relations
- Lay out the range and importance of term handling routines: term recognition, cleaning, normalization and enrichment

The term *E-Commerce* and compounds derived from it are going to appear incessantly in the following chapters. Unfortunately, there are almost as many perceptions of what these terms should convey as there are people using them (a statement which is supposed to indicate both their ubiquity and their vagueness). The sobering effect that went along with the end of the dot com boom's high flinging hopes took away the magical qualities of E-Commerce (making money out of barely anything). Over the last years, though, a digital infrastructure became imperative for broad sectors of businesses in all highly developed countries. In addition to this, companies such as Amazon, eBay or Google certainly make money out of something, and are so inherently entwined with the Web that one can hardly imagine them without it. What should be subsumed under the term E-Commerce and what excluded from it?

In a broad sense, E-Commerce could be defined as any method or means of enabling and enhancing commercial transactions by electronic communication technology. According to this broad sense, E-Commerce and E-Business would be interchangeable. As a consequence, many people also subsume fields such as electronic funds transfer, supply chain management, online transaction processing, data interchange, procurement, inventory management systems, customer relationship management and other processes under the rubric of E-Commerce. Their is a valid argument for grouping these fields under one rubric: Although all of these fields have evolved into specific branches with their own respective logic, they all center around the paradigm of transactions. However, if one factors out all auxiliary business processes out of E-Commerce, what remains is a stricter sense of E-Commerce. This narrower sense will be adhered to in what follows. According to this, E-Commerce consists of distributing, buying, selling, monitoring, promoting and marketing of products or services over communication systems such the Internet. The transaction-based model and its participants will be discussed in the following chapter[1].

From this point of view, an understanding of E-Commerce can be developed centered on terms of information and access. Three basic axiomatic assumptions on the relationship between E-Commerce and the Web illuminate why this view can be said to reflect the core of E-Commerce:

1. Commerce will soon not be separable from E-Commerce; in fact, the Web is the Yellow Pages of tomorrow

2. E-Commerce is not separable from textual representations, especially in the forms of terms. Term handling methods can be fruitfully applied to E-Commerce. By the same token, TE-Commerce might serve as a test-bed for a wide variety of linguistic issues.

3. For every E-Commerce application, it is feasible and necessary to capture relevant terms and their relationships.

Each claim deserves its own treatment. Apart from foreshadowing the methodical change in Part A from economical studies towards applied linguistics, these claims are also supposed to legitimate the approach in illustrating the importance of the topics they touch and, moreover, the conclusive logic relationships between them.

3.1. The diminishing discriminative power of the "E" in "E-Commerce"

Just as today no-one adds *Electric* to *lighting*, because all our home lighting except dinner candles are based on electricity, soon the notion of "E-Commerce" is going to be subsumed under a general concept of Commerce or business transactions. One may argue against the validity of this statement by referring to the billion people living still completely off-line, both in developed and in lesser developed countries around the globe. The oft-cited *digital divide* should rather be seen as a challenge than being accepted as a fate for off-liners. This holds true for both the global and the social digital divide. The heterogeneous global pace of Net penetration and usage, perhaps best illustrated by world maps featuring main traffic routing lines, IP or router density[2] is still a fact, yet there are indications that some less developed countries have accelerated gain rates[3]. For now, it suffices to state that E-Commerce is not inherently limited to desktop PCs or laptops, nor to any costly device that also functions

[1] A similar model of E-Commerce is developed in [Merz 2002]. General overviews on actor-based and incentive-based frameworks which form the broad background of the model presented here can be retrieved from [Law 1992] and other resources listed online at `http://www.lancs.ac.uk/fass/centres/css/ant/ant.htm [Nov. 1, 2006]`.

[2] `http://www.cybergeography.org/atlas/geographic.html [Nov. 1, 2006]`.

[3] For more details, see below Chapter Agents on different levels.

as status indications. There are many conceivable devices and means enabling fast electronic access to product and service information, and in some scenarios the "E" of E-Commerce might even hide in the back-end with the front-end consisting of printed matter[4].

In much the same manner as everyone today expects a business will have a telephone number, only niche or specialist businesses will be in a position to survive without being found via the Net. This is not to say that every small business needs a homepage and an own second level domain on which they display their goods and services. Conceivably, the Internet presentation of a small business might consist of nothing more than contact data allowing several offline and online methods to start transactions[5]. The crucial issue here is how these online presentations can be found. The business taking place in local stores will remain largely the same. What will in all probability change dramatically, however, is the way people get to know about these businesses. This elucidates how Commerce depends on search (the converse is also true, see below).

Until now, one of the most important means of acquiring contact data have been telephone books, namely White Pages and Yellow Pages[6]. White Pages provide basic contact information for almost all network participants, indexed by names. Yellow Pages are indexed by branches and feature both listings with contact information and more prominent placements. On the Net, telcoms and address aggregators usually offer White Page and Yellow Page services through one channel. Yet, the advantages of convergence — both in the form of multiple services packed in one channel and one service spread over several channels — also affect the distinction between general web Search Engines and Yellow Page web services[7]. While personal contact data such as telephone numbers are in general not found via general web Search Engines as they do not appear frequently (and even less consistently) on web sites, businesses, vendors, products and services can be found by them.

One thus faces three major shifts in searching for businesses. Firstly, from paper Yellow Pages to online directories with added value and easier (seamlessly online) means of contact. Secondly, a diversification of web queries that contain a growing number of queries which express needs conventionally posed to Yellow Pages: Business types or names in connection with geographical identifiers. Finally, a diversification of yellow page queries: no-one would expect to find a specific product type (such as a specific Samsung laptop) in a printed Yellow Pages. There is simply not enough space to include product keywords into the index. The same, however, is not true of online Yellow Page directories. Analysis of Yellow Page query logs reveal a significant

[4] For example, `Quoka.de` bundles classified ads from several local publishers, allowing to enter classified as on the Net that are then available online and in printouts.

[5] Cf. the ENUM scheme (`http://www.enum.org`) for a convergence of Telephone, Web and Email contact data.

[6] Yellow Pages is a trademark in many countries. This should not prevent its usage as a generic identifier for business directories sorted by branch categories.

[7] Recent studies in media adaptation and convergence processes can be found at the homepage of the Intermedia project, `http://www.intermedia.lmu.de/publikationen [Nov. 1, 2006]`

number of searches that do not fit into classification of business branches, but rather represent specific products, services or needs (see Part C).

One more note on another frequent conception of E-Commerce that deviates from the meaning proposed here. The retail sale of shippable goods with only a website as shopping place (Amazon.com's original line of business) may be the historic origin of E-Commerce, yet today a much broader variety of goods are traded online than just goods that fit through standard letter-boxes. While products that have a high value-to-weight ratio, are embarrassing to buy in person or ship to remote places still account for a considerable portion of E-Commerce transactions, new areas of online transactions have evolved up that do not even necessarily neeed any shipping component at all. High revenue sectors today also include media downloads, travel, financial and insurance products. The range of products and services searched for and transacted online does not differ fundamentally from those transacted off-line. A haircut will of course still be performed by a human hairdresser, yet searches related to hairdressing rank high both in generic Search Engine query logs as well as Yellow Pages query logs. Here are the top hairdressing and hair-care related searches from both types of logs, provided by Germany's goyellow.de and yahoo.de websites (see below, Term Spaces, for details on the query logs examined):

rank	*frequency*	*term*
650	10216	frisuren
1257	4751	haare
2113	2838	haar
2314	2575	haarpflege
2482	2400	friseur
2996	2007	haarausfall
3069	1961	haarfrisuren
3073	1958	kurzhaarfrisur
3356	1804	haarentfernung
3629	1666	friseurbedarf
5097	1219	frisur
6817	915	haarverlängerung
6900	906	friseure

Web query log (yahoo.de, 2003)

rank	*frequency*	*term*
2	44855	friseur
30	6953	frisör
85	2737	friseure
511	389	friseursalon
835	239	frisöre
1215	169	haarstudio
2278	93	friseurbedarf
2515	84	kraushaar
2555	83	haar
3496	61	haarentfernung
3666	58	haarverlängerung
3778	56	frisuer
4239	50	frisuren

Yellow Pages query log (goyellow.de, 2005)

Note that although there are no hairdressing-related queries in the top 500 ranks for the generic Web search query log (compared to three queries for the Yellow Pages log), the number of hairdressing-related queries in the top 4000 ranks is almost equal, ten versus twelve for the Yellow Pages logs. Apparently, the generic Web searches tend to lean towards information (a need that can be fulfilled online), whereas the Yellow Pages searches, which feature hairdressers as a top search (rank #2), tend to lean towards people and places (for a taxonomy of queries, see Part C). Yet, even the top queries presented here blur this distinction.

Finally, the annual W3B study illustrates in its 21st edition very graphically the diversification of online usage. It presented survey results in the ways people make use of the Web before Christmas[8]. Being the traditional peak-time for retail sales, more than half (59.6%) of the test persons plan to look for shopping ideas on the Net, and almost half of them (42.4%) plan to actually perform the shopping online. One in two is going to send Christmas greetings via E-Mail, one in three via E-Card. More than 10% of users plan to download seasonal images and another 5.9% looked forward to downloading Christmas-related music. In this context, it is, however, also worthwhile to note that the Christmas/end-of-year season lost some percents in the over-year revenues[9]. This can be interpreted as an indication that E-Commerce becomes more and more integrated in daily life.

While Commerce depends thus more and more on search, the converse is also true. Today's search providers, either general Web search or Local search, make a profit (if they do) through enabling business contacts between their users and the companies advertising on them. In a wider context, several types of online marketing have evolved

[8] Online at `http://www.w3b.org/ergebnisse/w3b21/ [Nov. 1, 2006]`.

[9] According to the GFK WebScope 2006 the ratio of revenues in the first half of the year to the revenues of the whole year has moved from 40% in 2002 to 48% in 2005. Cf. `http://www.gfk.de [Nov. 1, 2006]`.

in the last few years and not all of them take place on Search Engines. Among these thriving lines of business are, roughly sorted by market volume:

- Pay per click
- Affiliate marketing
- Search engine optimization
- Web banners
- Link campaigns
- E-mail and Newsletters
- Online viral marketing
- Content-driven marketing: Press releases, Blogs and Wiki etc.

All of these fields are to a large part term-based, and only the last and least important three one can be said to be representatives of a genuine *textual*-driven E-Commerce.

In summary, a wide variety of what is commonly understood as E-Commerce was shown to boil down to the concept of enabling and performing interactions between customers and companies. The matter quality that starts to be inherent to E-Commerce makes the distinctivess of the “E” in E-Commerce diminsh. According to this prediction, it will be rather the O-Commerce — the offline businesses — that one will need to pick out of the crowd.

3.2. From E-Commerce to TE-Commerce: The importance of terms

If E-Commerce is indeed an necessary further evolutionary phase of Commerce, one has to turn to the essence of Commerce to find the key properties of E-Commerce. A metaphor for Commerce is the market place, where needs and wishes on the one hand and products and services and the other hand are articulated. Bargaining begins with the prospective customer and dealer coming together, bringing their matching needs and offers along. It is not farfetched to state that finding such matches is a core part of Commerce, a challenge greatly facilitated by the advent of the Net. Vendors have to ensure they can be found by potential customers. In this respect, the concept of Yellow Pages as a paradigm of how to find companies by stating needs and wishes forms a central part of E-Commerce. How the transactions continue, once the contact is established, is another story, one which will be largely disregarded here. It is suffice to say that the classic conception of E-Commerce as Internet retail is founded on trust and a stable market environment — it requires trust to send money to someone never seen before for goods never touched before, but people are willing to do it if they have seen others who have previously done it without them facing problems. However, even if the actual transaction takes place via wire transfer and shipping of parcels or in a store, customers and dealers first have to be brought in contact. Speaking in the market place simile, as people are moving to the virtual market place, it becomes crucial to open up a booth there as well.

The conventional conception of the market place is characterized by the presence of transactors. As a communication and transaction platform it opens up place and time for sellers to present their offers, for buyers to declare their needs and for both to enter into transaction discussions and finally to exchange goods and money. In E-Commerce, writing is still the most stable, efficient and time-saving means of communication. This is not to denigrate the importance of other media such as images or videos, but these are much more one-way-media than is writing. It may be speculated whether the combination of ease of perceiving and ease of production is indeed a specific trait of writing, and whether this can be stated as a cultural universal. For the time being, there is no superior uniform way of entering, retrieving and displaying information available.

Keystrokes are a very efficient way to get across information, especially if fault-tolerant systems allow for typos and interaction takes place quickly, as for example with suggestion tools that start offering completions while the user still types[10].The selectional power of typed text is high, if measured in possible entries per time, yet it is redundant enough to allow even only remotely similar entries to be corrected by approximative matching technologies.

To summarize, E-Commerce interaction paths move in close connection with textual representations. Users make selections based on information presented textually and conduct the selections by either clicking on textual links or by typing in text. If one promise of the late nineties boom years has been thoroughly discredited, it is the then-predicted move towards non-textual representation modes. While images and video play a considerable role in E-Commerce systems, access is almost without exception based on textual tagging (see for example youtube.com). Image and video searches demonstrate how adding textual information to multimedia data allows efficient retrieval[11].

The importance of search for E-Commerce transactions can be experienced by anyone running a Web shop. A recent study revealed that about one in two online purchases starts with search[12], in certain vertical segments such as tourism the percentage is reportedly even higher. Searches leading to purchases do not only take place im-

[10] For demonstrations, refer to Google's query suggestion at `http://www.google.com/webhp?complete=1&hl=en [Nov. 1, 2006]`, SurfWax's LookAhead at `http://lookahead.surfwax.com/ [Nov. 1, 2006]` and the fault-tolerant Exorbyte MatchMaker suggest demo, presented at `http://www.exorbyte.com/MM_Suggest.htm [Nov. 1, 2006]`.

[11] To illustrate the contrast one may try current demos of image queries by sketch, for example `http://labs.systemone.at/retrievr [Nov. 1, 2006]`. It is almost impossible — at least in reasonable input time — to retrieve a picture seen previously.

[12] Source: Doubleclick's "Search Before the Purchase" study, released 2nd of March 2005. Online summary at `http://www.doubleclick.com/us/knowledge_central/documents/research/searchpurchase_0502.asp [Nov. 1, 2006]`.

mediately before the transactions, but often happen during several weeks prior to the purchase; in the tourism segment more than half of the buyer's final searches occurred at least two weeks before the transaction. In this context, generic keywords (such as "running shoes" as opposed to brand names) amount for the majority of searches. In particular in combination with brand names, for example "Nike running shoes", they generate high click-through rates, i.e. the ratio of clicks on a specific site to searches that bring this site up. Combinations of generic keywords and brand names account for only 1% of searches for apparel sites but 3.7% of their clicks.

Of course, the importance of textual representations is hardly a unique invention of the online age. Printed Yellow Page books are organized according to categories. Using them involves looking up the correct category (some category headers only serve as pointers to preferred categories) and then choosing from the business listings or placements. Both the selections of the user and the presentation of information operates on terms. The main change brought by online presentation of material therefore takes place not so much with regards to what is displayed simultaneously (considering that the usual digital displays are of comparable size to printed matter), but rather with regards to the spectrum of what might be displayed[13]. Choices between hundreds of thousands of keywords or index terms are possible in the online media, while delivering results with an acceptable perceived relevance for each choice. However, setting up frameworks to properly deal with such vast sets of terms is onerous. The richness of data available on the Web makes the harvesting of printed resources worthwhile only for a limited number of cases[14].

Term-driven E-Commerce, or TE-Commerce in short, is characterized by the crispness of information. It might change into a full textual driven E-Commerce over time. For now, terms are the common device to initiate interactions. The apparatus for this — search sites, catalogs, directories etc. — have rather short interaction cycles, either leading the user to the contact data or order form sought for or (more often than not) leaving the request unanswered[15]. In accordance with the temporal brevity of online interactions, the information display used for them is often crisp too: search queries are in general a couple of words and E-Commerce results regularly comprise address information, product snippets and partially filled out forms[16]. The reason behind this is clearly that a large proportion of transactions take place off the Web, via e-Mail inquiries, phone calls or face-to-face interactions. This leads to a paradoxical situation

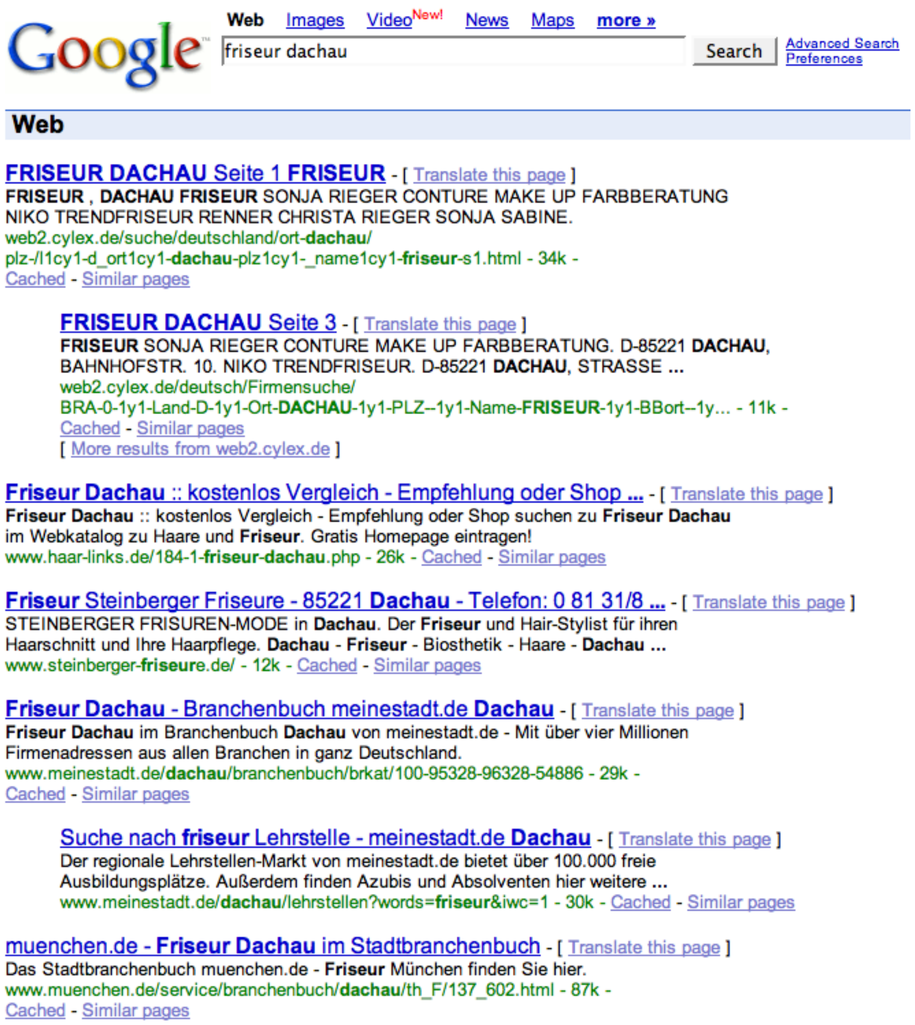
[13] To put it in a structuralist's way, the syntagmatic and paradigmatic axes.

[14] In fact, US Yellow Pages companies such as Amacai or Acxiom did the scanning and processing of printed Yellow Pages and thus could accumulate US-nationwide databases of businesses, classified according to their proprietary category systems or to Standard Industrial Classification (SIC) codes. However, with the constant flux of companies commencing and ceasing operations, these resources are usually only of limited worth after some years.

[15] User lab studies corroborate the limited time span of most search sessions, leading for example to the well-known result that listings after the first two result pages on Search Engines are hardly ever read. See [Machill/Welp 2003], 340-346.

[16] Arguably, the advent of user-generated media fills in these gaps. A striking illustration is for example `ciao.de`, where users not only comment on products, but also on other users' comments. This is indeed an autopoetic text production mechanism.

in terms of E-Commerce information availability: While a lot of valuable and reliable E-Commerce-relevant information is hidden in the vastness of online resources, many businesses do not reveal more than the most basic contact information, even if wrapped up in fancy graphics. For example, searching for hairdressers in Dachau yields numerous aggregator sites that may or may not contain helpful information, but very few genuine Dachau hairdressers' homepages:



friseur dachau

FRISEUR DACHAU Seite 1 FRISEUR - [Translate this page]
FRISEUR , DACHAU FRISEUR SONJA RIEGER CONTURE MAKE UP FARBBERATUNG NIKO TRENDFRISEUR RENNER CHRISTA RIEGER SONJA SABINE.
web2.cylex.de/suche/deutschland/ort-dachau/plz-/l1cy1-d_ort1cy1-dachau-plz1cy1-_name1cy1-friseur-s1.html - 34k -
Cached - Similar pages

FRISEUR DACHAU Seite 3 - [Translate this page]
FRISEUR SONJA RIEGER CONTURE MAKE UP FARBBERATUNG. D-85221 DACHAU, BAHNHOFSTR. 10. NIKO TRENDFRISEUR. D-85221 DACHAU, STRASSE ...
web2.cylex.de/deutsch/Firmensuche/BRA-0-1y1-Land-D-1y1-Ort-DACHAU-1y1-PLZ--1y1-Name-FRISEUR-1y1-BBort--1y... - 11k -
Cached - Similar pages
[More results from web2.cylex.de]

Friseur Dachau :: kostenlos Vergleich - Empfehlung oder Shop ... - [Translate this page]
Friseur Dachau :: kostenlos Vergleich - Empfehlung oder Shop suchen zu Friseur Dachau im Webkatalog zu Haare und Friseur. Gratis Homepage eintragen!
www.haar-links.de/184-1-friseur-dachau.php - 26k - Cached - Similar pages

Friseur Steinberger Friseure - 85221 Dachau - Telefon: 0 81 31/8 ... - [Translate this page]
STEINBERGER FRISUREN-MODE in Dachau. Der Friseur und Hair-Stylist für ihren Haarschnitt und Ihre Haarpflege. Dachau - Friseur - Biosthetik - Haare - Dachau ...
www.steinberger-friseure.de/ - 12k - Cached - Similar pages

Friseur Dachau - Branchenbuch meinestadt.de Dachau - [Translate this page]
Friseur Dachau im Branchenbuch Dachau von meinestadt.de - Mit über vier Millionen Firmenadressen aus allen Branchen in ganz Deutschland.
www.meinestadt.de/dachau/branchenbuch/brkat/100-95328-96328-54886 - 29k -
Cached - Similar pages

Suche nach friseur Lehrstelle - meinestadt.de Dachau - [Translate this page]
Der regionale Lehrstellen-Markt von meinestadt.de bietet über 100.000 freie Ausbildungsplätze. Außerdem finden Azubis und Absolventen hier weitere ...
www.meinestadt.de/dachau/lehrstellen?words=friseur&iwc=1 - 30k - Cached - Similar pages

muenchen.de - Friseur Dachau im Stadtbranchenbuch - [Translate this page]
Das Stadtbranchenbuch muenchen.de - Friseur München finden Sie hier.
www.muenchen.de/service/branchenbuch/dachau/th_F/137_602.html - 87k -
Cached - Similar pages

A screenshot of Google.com results for "Friseur Dachau" (hairdresser Dachau), taken on Sept., 15, 2006. Only one listing (the fourth from the top) actually refers to a Dachau hairdressers' homepage.

Given the scarcity of sound E-Commerce-related information online, hidden in the vastness of online data, one might start to feel like Buridan's ass, undecided which way to look. As terms, either as search queries or as navigational aids (anchor texts, i.e. the clickable text of a link), are the main vehicle for a user to convey needs, recognizing and analyzing correctly as many terms to relevant results as possible is crucial for the success of E-Commerce.

3.3. Term-driven E-Commerce Enhancements

Given the dominance and abundance of terms, the need to deal with them methodically becomes apparent. In particular, it is hoped to achieve the following goals through a better handling of terms:

- Facilitating handling of E-Commerce interactions, thereby enabling a consistently satisfactory user experience

- Matching customers' needs and businesses' offers more effectively and efficiently (increasing both the number and the quality of matches), thus inciting additional transactions

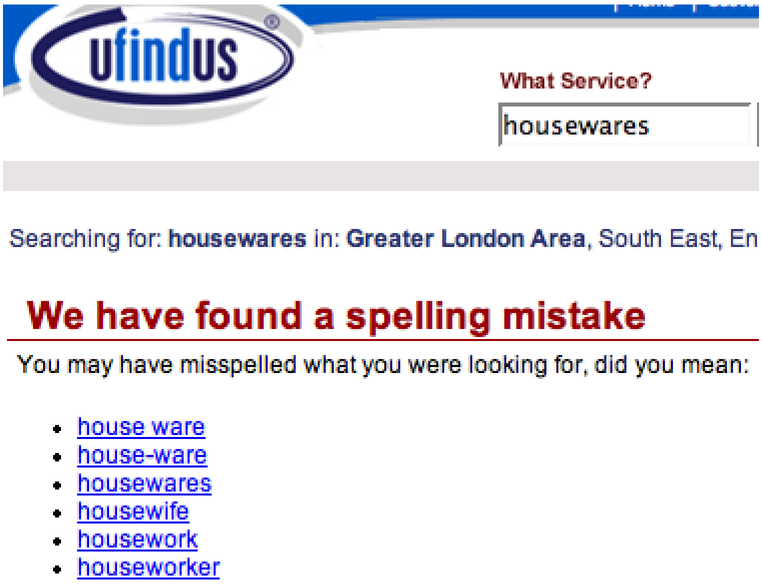
- Allowing a more precise monitoring of E-Commerce interactions, for example for purposes of marketing, product development, sales or regulatory measures such as fraud detection, tariff or taxation

In what follows, the steps necessary for a proper treatment of E-Commerce vocabulary and how applications could benefit from these steps are presented.

A first step is the *recognition of terms*. This includes segmenting textual input flow, for example downloaded Web sites or search engine query strings, into meaningful units and recognizing variate repetitions of these units[17]. This process makes observations on terms and keeps track of them. The discrimination between “already seen” units, “really new” ones and “somewhat familiar” terms plays a crucial role in setting up a system dealing with terms. In this thesis, the recognition of terms plays a minor role, as the corpora examined are non-prose textual resources and already offer a segmentation in terms, for example through individual queries or meta keywords.

Once recognition is performed, various types of *term handling* take place as a second step. One important way is the use of terms for retrieval which relies on a matching function from input terms to index terms in the set of retrievable data. The simplest matching function naturally is literal string identity, but it is neither a very precise nor at all a comprehensive matching method. Another way of term handling is the normalization of terms, i.e. matching variant forms by one canonic form. The advantage of canonic forms lies in removing redundancy from what the systems output. This is something different than telling the user what terms to choose and what not. For example, the following "Did you mean"-types of suggestions are not normalized with a detrimental effect on their benefits:

[17] See [Bourigault/Jacquemin/L'Homme 2001].



What Service? housewares

Searching for: **housewares** in: **Greater London Area**, South East, En

We have found a spelling mistake

You may have misspelled what you were looking for, did you mean:

- house ware
- house-ware
- housewares
- housewife
- housework
- houseworker

A screenshot of ufindus.com spelling corrections, taken on Sept. 15, 2006. Redundant entries (singular-plural, genitives etc) lower the perceived relevance of the suggested corrections.

In addition, *term grouping* (categorization, classification, clustering etc.) is another common application of term handling. As term groups themselves typically make use of textual representations also (for example, headings on Yellow Page sites are displayed as textual strings, though they are internally represented as IDs), grouping might be viewed as establishing a relationship $R : T \rightarrow T$, with the value set consisting of a subset $T_{class} \subseteq T$.

A final process, *spotting term relations*, step follows from the realization that E-Commerce terms are not isolated units with no connection to each other. A variety of different types of interconnections may be found between them, some of them universal relations such as synonymy (*notebook – laptop*), some specific to Commerce, such as brand name – product name (*ipod – mp3 player*). Despite numerous publications on single aspects of relations between terms, there is yet no integrative model of what relationships should be modeled for E-Commerce terminology and how they could be spotted (see below, Part B)[18].

How is it possible to accumulate any substantial part of E-commerce vocabulary? On first sight, the variety of terms seems to render such a task impossible. The same, however, could be said about language dictionaries and yet the creation of considerably covering electronic dictionaries was possible and has spawned many fruitful applications.

Five fundamental concepts of natural language processing (if not of any empiric observations) facilitate the labor involved in gathering vocabulary:

[18] For a monography on term variation, see [Jacquemin 2001].

- Dividing simple vs. complex units
- Recognizing variant repetitions
- Keeping track of known units vs. new units
- Working along frequency of observation
- Iterative deepening of the granularity of description

The basic axiomatic concept here is the division of simple and complex units. Any observed unit is either a simple unit or is build up from several simple units, forming a complex unit. This reduces the amount of what has to be described to simple units and the rules of how simple units combine to complex units. Depending on the nature of what is described, simple units might be of different sizes, for example spanning words, phrases or sentences. In TE-Commerce, brand names and product types could form suitable simple units of the vocabulary, while combinations such as "Siemens handy" and "Samsung notebook" are then analyzed by means of a rule stating that brand name plus product type forms an E-Commerce relevant term.

Starting from this separation of simple and complex units, variant repetitions might be observed for both types. Variant repetitions refer to the re-appearence of items already seen, although in a modified form*Even if the modification just lies in the fact that some amount of time has passed between the two occurrences.* For example, in the sentence "Samsung notebooks in general and the X-05 notebook in particular are known for their light weight", *notebooks* reappears in the singular form. For TE-Commerce, as for natural language in general, the most prominent kinds of variations are those related to the form of units (this connects to the branch of morphology), the ordering of units (this is one part of syntax) and substituting lexical units while preserving the meaning (this connects to semantics).

Keeping track of the observations already made is another way of putting the idea of setting up TE-Commerce lexica. Through devices that can detect variations, many observations that seem new at a first glance may really be regarded as another instance of variant repetition. This makes genuine new units even more interesting, as these refer to emergent phenomena. In TE-Commerce, the introduction of a new class of products (such as *mp3 players*) can be regarded as an instance of introducing a genuine new unit[19].

If one keeps track of observations over a certain time span, some observations will occur only once and others repetitively; among the repetitions, some occur more often than others. Keeping the list of number of occurrences over a given span of time results in a frequency list, i.e. the list of all distinct occurrences together with the number of times they appeared.

In setting up such lists it is possible to discern new trends, periodical fluctuations and to determine the core set of terms that recur over a large span of time. For instance,

[19] Obviously, there will always be ways of incorporating these newly-emerged products into already existing groups. In the case of *mp3 player*, one could look for the common heading for products such as *cd player* or *walkman* and insert them at this place.

query logs from the beginning of a week differ from those of week-ends, in size and in content. Frequency lists of summer queries differ from those of winter queries. The shorter the log time becomes, the greater the impact of short-time trends: One day of a query-log features high frequency “words of the day”, derived for example from events that recently took place[20]. On the other hand, if gathered over a longer time, the periodic cycles are affected by general trends, by the influx of new terms needed for innovative products and by long-term changes in user behavior. Some aspects of comparison between frequency lists will be studied in more detail in the chapter on Term Spaces below.

Yet frequency lists serve also as means of prioritizing labor. They make it possible to begin with high-frequency terms and then continuously work down. Working through the alphabet when preparing terms for E-Commerce applications would either only allow completed lists before going live with the enhancements or leave the users with a puzzling break in the system’s behavior from one letter to the next. Both options are not feasible. Similar to interlaced JPEG images that allow a first impression of the full picture before the download is complete, working along a frequency list takes care of the frequently-asked terms first. It has to be conceded that this is an idealized view, as looking at previously-seen data might not always lead to valid predictions on future user behavior. Moreover, users do not have this particular frequency list in mind, so it will seem puzzling why the system would deal with one term perfectly, yet fails with another which only differs in a measure not transparent to the user.

A last principle that facilitates the process of describing TE-Commerce units is to allow iterative amendments to the lexicon. As it is not feasible to describe everything in full depth initially, one may start with a basic set of descriptors and work both in the dimension of adding new units as well as populating descriptors for ones already seen. Another rationale for following this method is that the selection of features that need to be described should result from the observation of many instances, rather than stem from a priori considerations.

[20] For examples, browse Google’s Zeitgeist at `http://www.google.com/intl/en/press/zeitgeist.html [Nov. 1, 2006]`

4. TE-Commerce Agents

After introducing the notion of Term-driven E-Commerce, it is only natural to ask what players and stake-holders build up the world of TE-Commerce. The necessary steps involved in examining TE-commerce agents are as follows:

- Setting up a framework in which all relevant stake-holders of TE-Commerce are integrated
- Introducing the main characteristics and statistics on these stake-holders
- Commenting on important trends and innovations affecting these agents
- Commenting on important threats and opportunities affecting these agents
- Reviewing research on these agents, especially on users, in the light of its relevance to TE-Commerce

The figures and characteristics presented below cover the global, the European and the German market. In comparing these data, the specifics of the German market are outlined. Moreover, the differences observed between the different markets hint at the variety of shapes E-Commerce can take in different settings. Obviously, such a comparison is only applicable to markets that are in principle fit for E-Commerce, thereby excluding developing countries which lack the prerequisites of a digital infrastructure and the necessary social and legal framework. While E-Commerce parameters such as what products and services are sought for, how they are shipped, who runs E-Commerce businesses or what payment methods are preferred are by no means uniform across the globe, the core issues of Term-driven E-Commerce remain remarkably stable. In all markets and for all languages, E-Commerce requires term handling, including term spotting and recognition of variants and related terms. While the linguistic methods and examples provided below are related to both German and English, many observations and conclusions should apply to much larger range of languages, markets and cultures.

While in this and the following chapter the abstract model of TE-Commerce is backed up with figures and enriched with details, it still remains a model, meaning that the agents introduced do not map in all cases to real-life person or organizations. Persons or organization may in fact fulfill several roles in the model. For example, a particular private homepage owner does not cease to be a user most of the time even if acting as a content provider in the particular role of a homepage owner. Moreover, agents are not to necessarily restricted to humans or organizations. Technical artifacts, for example the algorithmic ranking mechanisms of Search Engines, might be as well considered as agents or at least as components of agents[1].

Introducing the agents involved in TE-Commerce sets out from the paradigm of the virtual market place, understood as the interface where buyers, sellers and intermediate agents meet, communicate and transact[2] The main differentiating characteristics of virtual market platforms are the instant information and interaction cycles that allow both spatially and temporally flexible market activities. The advantages of spatial and temporal flexibility have to bought on the other hand by a rigidity on the form with which information and interaction are exchanged. On the Net, this form is typically a packet that is governed by standard protocols[3]. Whereas traditional market places allowed for heterogenous means of exchange, the virtual market requires a uniform and compact representation of offers and needs. This need for a compact representation that can be exchanged via packets is one origin of the importance of textual encoded information in the virtual market. An illustrative example is provided by the comparison between a pre-digital-age flea market and eBay-like sites.

The *flea market* takes place at a physical place; sellers might travel around to participate in different flea markets all over the country, but this is cumbersome (buyers will even be less prone to do so). Various impressions will influence the purchase of items — not only based on the items themselves, but also whether the potential buyer likes the face of the seller, the arrangement of items on the table, the company of other people looking at these items, the weather shading a favorable light on the items and so on. Compared to a virtual market, there is a remarkable variety of media and senses involved. On virtual markets, apart from images of products, all is text: the description of the product, the categorization of the product inside the product tree, even the rating for the seller is represented in the form of short textual evaluations (for example on eBay's ViewFeedback pages). The physical process of packing the purchased item in a parcel and sending it off is in many cases the only remnant of the physical world. Without face-to-face contact between the different market place agents, their digital messengers — textual descriptions, figures, images — become the only way of offering and seeking for goods.

4.1. Agents on three levels

As in any market place, the primary agents on virtual market place described above are buyers and sellers — or, more generally, those looking for products and services (consumers) and those offering them (businesses). Buyers and sellers are 1^{st} degree agents. These agents can in reality be represented by all different types of entities: individual persons, companies with various legal status, public agencies and many more.

[1] This complies to the actor-network theory first developed by Bruno Latour (see [Latour 1991]).

[2] For general theories on market places, see standard textbooks in macroeconomics, for example [Pindyck/Rubinfeld 2004].

[3] For a treatise on packet-switched networks, see Noam Eli's article on nano-regulation, online at `http://www.citi.columbia.edu/elinoam/articles/con_info_money.htm [Nov. 1, 2006]`.

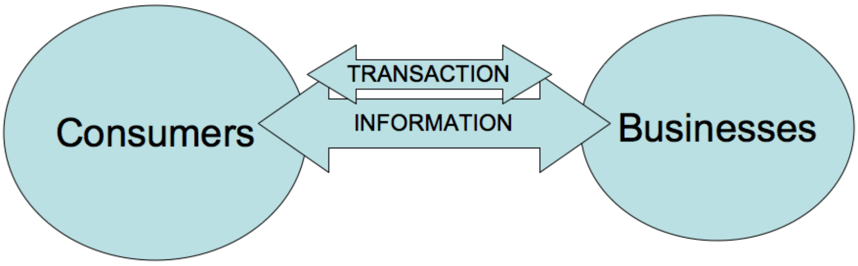
These two groups are the immediate participants of the virtual market place. By introducing two sets of entities, a new definition of a virtual market place can be obtained: The virtual market place may be viewed as the Cartesian product of buyers (B) and sellers (S):

$$M : U \times B$$

These set contains all the potential relations between customers and users. All sort of interactions (communicating, expressing interest, bargaining, purchasing, complaining etc) form then subsets of M, with transactions Ti of a kind i:

$$Ti \subseteq M.$$

This leads to the following schematic summary of the first two agents of TE-Commerce:



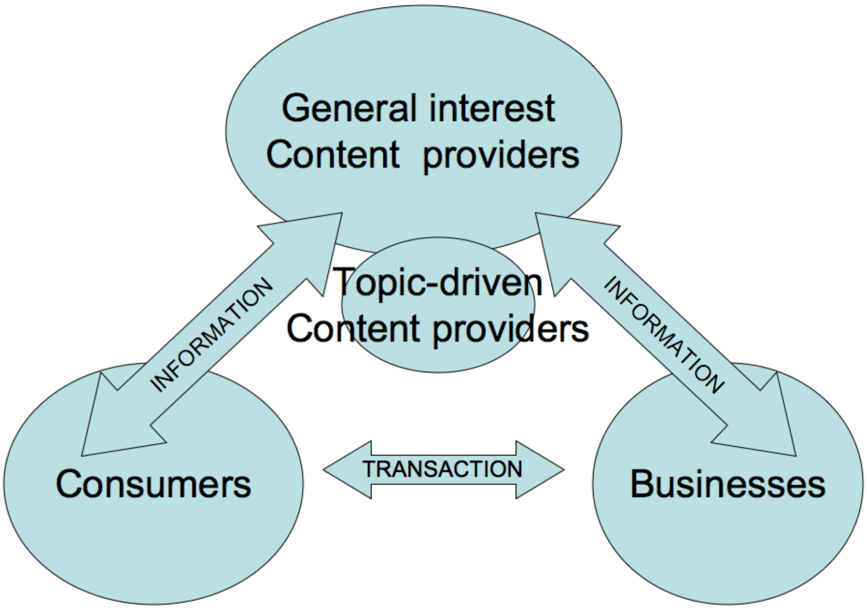
Consumers and businesses as principle agents of TE-Commerce

The picture of first-level-agents on the virtual market place would not be complete, however, without introducing content providers. This agent is a genuine contribution of the virtual market place; it has no correspondence in the flea market model.

Content providers can provide one of two functions in the TE-commerce model, either gathering (or channeling as it sometimes is called, though this springs from a much too strict steering metaphor) users in one place or bundling users with specific topical interests, backgrounds and needs. These functions are provided by different types of content providers. A popular news website for example gathers users with very little discriminative topical characteristics, apart from very broad political or sociographic characteristics (liberal vs. conservative online newspapers come to mind). The opposite holds for example for the usually smaller sites dedicated to one topic (fansites, sport or hobby sites, sites with a local scope etc)[4].

In the model of TE-Commerce, content providers either function as businesses if they provide paid content, or as aggregators of users. By virtue of that aggregation, a content provider might refinance itself through displaying advertisements. Both the gathering of many users as well as the accumulation of users with very specific interests can be monetized as space for advertisements.

[4] This is not to say that a large website may not host pages devoted in an equal grade to one topic, yet the entry page remains an accumulator of unspecific traffic.



Content providers and TE-Commerce first level agents

Whether agents participate immediately on E-Commerce activities or aggregate such immediate agents leads to a subdivision of agents into 1^{st} degree agents and higher degree agents. As immediate participants, consumers and companies are 1^{st} degree agents. Aggregators such as Search Engines operate on a level above the immediate participants; they will be referred to as 2^{nd} level agents. In what follows, even third level agents such as search technology providers or regulatory bodies are introduced on an even higher level, serving or controlling second level agents.

In general, 2^{nd} degree agents are entities enhancing the matching between 1^{st} degree agents. They are not immediate participants in the market-place, but both aim at and ultimately profit from successful matchings. Among such 2^{nd} degree agents are Yellow Pages Directories, Search Engines, and also Search Engine Marketing companies.

Yellow Pages Directories brings users and companies together. From the perspective of bringing together consumers and companies, there is no fundamental difference between Yellow Page directories and Search Engines. Both are used for finding companies offering the sought-for product or service and the functionality of Search Engines to retrieve URLs or browse non-commercial topics is from this viewpoint merely an additional benefit that Search Engines provide to users. Today's business model of Search Engines mainly depends on online marketing in its various forms (paid placement, paid inclusion, pay-per-click etc). However, the increasing possibility to monetize content sites via targeted advertising will without doubt also play into the hands of search providers, as the quantity and quality of traffic they direct to content sites directly determines the sites' economic success. This would bridge the difference between Search Engines functional place — that of an intermediary between users and content sites — and their typical business model of today, based on advertising revenues. While the business model works well for the big Web Search Engines[5], it is logical step trying to convert more and more content providers into paying customers, especially if the traffic provided to them directly leverages revenues for the content provider. The possibility of exactly measuring the economic benefit for a content site that the Search Engine's traffic brought to it allows a mutually beneficial situation. One popular way of transforming it into a contract model is revenue sharing.

While Search Engines have a benefit in providing free results to users (the more users regularly use the Search Engine, the more high qualified traffic it can dispatch to advertisers), it is not ruled out — especially under the given oligopoly of Search Engines (see below) — that they will at one stage try to monetize the users, for example through bulk sales of query capacities to Internet Service Providers. In such a scenario, a user could either choose to stick to her Internet access and only access a limited functionality of the Search Engine, or upgrade the Internet access with a flat fee for queries that are answered based on the full resources and capabilities of the Search Engine. These possible developments and other issued related to the power of Search Engines as information access providers have just recently received attention from media politics[6].

The most prototypical $2nd$ level agents, one which is dealing with users, companies and content providers, are Search Engine Marketing companies, often abbreviated SEMs. While in reality, the big Web Search Engines have their own SEM division, either acquired (Overture-Yahoo) or largely self-developed (Google's Adwords and MSN's AdCenter), a SEM plays a different role in the framework evolved above. Through delivering ads to users via content sites, SEM bring users, companies and content sites together.

While the framework is in principle closed with the two first levels (meta Search Engines for example do not add anything new to it; they still bring users and content together), the agents concerned with the framework itself have also been taken into account in sketching out the full picture. These 3^{rd} degree agents are concerned with facilitating, monitoring and regulating the interactions between agents on the first two levels.

Whether 3^{rd} degree agents operate on existing structures such as Search Engine Optimizers or seek innovative methods to enhance search performance, such as Search Technology providers do, they are not immediately involved into bringing needs and offers to a match, but rather provide solutions and services working as a leverage to search performance. A separate group of 3^{rd} degree agents is built up by entities whose function lies in supervising the framework and detecting players who are breaching its rules. These entities can be private watch-dog organizations such as Search Engine Watch[7], but also official regulatory agencies. Using such flexible inclusion criteria, one should also count the fraud-detection groups within pay-per-click providers as a third

[5] Google's advertising wins run at about 733 million dollars in the *three months* from July to September 2006. See `investor.google.com/pdf/20050930_10-Q.pdf [Nov. 1, 2006]`.

[6] See as one example, Germany's Green party pamphlet *Suchmaschinen: Das Tor zum Netz*, online at `http://www.gruene-fraktion.de/cms/publikationen/dokbin/63/63265.pdf [Nov 1, 2006]`.

[7] `http://www.searchenginewatch.com`.

degree agent.

When examining these agents, the question of what methodological approach is best suited to determine their characteristics arise. It stands to reason that there is no single best solution applicable to all different agents. Different empirical methods, either based on surveying users or analyzing their trails in the form of log data will be discussed.

4.2. Users

Users certainly comprise the single most decisive agent in the TE-Commerce framework. However, the name *users* does not reveal much; there is even a shade of circularity in pointing to their importance. However, this is mostly due to the overly general name to which there is no commonly accepted alternative. To specify users as one agent in the TE-Commerce model in more detail, different facets of users will be examined.

4.2.1. User statistics

In the following section, general statistics on German, European and world-wide Net penetration and usage are presented. Starting with Internet access rates, the focus then moves onto specific types of Internet usages, namely E-Commerce activities, searching and finally E-commerce-related searches.

Onliners

Well-known data sources on German online behavior are provided by the Destatis (Statistisches Bundesamt) — the official statistics agency in German —[8], the ARD/ZDF-Online Studies[9] and the W3B Studies[10]. While there are numerous statistics for Internet penetration on large geographical scales, these aggregated views do not provide the same level of granularity in observations, especially not covering a longer period of time, as the above mentioned succession of surveys conducted for the German market do.

In general, Internet usage in Germany has seen a consolidation in recent years with a sinking growth rate in online population. The number of off-liners did not substantially shrink and is not likely to do, unless Net access comes in a different shape, cloaked in traditional devices such as TV sets. However, the group of offliners cannot simply be identified as being only elder and less well-to-do people. Many offliners choose to stay disconnected from the Internet willingly[11]. In this context, it is elucidating to

[8] http://www.destatis.de.

[9] http://www.daserste.de/studie/.

[10] http://www.w3b.com.

[11] See "Digital Divide could deepening", online at http://news.bbc.co.uk/1/hi/technology/6085412.stm [Nov. 1, 2006].

note how the Net has lost the property of being a medium for the youth over the last decade.

Comparing the age distribution of Net users in Germany from 1995 to 2005, it is striking how the percentage of young users has declined, thus making the Net users more similar to the overall population[12]:

Age	*User distribution in 95*	*2000*	*2005*
19y and below	N/A	4.8%	4.0%
20y to 29y	62,6%	30.4%	22.5%
30y to 39y	N/A	38.9%	25.8%
40y to 49y	N/A	18.6%	25.9%
50y and above	2.5%	12.4%	21.9%

On an European scale, German represents the upper average of EU countries in terms of broadband penetration, onliners and percentage of E-Commerce users[13].

Globally, the distribution of Net penetration is rather skewed; penetration rates span from 2.6% (Africa) to 68.6% (North America)[14]. However, if the growth rate is also taken into account, one can observe approximately an inverse proportional relationship between penetration rate and usage growth. Although one cannot infer from these figures that the digital divide will close down fully, they indicate that it has not become larger in the last five years:

Continent	*Internet penetration 2005*	*Usage growth 2000-2005*
Africa	2.6 %	423.9 %
Asia	10.4 %	232.8 %
Europe	36.4 %	179.8 %
Middle East	9.6 %	454.2 %
North America	68.6 %	110.4 %
Latin America/Caribbean	14.7 %	350.5 %
Oceania / Australia	52.6 %	134.6 %
World total	16.0%	189.0%

Within one society of one country, a social digital divide that deprives in general the poor, lesser educated or female part of the population from Internet access and usage can be often observed. In general, such deprivation seems to follow general discrimination that holds for most aspects of public life, especially commercial life. The recent figures for Germany reveal a closing down of the male/female digital divide. Although the ratio of all male users to female users runs still at ca. 55:45 (figures for 2005[15]), the ratio of all teenage male users to female users is virtually 50:50[16].

[12] See W3B study 21, online at http://www.w3b.org/ergebnisse/w3b21 [Nov. 1, 2006].

[13] See http://www.destatis.de/presse/deutsch/pk/2005/ Statement_IKT.pdf [Nov. 1, 2006].

[14] See http://www.internetworldstats.com/stats.htm [Nov. 1, 2006].

[15] See ARD/ZDF Online-Studie 2005.

[16] See the JIM study 2006, online at www.mpfs.de/fileadmin/JIM-pdf06/JIM-Studie_2006.pdf [Nov. 1, 2006].

On a personal level, there are indications from surveys that most users are accustomed to the Net to such an extent that depriving them of online connections causes considerable strain. The "Internet Deprivation Study"(Ipsos-Insight) demonstrated how users accustomed to the Net became psychologically strained and restless if it they are deprived from it. Test subjects were reluctant to resort to off-line alternatives such as telephone books[17].

Summing it up, it is not only remarkable how quick the Net has achieved high penetration rates in developed countries, but also how thoroughly and apparently irreversibly it has changed information and communication patterns both on a business as well as personal level.

E-Commerce users

With a deep-ingrained Internet usage shaping many people's daily life, it is obvious that different spheres of live are affected through the Net. One of these spheres is shopping.

The number of online shoppers in Europe is predicted to increase from 100 million to 174 million by 2011, according to a Forrester Research study. In total, this would lead to an E-Commerce market in Europe worth 263 billion EUR[18]. Until 2008, Germany's market is expected to grow to over 85 billion EUR in 2008[19].

Although E-Commerce is already widely established in European countries, many voices still point to the typical worries and shortcomings that prevent people from using E-Commerce[20]:

- Presence of shopping guide
- Happy with the current offline retail — no need for change is felt
- Physical experiences of purchases
- Distrust in paying methods

However, in the broader understanding of E-Commerce that goes beyond mail-order and also includes interaction enabling in the pre-sales and after-sales phase, most of these and similar claims no longer prevail. In contrast, information online has a considerable impact on deciding what to buy in many domains. The 20th W3B study revealed that more than two third of Net users search for product information and comparison, while over 60% of these went to an offline shop for the actual transaction[21].

[17] See `http://docs.yahoo.com/docs/pr/release1183.html f [Nov. 1, 2006]`.

[18] See `http://www.forrester.com/Research/Document/Excerpt/ 0,7211,38297,00.html [Nov. 1, 2006]`.

[19] `http://www.infoedge.com/product_type.asp?product=EM-2222& c1=nav&source=maintopic [Nov. 1, 2006]`.

[20] See `http://www.destatis.de`.

[21] http://www.w3b.org/ergebnisse/w3b20/

Other studies report the percentage of onliners that use the Internet to get informed on products and services at 82%[22].

The following figures are taken from the same W3B survey and indicate the percentage of online users that stated online information had been *very important* in deciding purchases.

Percent of Users	*Domain*
46,2%	Automotive
40,5%	Phones and Mobiles
36,0%	Entertainment Electronic
24,1%	Domestic Electronic
15,6%	Furniture
12,9%	Pharmaceutics
11,1%	Fashion
9,1%	Cosmetics

While the top position for Automotive is explainable by the high item costs, the selection of other domains is a wide range of the so-called Slow Moving Consumer Goods (in contrast to Fast Moving Consumer Goods typically offered in a supermarket). Some of these domains are also areas of high competition between paid keyword advertisers (see Part D). The broad range of these domains should once again illustrate that E-Commerce is not restricted to retail through letterboxes.

E-Commerce related Search Engine usage

Browsing the Web without any use and help of search engines can be a most eclectic experience. Doubtless, there are people who do not go beyond their providers' portal content and link offers, not aware of the Net's essential pull mode that is invoked by typing in queries. For the year 2002, a 91 % usage of search engines for internet users has been reported [23].

It is conceivable that some users do not even use the URL field of their browser, but type in URLs in a Search Engine to get a clickable link. The high occurrence of perfect URLs in a query log (about 3%-5% of all query tokens, see Part C for details) suggest that this behavior is not totally uncommon.

The percentage of E-Commerce-related queries is hard to determine. While [Spink et al 2002] report 12.7% *Commerce, travel, employment, or economy* topic queries on Alltheweb in 2002, a report by U.S. Bancorp Piper Jaffray state that 36% of Web queries have a commercial background[24]. However, [Spink et al 2002] other topics (such as *Computers or Internet*) might contain E-Commerce-relevant queries as well, however. Looking at the top 193.195 Web search queries taken from a German search engine query log, however, only a meagre 24.5 % do not produce adwords on

[22] See `http://www.destatis.de`.

[23] See [Machill/Welp 2003].

[24] See [Rashtchy/Avilio 2003].

Google, and many of these for policy reasons, for example against advertising hardcore pornography or gambling[25]. This means that 75.5 % of top searches have a potential E-Commerce value, given that advertisers decided to reserve these keywords for their ads. It should be noted that most gaps are because of Google policies (adult terms or brand names). Detailed figures of this campaign database will be presented below in Part D.

A complementary figure runs at 70% of all online transactions springing from a search query[26]. This means that at the beginning of an online transaction (purchase, download, subscription etc.) a search query of the user initially opened up the way towards the vendor's site. A GFK study in 2005 revealed that out of the 27.4 million Germans that use the Net for gathering information before purchasing items — both at online and offline shops—, the vast majority (about 90%) go to Google first[27].

On online shops, it is reported that more than 50% of users go straight to the search box before looking at navigational elements and that one third of users experience a failure in searching for a product, even though it is in stock[28].

In summing these figures up, it is indisputable that firstly E-Commerce-related activities have a considerable share in users' interests on the Net and secondly, E-Commerce cannot be disentangled from search.

4.2.2. User related research

Exploring characteristics of users leads to a challenge in methodology which does not occur when surveying other groups: As online media usually store the usage conducted via them, these stored data offer insights into users as comprehensively and with as much detail as possible. Conventional methods of determining characteristics of groups have to rely on the representativeness of the sample accessible for indirect methods such as interviewing, testing and observing the persons in the sample group. In the digital age that brought with it the availability of complete records storing users' entries and selections, these methods have to compete with analysis of the complete factual data.

To put this into its proper perspective, one could imagine a new type of TV-set that recorded every program change and sent these records back to the broadcasting station. Doubtlessly, there would be no further need for determining audience ratings by observing or interviewing a sample group if the complete data were available. It is hardly conceivable how any other data based on samples and surveys could do more

[25] Tested in September and October 2006. The queries represent the top searched queries from Yahoo in Germany in the first quarter of 2003. See below, Part A, Term Spaces, for details on this log which is there abbreviated with YG.

[26] Forrester Research / IAB UK, online at www.iabuk.net/images/Overture, search marketing_452.ppt [Nov. 1, 2006].

[27] http://www.golem.de/0604/44623.html

[28] See "Kein Problem mit Bürostülen und Höhrbüchern", ECC, online at www.ecc-handel.de/kein_problem_mit_buerostuelen_und_hoerhrbuechern.php [Nov. 1, 2006]. Studie über die Qualität der Produktsuche in Online-Shops, Pressemitteilung luna-park, Nov., 23, 2005, online at openpr.de/news/69440.html [Nov. 1, 2006].

than refine or add more details to the complete records. Yet this is the situation with E-Commerce and especially E-Commerce-related search, two domains in which the importance of logs are not yet universally seen[29]

It has to be conceded that the cross-relations from user data to the offline background of users have their value. Such relations may (fortunately) only be determined for sample groups of users. From a privacy point of view, it is even problematic to relate queries to individual users disguised by IDs (see the case of the AOL search log below). However, as will be demonstrated in depth in Part C, the accumulated data of searches alone (without any relation between individual users and queries) provides a valuable knowledge repository to which surveys or tests on sample groups might add complementary information, yet which can not be surpassed in terms of providing a full picture of user habits.

In the following section, a short summary of empiric approaches (both surveys and lab experiments) based on user samples is presented, together with opportunities and limitations inherent to these approaches.

Interview based methods

Most of the long-established studies are based on interviews, which themselves are usually based on questionnaires. While open forms of interviewing such as group-discussions are suitable for in-depth exploration of users' knowledge and feelings they are not suitable to be conducted out on a large scale.

In both cases — questionnaires and open interview forms —, however, the unbiased recruitment of participants is a challenge. Using web-based questionnaires allows the recruitment of many participants at a low cost, but at the price of a rather skewed selection of participants. While offline interviews — generally conducted through computer-aided telephone interviews (CATI) — can use socio-demographic data in order to achieve a representative sample of the total population, Web questionnaires such as the W3B studies can only rely on the data self-selected users provide and what they deliberately reveal.

This leads to a further problem of any empiric approaches with study subjects that are aware of being observed — their behavior will be different than when not watched. One aspect of this is called social desirability bias, i.e. a tendency to answer in a way the interviewee expects are evaluated positively in society. However, the differences that spring from the multiple interactions between interviewer and interviewee go beyond values. The narrative drive of humans, i.e. the desire to deliver a coherent story, could also contribute to discrepancies between real behavior and reported behavior. Following this line, interviewees might also feel the need to present themselves as more thoughtful and reflective than they are in reality.

The strength of interviews for user studies lies in connecting online facts (for example

[29] A counter-example is `www.amazon.com` that started to accumulate data from the very beginning and could thereby create cross-selling and recommendation features. See the interview (in German) with Andreas Weigend, Amazon's Chief Scientist until 2004, online at `http://www.weigend.com/WeigendIMPULS2005.pdf [Nov. 1, 2006]`.

what queries have been entered or how they are evaluated on the search site) with offline facts (for example educational background). Open forms of interviews also allow investigators to get close-up pictures of individual users and determine the most important trends and issues concerning user experiences.

Lab experiments

An even more detailed inspection of user behavior can be achieved with laboratory equipment, especially monitoring and surveillance equipment. Video cameras (filming the test person or the test person's screen or both) and/or installed logging software record the sessions that are later on put into a log, allowing the registration of each single activity of the user according to a coding guidebook. In addition, users are instructed to verbalize their thoughts and strategies, for example by being asked to think aloud and comment their actions and thoughts.

An important drawback of lab experiments is their suffering from social desirability bias. However, as they can be designed to facilitate the tested persons to fall back to trained usage patterns, one could expect that users fall back to more natural responses, at least after some time. Lab experiments require costly equipment and manual efforts in operating and evaluating the experiments[30]. This usually limits the sample size in lab experiments which makes the selection of a balanced — in terms of sociodemographic background and Net expertise — set of candidates a demanding task. What lab experiments lack in representativeness of the examined sample, they are supposed to compensate through an in-depth analysis of users' strategies and their reflections. Two rather large German lab studies with 160 and 66 users examined can be found in [Machill/Welp 2003] and [Hoelscher 2002], see also [Hoelscher/Strube 2000] for a similar lab study.

Experimental methods can be combined with interviews in order to give a broader picture of the individual user. They are both helpful before the actual experiment to find out sentiments, usage patterns and experience, but also after the tests to allow the users to reflect on the experiences in the lab.

4.2.3. Users, defined by their trails

The digital environment allows a new approach to user study, given that it made it possible to store every action of a user on a site. In contrast to the artificial surrogate of lab experiments, a query log contains thus the full picture of search. However, it requires some analysis technologies to extract meaningful data out of the sheer mass of recorded searches. While it is possible to get lost in the size of the log, a more severe limitation lies in the property of log information to always represent historic data which validity for the future is not guaranteed. The largest query log presented here span six months (see below), a time-span during which considerable changes (new events, new products or brands) can take place.

[30] See also [Lewandowski 2005], Chapter 2.6.

Another drawback of logs was pointed out by [Hoelscher/Strube 2000]: Since the data is anonymous, one does not know anything about the context of the individual user, namely the kind of information problem she was trying to solve or her level of experience.

This is connected with the concerns that analyzing logs might raise with regards to users' privacy. As many users type in personal data (for example vanity searches), it is sometimes possible to deduce the identity by the search uttered[31]. In this context, it is interesting to note how difficult it is to assess the worth of query logs. When AOL released query log data of about 600.000 individual users in July 2006, public outcry led to a quick withdrawal of these data. Not surprisingly, they have later spread in the digital world From the outset, the AOL data seemed harmless, because users had been disguised by IDs[32]. However, a closer look on the AOL data revealed the possibility of identification for many of the users (in a process resembling a dragnet investigation). Clearly, queries do not only reveal what people want to know, but also their interests, background and previous knowledge.

There is a very rich source of knowledge hidden in the queries accumulated by millions of users[33]. Although a query such as *cheap flight london* would in general be analyzed as questions (even if the question particles are omitted, i.e. as a reduced form of "where can I book cheap flights to london?") or directives ("show me cheap flights to london"), it reveals nevertheless an assumption about state of affairs, here that London is a place one can reach by flying. Aggregated queries present aspects of common world knowledge in a very concise way. Their worth for that purpose is not only their sheer number — where else could one obtain 500 million utterances per day — but also their deliberateness, as most of them are triggered by genuine informational needs. In Part C, different types of queries will be introduced, together with strategies how to detect and resolve them.

Apart from human users, there can be found in every query-log some entries which obviously are the result of automated querying. Queries dispatched by scripts are recognizable in an accumulated log (without information on IP addresses and timestamps) by systematic pattern variation, for example a numeric or alphabetical pass through a list, and a frequency much higher than would be expected given the length and weirdness of the query term (see also below: Term Spaces). This will, of course, not detect all automated querying, but it helps to reduce the noise of non-human queries in a log.

Search engines do in general not welcome that their server resources are used by anyone else than human users (robots and spiders consume bandwidth, but will surely not react to advertisements). They usually limit the number of searches for a given web agent. The APIs provided by Search Engines allow a legitimate scraping of results for the allowed number of queries per day.

[31] If more textual material than in queries is available, identification of users might also work through stylistic anaylsis. See also [Novak/Raghavan/Tomkins 2004].

[32] In the late 90's the Excite log provided the same fields, but it did not stir any public outcry then.

[33] See also [Scholer / Williams 2002] and [Lam/Pennock/Cosley/Lawrence 2003].

4.3. Businesses' websites

In the model developed here, the online presentation of a business should be differentiated from online selling channels. Therefore, only self-presentation and promotion, along with contact data, is considered as a genuine businesses' website. In this context, one has to exclude not only purely private websites (private homepages, fansites, sites by special interest groups), but also the front-end of service or selling channels. For example, the proper business web site for Thomson Directories is therefore their corporate site at thomsondirectories.com, not their local search front-end at thomsonlocal.com. It happens both that one website hosts several businesses (for example dealers of a chain) as well as one business being hosted by several websites, either mirrors or websites dedicated to different lines of businesses.

It is not a trifle task to locate the homepage of a business, except for very large businesses that are well linked-to. Among the websites of businesses are also numerous very small companies, consisting of only one person. Adding to the this, a considerable percentage of websites has not changed for a long time, making it unclear whether the company is still in business. Many websites only list the company name and address, but the website is not operated by it. There are also a considerable number of spam click-through sites that accumulate keywords and company listings without providing any substantial benefit to the user.

These borderline cases make it not only hard to extract the extraction of URLs for business websites, but also pose methodological difficulties on the question of when does a webpage qualify as a company homepage and what special cases need to be considered.

4.3.1. Freshness and availability of businesses on the Web

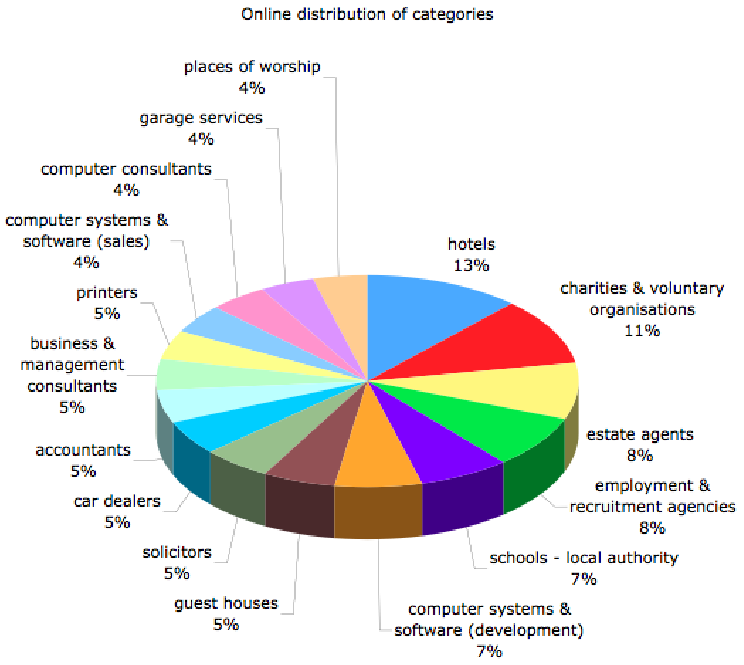
As laid out, business websites come in all different flavors, ranging from template-based one-pagers to huge sites with thousands of pages. In summing up the size by the number of subpages under a given second level domain that belongs to a company, several simplifications have to be acknowledged. Firstly, the company's website might span several second level URLs and secondly, the number of sub-pages might not always be a good indication of the depth of a website, for example if many identical subpages exist or the bulk of the content has no static URL to it.

Apart from the size of a website, it is also how up-to-date the website is what indicates how intensively it is being used and promoted by the business operating it. For the tests below, the *Last-modified* information from the HTTP header was extracted. While again this is not in all cases reliable, as it rather underestimates the number of companies with no fresh website[34]. To validate the information from

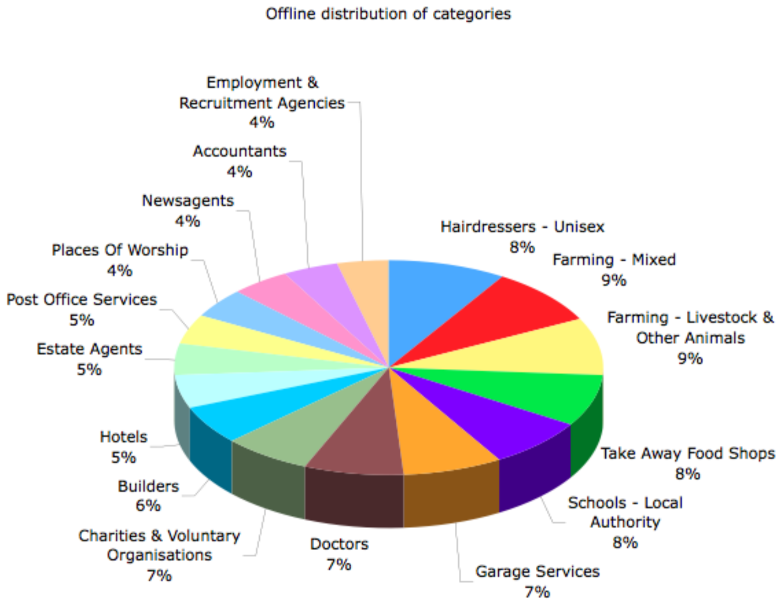
[34] It is more probable that webpages with unchanged content are automatically modified — for example by a date field in the source code — than that webpages with fresh content are delivered with wrong last-modified headers. Moreover, the RFC 2616 specifications do not permit *a Last-Modified date which is later than the server's time of message origination*, see `http://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html [Nov. 1, 2006]`.

the HTTP header, a regular expression was used to find *last modified* elements on the page text. These experiments backed up .

A last aspect discussed at this point is the distribution of businesses into broad industry and branch sectors. On the basis of UK company data provided by Infoserve and enriched with URLs, it was possible to set up two pies, one for the distribution of all companies (with and without URLs) and one for all companies with URL. The difference illustrates which sectors are more advanced in setting up virtual company representations:



Online distribution of categories.



Offline distribution of categories.

Note that hairdressers, farmers and take-away food shops which are very prominent in the offline pie are demoted to lower ranks in the online pie. On the other hand, hotels, charities and estate agents are significantly more prominent online than offline.

4.3.2. Web of German businesses

A recent Eurostat survey pointed out that 72% of German businesses have a website. In the German market there are about 3 million taxable companies[35] and about 700.000 additional contractors. The figure of 3.7 million can also be derived by accumulating all non-resident entries in telephone books. Only about one and a half million, however, have a legal form such as AG or GmbH and can thus be called a company proper.

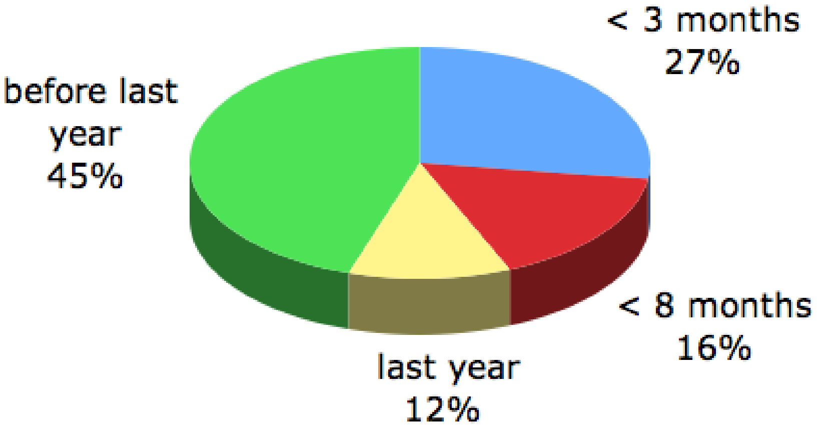
In fact, extensive tests on the German Web conducted at the CIS resulted in about 1 million business home pages retrievable through crawls, also using Search Engines for piggy-backing. The rest are either very weakly linked or only appear in the hidden web, for example pages that are hosted on a yellow page provider. It is hardly conceivable, though, that there are another million business home pages — as could be expected by multiplying 3 million taxable companies with 72%. Even with a lower number of all existing business homepages it is still remarkable how many business homepages are almost impossible to find via search agents and link hubs[36]

The pie of German business sites, derived from a sample of 356.000 company homepages and broken down by last-modified-date (based on server headings) is as follows:

[35] See http://www.destatis.de/basis/d/fist/fist011.php

[36] There are 10 million second level .de-domains registered, see `http://www.denic.de/de/domains/statistiken/index.html`.

Date of last modification	*percentage*
younger than 3 months	27%
older than 3 and younger than 8 months	16%
last year	12%
before last year	45%



Distribution of freshness among DE company homepages

4.3.3. Web of UK businesses

According to a Eurostat's survey conducted in 2005, 74% of UK businesses have a web site. They overtook Germany through a +8% surge in 2004. The largest Yellow Pages directories in the UK list about 2.7 million businesses, including branches of dealer chains.

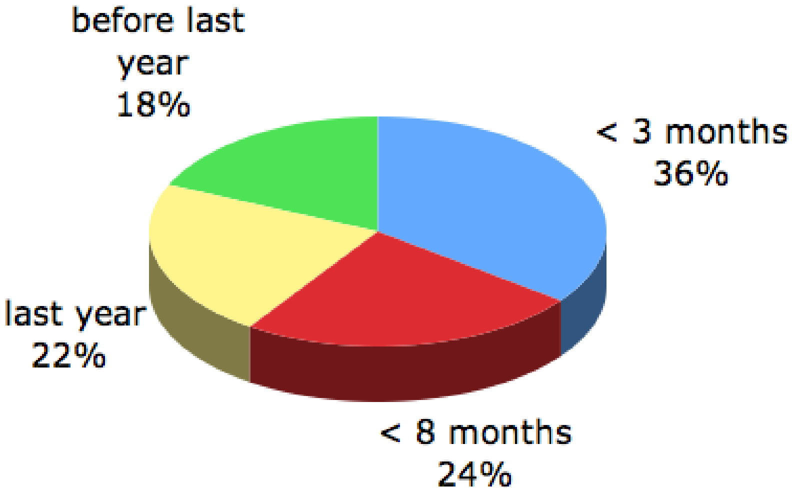
Without factoring out the companies proper from this set, one would expect roughly 2 million business homepages. However, only about 750.000 are prominently linked or appear on Search Engines. Several conclusions can be drawn from this discrepancy. Firstly, the percentage of business homepages is certainly less than 74% of all businesses addresses (=records in a YP directories). Secondly, there is an enormous asset in the location of the remaining homepages that cannot be found on Search Engines.

About 5.3 million .uk second level domains are currently registered[37]. The ratio of registered domains thus roughly corresponds to the ratio of findable business homepages.

The pie of UK business sites, derived from a sample of 194.000 websites and broken down by last-modified-date (based on server headings) is as follows:

[37] See `http://www.denic.de/de/domains/statistiken/ domainvergleich_tlds/index.html` [Nov. 1, 2006].

Date of last modification	*percentage*
younger than 3 months	36%
older than 3 and younger than 8 months	24%
last year	22%
before last year	18%



Distribution of freshness among UK company homepages

4.3.4. Web Shops

Shopping search engines in Germany such as `shop.de` or `preissuchmaschine.de` list several ten thousands of online shops (shop.de claims 45.000 shops and pangora over shops). This number represents a very broad range of different companies, from one-man, part-time outlets to multi-million stores such as `pearl.de` or `conrad.de`. Typically, a web shop specializes in one branch, for example beauty products or electronics. Some branches are highly regulated in Germany with pharmacies being a prominent example. Shops have to rely on a functioning shopping infrastructure, i.e. procedures on the back-end and front-end that allow selecting goods, purchasing and paying for them without hassles. Smaller shops especially use generic shopping software to fulfill this task. Providing scaleable and trustworthy billing with multiple billing methods (credit card, wire transfer or cash on delivery) is by no means a trifling task. In many cases, the web space provider also supplies out-of-the-box shopping systems. However, a running infrastructure is not enough. Shop sites have to provide easy, intuitive and attractive access to their goods.

While the transactional-related modules of most online shops looks very similar (usually centered on the shopping cart metaphore), the question of how to arrange products on virtual shelves is answered with very varying success. A common layout for online shops is a hierarchical categorization of products a presentation of goods via result and record pages, a shopping cart link and a check-out link, often placed at the top on the right hand side[38]. Apparently, there is still ample room for improvements on the usability of shopping sites, though. A study reported more than 40% of purchase interactions being aborted because customers experience technical difficulties or do not know what do next at one stage of the transaction process[39].

Another field of possible improvements relates to guided navigation and cross-selling functionality. Guided navigation tries to support users in finding what they have looked for on online stores while at the same time preferring to present those items that are known to generate high revenues (much like human shopping guides)[40]. Cross-selling is basically a kind of very targeted advertising. Once a product has been purchased cross-selling tries to sell products related to it — for example, if a customer bought a notebook pointing to a notebook bag that is also in store.

According to the same Novomind study reported above, both guiding the user through the offers or exploiting cross-selling are not widespread on typical German online shops. Moreover, searching functionality is in most cases very basic at best, punishing many common variations with disparate or zero results (see also Part D for a test battery). As platforms bringing products and customers together, online shops face many of the challenges with Search Engines and need term handling to a similar extent. It is not sufficient for a shopping site to receive traffic, it has also ensure that it can be converted. One component of achieving high conversion rates is to provide an easy access to offers which is at the core a term matching task, either as matching user searches to product descriptors or deploying navigational elements that match what users look for (see also below Part A, TE-Commerce Interactions).

4.4. Content Providers

While customers and businesses belong to any market place whether offline or online, it is also necessary to more specific Web agents into account. Web content providers have to be examined in order to give a full picture. They appear in the model as companies if they offer paid content. In this function, they provide a service not fundamentally different from other downloadable services, such as paid music download. However, content sites are also places attracting users and therefore offer valuable advertisement spaces. In their function as entities gathering users, they fit into the model as a manifestation of users. Related to this view is the old promise of the Web as a truly participatory media where the difference between those producing and those consuming texts is bridged [41]. Recently, the notion of personal content and

[38] Amazon.com once conducted a test with presenting half of their customers a layout version that had the check-out link on the left margin and the other half with a version that had it on the right margin. The latter version produced significantly more revenues (speculatively, because the preferred reading direction expects an exit point at the right hand side?). See Andreas Weigend, IMPULS interview 2005. This is again an example how data accumulation on the Net can be used to find out about user habits in a way that surpasses ordinary surveys in reliability and efficiency.

[39] http://www.novomind.com/index_ht.html?press/2005/rel_77.html

[40] Two of the most widespread solutions for guided navigation in the context of shopping sites are Celebros (www.celebros.com) and Endeca (www.endeca.com)

[41] Roland Barthes' idea of an open, "writable" text resurfaced in the wake of the Web.

users' communities has gained attraction also from the point of view of commercial exploitation[42]. In summary, content sites on the Web can fulfill several functions at once. When offering paid content they act as companies dealing with virtual but nonetheless bona fide goods. Steering user traffic, they appear as one manifestation of users. In both respects, content sites are related closely to entities of the 1^{st} level and can be numbered among them.

One special issue related to content providers needs to be discussed in more detail. The point of entry on the content providers site was subject to legal controversy that was led under the name *deep link*. Deep links refer to accessing a content page without passing through the main page of its site and without using the navigational aids supplied on the main page. Many content site operators stated that they would like to see users starting at their homepage and combatted the use of deep linking that allows to go directly to a content page. One line of argumentation that persistently surfaced in the discussion goes that revenues for the content page operator decrease if users bypass the advertisements on the front page. A principal court decisions in Germany (BGH 2003) has only recently ruled out the possibility of disallowing deep links. A few years back, web services such as Paperboy and Paperazzi have suffered severely or even broken down because content providers would not allow deep linking and started legal wars against it.

While at that time — when static banner ads had been the substantial income colon of content providers — such action might have made sense, as any fraying of traffic had direct financial impact, this preoccupation has largely been made obsolete by new advertising models. The controversy has nowadays moved on, for example to the question whether newspaper content is allowed to show up on an aggregate view (like on news.google.com). With regards to the TE-Commerce model that is developed in this and the following chapter, such controversies highlight how tightly interconnected searching, advertising and providing content really are.

One vast source of content that is in general not yet accessible through Search Engines lies in a rather classic form of social media, around long before the advent of Web 2.0. Guest-books and forums fall into this group. However, through restrictive `robots.txt`, mandatory log-in or non persistent URLs, a lot of valuable information they contain gets lost. While for some topics, there is undoubtedly a tendency of participants to protect their privacy and keep among themselves (for example parents of physically challenged children), there should be a way of making the content searchable without affecting the privacy of the contributors. One possible way could be caching the content with disguised identities, such as in Google's archive of newsgroups.

[42] Referring to the purchase of the Web photo community Flickr, Yahoo's head Terry Semel called so-called social media a "gigantic piece" of the company's strategy, see [Schonfeld 2005]. There are numerous other examples of so-called social media sites - such as blogs, community portals, special interest sites - being bought by large Web companies. See below, Web 2.0 and Convergence Phenomena

4.5. Making matches: the Search Agents

The main property of second level of agents is to bring first level agents together. The primary task of search agents as one of the 2^{nd} level agents is to couple users and content: Users can retrieve webpages based on the queries they enter or a selection of catalog links, and content starts to become accessible on the Web if it can be found through Search Agents. Starting from generic Web Search Engines such as Google, Yahoo or MSN, different specialized Search Agents are examined subsequently.

4.5.1. Generic Web Search Engines

The number of self-managed generic Web Search Engines that build up their own index by crawling billions of Web pages became very limited. Maintaining a full-blown web index requires enormous hardware equipment (rumors have Google possessing about 450.000 single servers[43] leading to investment expenses and running costs that are in general not affordable for a new player[44]. Although all major Web Search Engines can be used without charge as of today, they clearly offer a service that has economic worth both for users and for content providers, even excluding Search Engine Marketing that forms the main source of revenue for Search Engines today.

In the oligopoly that characterizes the world of Search Engines today, it is conceivable that search services might be charged at some point once a viable way of micropayment is found. Providing search services is doubtlessly a value-creating activity (and requires substantial funding, if only for hardware and bandwidth[45]. Indications to the potential monetizing capabilities of search services (again, not to be confused with online marketing) are for example the rising number of search equity investment deals that increased from eight in 2001 to 27 in 2004 and 31 one year later[46], resulting in about 150 million EUR worth of venture capital invested in new Search Engines, mostly however for niche searches.

To illustrate the stage of the Search market between monopoly and oligopoly, here are figures both for Germany and the US referring to the number of searches conducted on different Search Engines:

Figures for the top Search Engines in the German market[47]:

[43] See [Markoff/Hansell 2006].

[44] To illustrate the costs connected with creating a full-blown Search Engine: The development new European Search Engine quareo is backed with 400 million EUR funds

[45] The open-source Search Engine Yacy tries to rally the necessary resources through distributing the crawling and indexing process to its users

[46] Source: VentureOne and Ernst & Young. See `http://news.com.com/Can+there+be+another+Google/2100-1025_3-5983371.html [Nov. 1, 2006]`

[47] http://www.webhits.de/deutsch/index.shtml?/deutsch/webstats.html, based on an analysis of 41.000 searches.

Search Engine	*Percent of all searches in Germany*
Google	85.9%
Yahoo!	3.4%
MSN	3.4%
AOL	1.9%
T-Online	1.3%
Lycos	0.6%

Figures for the top Search Engines in the US market[48]:

Search Engine	*Percent of all searches in the US*
Google	44.1%
Yahoo!	27.9%
MSN	12.9%
AOL	6.7%
Ask	5.3%

These searches mainly originate from the site of the Search Engine itself, i.e. from users visiting `google.com` or `yahoo.com` and enter their query. While the broad majority of searches is done on the site of the Search Engine itself, about 12% originate from browser toolbars[49].

In Germany, as in other European countries, the market is characterized by a Google monopoly — the figures of Google usage are sometimes reported even higher than 90%. In the US, an oligopoly among three major players — Google, Yahoo, MSN — can be observed, given that AOL uses Google results as the basis of their search service.

4.5.2. Vertical and Topical Search Engines

While today there are only a handful of Web Search Engines left that keep their their own index and have a considerable reach, numerous specialized search providers have emerged since the early days of the Web[50]. Specializing can either be done by restricting the search to a specific business, topic or locality. As the distinctions between vertical and topical Search Engines are rather blurred (following a common interpretation of the terms, vertical searches are addressing b2b and topical searches b2c), they will be treated as one.

As a search agent for one domain[51], vertical Search Engines are fed by vertical spidering or businesses submitting their data. These options are not mutually exclusive — a common practice is to include paid listings on top of the results from the vertical spidering.

[48] http://www.comscore.com/press/release.asp?press=914

[49] http://www.seroundtable.com/archives/003381.html

[50] For example, a company called Accommodation Search Network ran hotels.com as early as in 1999; see `http://web.archive.org/web/19991127074842/http://www.hotels.com/index.html [Nov. 1, 2006]`

[51] See below, Part B, Lexical Units, for an explanation of the domain concept

While creating an index through vertical spidering requires the extraction of features from web pages — either through a general learning algorithm or fine-tuned grammars —, the latter approach can resort to a detailed form presented to businesses when submitting their data. This self-description process is not always fully reliable, though, either because businesses do not meticulously fill out these forms or because they see an advantage in not truthfully reporting these details.

On the level of query interfaces, today's vertical search agents often use complex search interfaces with many different search parameters, specific to the domain they cover (in the hotel domain for example hotel ratings, size of rooms, smoker/non-smoker etc). The processing of free-text queries with regards to the c analysis

Query-triggered special searches appear also on the site of general Web Search Engines, for example with Yahoo Channels. Regardless of the URL of a vertical search agent, it requires both a specialized indexing and a specialized query interface to be ranked as vertical or topical search.

In general, a vertical Search Engine is expected to provide both a broader coverage of the domain and especially more fine-grained results than a general web search can offer. Another effect that helps to boost performance of specialized search engines is a direct result of term handling: Many ambiguous queries are quite distinct in one domain. For instance, searching for *speakers* on a home audio search restricts the two possible senses (human profession and audio equipment) to one.

4.5.3. Local Search

In addition to domain-related specializations, local searches focus on specific places. A fundamental difference between the two specializations, however, lies in the fact that every topic requires a new topical search, whereas local search devices are in general supposed to cover all localities[52].

Local search on the web emerged from different backgrounds: Directory Assistance (for example Varetis/goyellow.de in Germany), Yellow Page providers (gelbeseiten.de) or web portals (web.de). Although the market in Germany is still dominated by Gelbe Seiten — jointly issued by DeTeMedien and 16 regional telephone book publishers — with a share in local searches reported at about 95%[53], many businesses are pressing into the field, as it is expected to grow considerably in the next years. A prediction of the size of the German market for local search in 2009 runs at 298 million EUR[54].

Convergence phenomena between web search and yellow pages are already visible, both from the direction of Web search going into yellow pages (such as web.de) and

[52] The knowledge of a specific locality lies in general within the sales force of Yellow Pages providers. Social generated content has recently been considered as a possible alternative, see for example `www.kiji.de` and `www.qype.com`.

[53] This figure was cited by Goyellow's lawyers when successfully challenging the trademark on Gelbe Seiten. See `http://openpr.de/news/95638/ direct-Beiten-Burkhardt-Der-Fall-der-Monopolmarken-Gelbe-Seiten- und-Yellow-Pages-geloescht.html [Nov. 1, 2006]`.

[54] According to Hanns Kronenberg of `muenchen.de`, based on a study by Kelsey Group. See also `http://www.silicon.de/enid/b2b/13141 [Nov. 1, 2006]`.

from yellow page into Web search (herold.at). This results in a mix of Yellow pages directories and Web search companies as the leading German local search providers:

- Gelbeseiten.de
- AllesKlar (meinestadt.de)
- web.de
- Goyellow.de
- T-Info (suchen.de)

Local search on the Web allows to establish a pay-per-performance billing model that lets the advertiser pay for either impressions, clicks, phone calls (through phone number tracking, see below) or other actions such as downloading the address into Outlook or sending the address via E-mail. In this respect, but even more so through its capability of delivering targeted results, local search on the Web goes way beyond conventional printed Yellow Pages.

Today, one of the main assets in setting up a local search are commonly considered to be address and map or satellite image data. As it is not feasible to accumulate again millions of addresses (that in addition have to be up-to-date), usually a local search provider has to buy addresses if they do not already possess them. Today's in-vogue satellite images, apart from providing a nice visual effect, might contribute to the perceived relevance of results as they corroborate the physical existence of the business. In this, they would fulfill a part of the function of listing a business homepage. Yet allowing the user to inspect the homepage of a business on the local search site goes beyond confirming the existence of the business — it also allows the user to get a first impression of the business and its products and services.

4.6. Search Marketing

Search marketing (SEM) is a powerful way of promoting products using the channels search engines offer. The term comprises different methods of promotion, just as search engines have evolved into a variety of services, including paid inclusion, paid placement and sponsored links. In the following, the focus will be on the latter type, not only as it is the fastest growing, but also as the two other options have lost most of their relevance. Paid placement without indicating the status as an advertisement breaches German regulatory laws. Paid inclusion is legal, but no major generic Web Search Engine can afford to build up an index solely with paying customers[55], making paid inclusion only attractive for website operators because of the speed of inclusion.

This leads to to sponsored links — usually based on a pay-per-performance base — as the dominant Search Marketing device. Sponsored links can appear either on

[55] GoTo, the company that became then Overture, started in the second half of the Nineties with exactly this model.

the site of the Search Engine or on content sites. Both cases will be examined in the following two sections.

4.6.1. Keyword triggered advertising

In keyword triggered advertising, businesses book keywords for which they want their ads to appear. If a user types in a query that can be related to this keyword, the ad booked for it appears. A click on the ad is charged from the advertiser by the Search Engine Marketing company and shared between the SEM and the content provider on which the ad appeared. Ads are ranked according to various mechanisms including what the advertiser is willing to pay for a click. As the ads only appear for users who have typed in specific keywords, they have far higher click rates than static banner ads. Fine-grained campaigns might reach click rates well at 5%, i.e. every twentieth time the keyword was entered it is also clicked[56].

SEM has turned out to be a dynamically growing market: in the time between 2004 and 2005 its volume doubled in Germany from 110 million EUR to 220 million[57]. The prediction for 2008 run at 300 million EUR[58]. However, there are some signs of consolidation and concentration of processes. In 2004, three major events shaped the current global market situation:

- Yahoo purchases Overture in January 04
- A merger between FindWhat and Espotting creates Miva in February 04
- MSN terminates the contract with Looksmart and starts its own pay-per-click program in June 2006

With the three major SEMs that are now identical to the major Web Search Engines, their market shares in paid search follow largely the market share in Search. ComScore Networks reported at the end of 2004 a share in the US paid search market of 35% for Google, 32% for Yahoo and 16% MSN[59]

Considering the European market, it are also Miva, QualiGo and Mirago that deserve to be mentioned. It has to be noted, however, that these do not deliver ads to major Search Engines. Furthermore, there are numerous companies that organize campaigns for advertisers such as adpepper.com or advertisers.com. In certain niches smaller players are still active, for example Kanoodle with its contextual advertising program especially on financial-related sites[60].

One of the most crucial issues in SEM is click fraud — malicious clicks on ads that have no worth for advertisers. Estimations of the extend of click fraud rank as high as

[56] See [?].

[57] Explido Web Marketing, see `www.explido-webmarketing.de /pdf/SPIXX_Jahresrueckblick _2005.pdf [Nov. 1, 2006].`

[58] According to Forrester Research

[59] See `http://www.alwayson-network.com/ comments.php?id=9211_0_6_0_C. [Nov. 1, 2006].`

[60] Also on MSN spaces as of October 2006.

10-15% of all clicks[61]. Although all major SEMs use technologies to detect click fraud, both sophisticated auto-click software (often using hijacked computers) and low-paid manual clicking threaten the confidence in keyword advertising. As a technological device can only partly be the remedy — ultimately, distinguishing between bona fide clicks and malicious clicks would require a brain scan of the user —, the pressure into developing new revenue models going beyond pay-per-click grows. One solution lies in pay-per-lead, for example based on successful transactions such as purchasing an item on a web shop or downloading a software. Miva introduced an interesting extension of this concept into the offline world. By displaying a special call-through phone number to users on the website (either Miva's or that of a content affilate), it is possible to charge advertisers on a pay-per-call basis, as all calls on this number are due to the advertisement on Miva's network. Despite click frauds and legal issues such as using brands as keywords, keyword advertising is since the start of MSN's AdCenter provided by all three major Web Search Engines.

It is worthwhile to take a look on how Microsoft as a newcomer in the SEM business tries to rival the established keyword advertising schemes of Google and Yahoo. Microsoft tries to take targeting even further by delivering ads only on specific times and only to user groups defined by age, gender or ZIP code. Given that targeting can be controlled by monitoring click and conversion rates, even a less than 100 % accuracy in demographic targeting is acceptable, because it will either have a positive effect, non effect at all, or can be turned off soon if it shows to have a detrimental effect[62]. Moreover, a boost in click rates leads immediately to a corresponding boost in revenues for the SEM — 10% more clicks translate into a 10% increase in revenues, assuming that advertisers are able and willing to pay for the increased traffic. Microsoft intends to send ads also to mobile devices and Xboxes. It

The two smaller competitors in the German market, Miva and Qualigo do not deliver their ads to widely used Search Engines, but focus on content sites (especially Miva with affiliates such as `zeit.de, falk.de, freenet.de, sat1.de, kabel1.de`, `markt.de)` or smaller Search Engines (especially QualiGo, delivering ads to `blitzsuche.rp-online.de`, `tricus.de`, the German search on `ixquick.com, websuche.de` or `alluna.de`).

4.6.2. Content triggered advertising

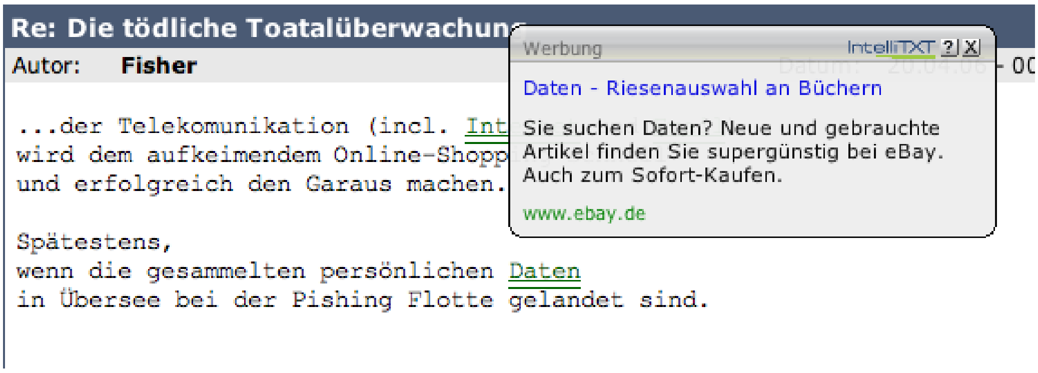
The flexibility and promptness of the virtual market allows delivering a new form of advertising on a large scale. Content-triggered advertising refers to the dispatching of ads to pages based on their content. While targeting has always been a key concept in advertising, the automatic delivery of ads based on content properties where it will be displayed, is a relatively young concept in E-Commerce. The opportunity to drive traffic that is qualified on a very granular level (consider the example of

[61] http://www.businessweek.com/magazine/content/06_40/b4003001.htm

[62] Microsoft spokesperson Winfield said that Microsoft's demographic "abilities are far from perfect, but even information that is 25 percent accurate is useful." See `www.zdnet.co.uk/misc/print/0,1000000169,39237026 -39001068c,00.htm [Nov. 1, 2006]`.

advertisements of Turkey hotels on the margin of a reportage on the Antalyan coast) makes this concept extremely promising. Albeit its promises, it is also one that has to face tremendous difficulties regarding how to capture what makes content a good and target-specific context for ads. This especially applies to those contexts that make the ad display seem odd, out-of-place or even embarrassing.

There are plenty of opportunities to ruin the effect of an advertisement by the context in which it appears. Often cited examples include promoting airlines on articles on plane crashes or scuba lessons on articles on shark attacks. Yet even if topics of death, terrorism, catastrophes, havaries etc. are recognized and discarded, there is still ample room for advertisements looking strange. While it is possible to measure the success of contextual ad campaigns by the ratio of clicks to page impressions, it is not easy to deduct what the reasons for failure could have been and how one should learn from them.



Re: Die tödliche Toatalüberwachun

Autor: Fisher

Datum: 20.04.06 - 00

...der Telekomunikation (incl. Int
wird dem aufkeimendem Online-Shopp
und erfolgreich den Garaus machen.

Spätestens,
wenn die gesammelten persönlichen Daten
in Übersee bei der Pishing Flotte gelandet sind.

Daten - Riesenauswahl an Büchern

Sie suchen Daten? Neue und gebrauchte Artikel finden Sie supergünstig bei eBay. Auch zum Sofort-Kaufen.

www.ebay.de

A screenshot of contextual advertising for data, found on http://forum.golem.de/read.php?10270,608461,609082#msg-609082, taken on 09/15/06.

In the case depicted above, the effect of contextual advertising is harmed in even three ways: not only is the context negatively connotated (the commentator expresses fears of phishers getting hold of his personal data), but the ad itself is also one of exemplary irrelevance. Why should someone wants to buy *daten* (data) on eBay? - and just to top it off, clicking on the ad leads to hundreds of thousands of results, just because *daten* appears in a common disclaimer on auctions about the protection of the buyer's customer data.

It seems hardly convincing that the shallow features usually selected for Machine Learning will be sufficient to build a stable and robust prognosis model for the success of contextual advertising that works largely on an emotive and sentimental sphere. Likewise, trying to locate the content of a text into a taxonomy in order to pull out appropriate ads while simultaneously detect negative contexts requires sophisticated matching functionality is still a major challenge as can be seen by the many examples of out-of-place contextual advertising.

In a similar way how SEMs and the major Search Engines entered tight relations and in many cases merged, also contextual advertising providers have been attractive partners for Search Engines. Quigo, one of the first providers of contextual advertising, became integrated into Yahoo and Google purchased both contextual advertising technology provider Applied Semantics. Such moves are not surprising of one considers the total market volume in contextual advertising which is estimated at 1.4 billion USD by 2008, growing at a predicted average rate of 84% in 2004[63].

To provide relevant advertising on a content page, more than just the on-page content can be and should be taken into account. The quality of traffic is not only determined by the content on which it lands, but also by the selections users made in order to get to this page. As will be elaborated in the next chapter, access by search and access by browsing are the two dominant models for reaching content. Both models translate into contextual advertising strategies. If a content page is accessed through a search, than the selection of which advertisement to show could be made through the query. For access by browsing, contextual advertising mechanisms could mine internal anchor texts, as these represent the selections the user made to get to the content page.

4.7. TE-Commerce Technology providers

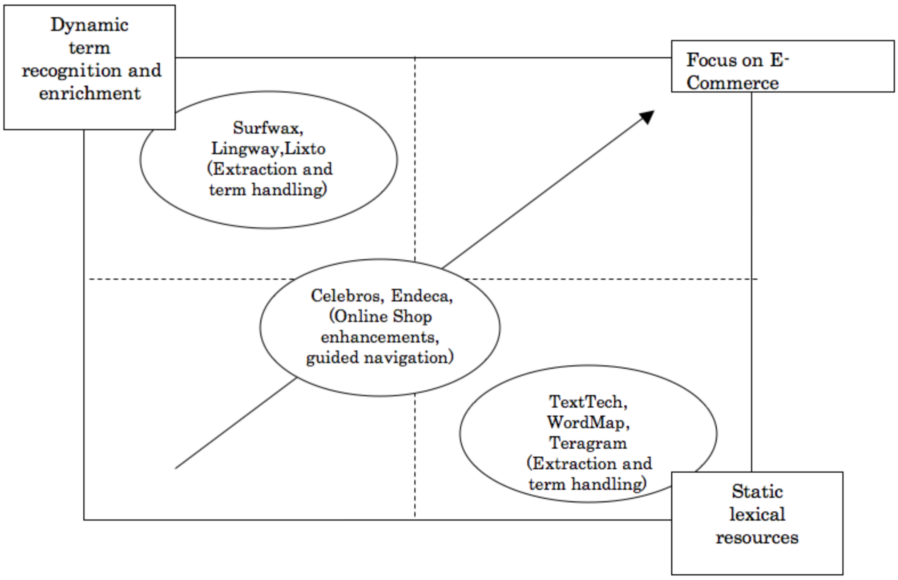
TE-Commerce technology providers are 3^{rd} level agents facilitate the interactions of lower level agents by means of technical artifact. The providing of a system infrastructure such as a billing system or a shopping cart lies outside the framework of TE-Commerce, just as do Internet Service providers or router manufacturers.

Genuine TE-Commerce technology providers enhance the handling of terms for other agents. These technology providers cluster in several groups based upon at which point of the TE-Commerce model their contributions take place and how these contributions are integrated. One group specializes in setting up lexical resources, especially ontologies, that are integrated at the site of the Search provider. Such resources are language-specific and in general need enormous amounts of working hours to set up[64]. In contrast to this first group, a second group of technology providers focuses on the dynamic processing of TE-Commerce interactions. Examples of what the second group may provide are algorithms for orthographic approximate matching, automatic information extraction from Web pages, semantic matching, auto-classification etc. A third group specifically addresses online shops and aims at increasing the shop operators's revenues, for example by guided navigation, product recommendations or conversion tracking.

In the scheme below, the term-related technology providers are grouped into these three clusters as described above with few sample company representing the group.

[63] Jupiter Research 2003. See blog.zdnet.com

[64] David Crystal reports a 8 million USD investment in setting up the lexical resources for Crystal Semantics. See `www.crystalsemantics.com`.


Dynamic term recognition and enrichment
Focus on E-Commerce
Surfwax, Lingway,Lixto (Extraction and term handling)
Celebros, Endeca, (Online Shop enhancements, guided navigation)
TextTech, WordMap, Teragram (Extraction and term handling)
Static lexical resources

Many TE-Commerce technology providers above have evolved from university research or are supported by academic researchers. In recent years, a considerable number of TE-Commerce technology providers have been acquired by larger technology providers or 2^{nd} level agents. Some examples of the past:

- Ontology specialist Synapse has been acquired by Convera
- April 2003: Applied Semantics, a specialist in context-sensitive advertisement, has been acquired by Google
- October 2003: Sprinks with its has been acquired by Google
- Lexicography specialists Crystal Semantics has been acquired by Ad Pepper.
- April 2006: Google acquires the Orion search algorithm that provides contextually related searches[65]

4.8. Web 2.0 and Convergence Phenomena

While a lot of media coverage given to the so-called Web 2.0 rather attributes to a hype than a true paradigm shift, there are undoubtedly genuine thriving examples of social and participative content. Restricting the meaning of Web 2.0 to its core aspect of user generated content — thus disregarding other aspects commonly subsumed unter Web 2.0 such as the blending of applications into websites, the mash-ups, the proposed new lightweight business models etc. — makes integration into the TE-Commerce model

[65] http://blogs.zdnet.com/Google/index.php?p=157 [Nov. 1, 2006]

developed above easy[66]. With the advent of Web 2.0, users and content providers become closer, sometimes even indistinguishable from each other. Yet, as will be seen below the blending between users and content providers is not the only converging phenomena that Web 2.0 brings forth, especially with regards to E-Commerce. It also affects the interplay between searching and content. User-generated content is not viable without advanced access, organization and search functions.

4.8.1. Web 2.0 and E-Commerce

The most striking observation regarding the relationship between Web 2.0 and E-Commerce is arguably how much effort big E-Commerce players such as Amazon, eBay, Yahoo or Google put into user generated content. After the 1998's acquisition of imdb.com by Amazon[67], there is much activity in the last years related to sites bearing considerable social content, such as flickr.com, bought by Yahoo[68], or myspace.com, bought by News Corp. This did not only demonstrate that user generated content is seen a a valuable asset, but also fueled the hopes in a new Web boom. One reason behind this interest certainly lies in the impressive page impression numbers of the mentioned websites[69]. In addition, the shift towards participative forms of content seem inevitably, thus making it attractive to join into it. Many large companies now allow participative content on their websites, especially blogs from employees or management personal. Put in a very general formula, the virtual market allows exchange of information and sentiments among many more participants than was possible offline, plus that these exchanges can be archived and inspected long after they have taken place.

Another crucial shift in Web 2.0, as pointed out by Berners-Lee, is a new form of refinancing content that is centered on the concept of the tail. While the 90s saw mainly static ads on websites with high visit numbers as a refinancing model, the targeted delivery of advertisements to hundreds of thousand websites allows a monetization of the long tail of websites with middle and low visit numbers. As Berners-Lee put it into a lesson that is according to him part of the Web 2.0 concept: “leverage customer-self service and algorithmic data management to reach out to the entire web, to the edges and not just the center, to the long tail and not just the head”. Following this line of thought, the monetization potential of a website no longer depends only on the total number of visits, but also on the quality of traffic it generates.

[66] While the name Web 2.0 brought up all different kinds of definitions, the explanations by Tim Berners Lee, online at `http://www.oreillynet.com/pub/ a/oreilly/tim/news/2005/09/30/what-is-web-20.html [Nov. 1, 2006]`, have gained some authority as a fundament to Web 2.0

[67] See `http://www.prnewswire.co.uk/cgi /news/release? id=37602 [Nov. 1, 2006]`.

[68] Aptly, this acquisition was first announced on Flickr's own blog, http://blog.flickr.com.

[69] Myspace.com ranks currently [October 2006] as number 2 website worldwide according to `alexa.com`.

4.8.2. Importance of Term Handling in Web 2.0

In Web 2.0 with its open access to create and modify webpages, edit rules are impossible to enforce. User-generated content is usually not curated by professionals and in addition often written in an ad-hoc manner, with many typographical errors and usage of non-standard vocabulary. These changes affecting the material that needs to be indexed has to be taken into account in the search process.

Tagging as a substitute to full text search suffers from the same problems (ad-hoc setting of tags, typographical errors, non-standard vocabulary), yet in an even aggravated manner. Tags are usually single words and therefore in many cases highly varying and ambiguous[70]. As Adam Mathes illustrated in the case of searching for the tag *filtering* on `del.icio.us`, completely different meanings have been subsumed under the tag. Below is the list of webpage abstracts that had resulted from searching for *filtering*[71]:

- Last.FM - Your personal music network - Personalized online radio station
- InfoWorld: Collaborative knowledge gardening
- Wired 12.10: The Long Tail
- "Oh My God It Burns!" Practical Applications of the Philosophers stone. For drunks. Brita filter makes bad vodka into good vodka
- Introduction to Bayesian Filtering

Tags are not only ambiguous in many cases, but also exhibit orthographic, morphological and semantic variation. Searching for *accomodation* and *accommodation* or *video recorder* and *vcr* yields disparate results on current large user-generated content sites with a tagging system such as flickr.com, del.icio.us or youtube.com. As the distribution of tags follows the classic Zipf curve, there are many rarely used tags which are often just variations of commonly used tags. Tapping into the content labeled by rare tags, for example for search refinement or drill-down, would provide a much richer retrieving experience to the user.

The need for normalizing and cleaning tags is even higher if tags should be used to build up any kind of ontology. Letting users set up a shared ontology by themselves has considerable advantages, especially considering that professionally curated ontologies are costly and often deviate from what people expect it to look like. However, if a user-generated ontology is abundant in redundancy and inconsistencies, its contributions will be of very limited worth[72]

While Web 2.0 is often seen in connection with the Semantic Web, it certainly deviates from the Semantic Web's original concept of applying semantics to content. In

[70] David Crystal points out nice examples such as *depression* — era, geographical formation or psychic state.

[71] http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

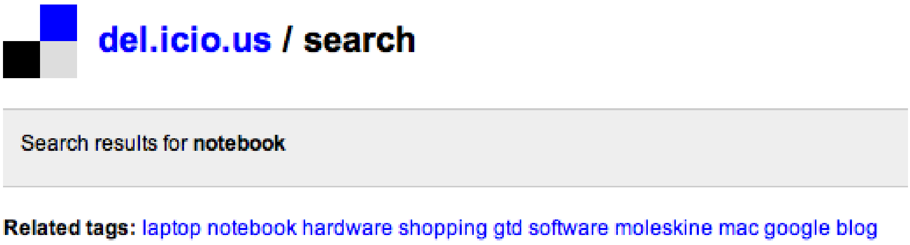
[72] If multiple users generate content or ontologies, adding nodes, especially in popular areas, will happen quickly, but systematic modifications and balance do not follow automatically.

contrast to the Semantic Web is hardly separable from ontologies and inference mechanisms[73], Web 2.0 builds upon collaborative classifications, often called folksonomies. Well-known examples of folksonomies are social bookmarking sites such as del.icio.us or the category system on Wikipedia (see above).

The main advantage of tagging in comparison to a priori ontologies is often seen in the simpleness, robustness and its origin in actual usage rather than in a rigid normative ontology. There are even voices that see the lack of variance recognition as an innate strength to folksonomies:

> Aside: I think the lack of hierarchy, synonym control and semantic precision are precisely why it works. Free typing loose associations is just a lot easier than making a decision about the degree of match to a pre-defined category (especially hierarchical ones). Its like 90% of the value of a proper taxonomy but 10 times simpler." [74].

However, the lack of explicit relationships often lets folksonomies contain nothing more than loosely associated single word. For example, searching *notebook* on del.icio.us results in the following related tags:



del.icio.us / search

Search results for **notebook**

Related tags: laptop notebook hardware shopping gtd software moleskine mac google blog

Among the related tags presented here is a synonym to *notebook* (*laptop*) and its singular variant, but also very general tags such as *shopping*, *software* or *blog*. Sometimes even "junk" tags such as *laquo* (obviously originated in the HTML entity) are produced (searching after *northwest airlines* on Oct. 1, 2006). The quality of the related tags thus differs greatly. A variance handling routine could work without any intrusion, for example by just grouping equivalent tags without forcing any control on the usage of tag variants.

Tagging is often considered to be an approximation of how human minds store and access knowledge. It is often just a small piece of information that is needed to retrieve a memory that is inaccessible via a systematic search. For example, if one is looking for a translation of a term into a foreign language once learned (say, the English word for *Strassenbahn*), the key to retrieving the translation is usually either a phonetic chunks, but might also be a related term (even if weirdly related on the basis of associations or personal memories, in the case of *streetcar* maybe a scene with Elizabeth Taylor

[73] See for example [Davies/Fensel/van Harmelen 2003].

[74] Stewart Butterfield's Sylloge blog, see `http://www.sylloge.com/personal/2004/08/folksonomy-social-classification-great.html` [Nov. 1, 2006].

of the movie, “A streetcar named desire”). The possibility that the mind starts the retrieving process by scanning an ontology, starting with “entities”, moving down to “concrete objects”, to “means of transport”, to “public transportation” and so on is hardly convincing. An empiric indication of that lies in the low usage frequency of many highest-level ontology labels, such as “entity”, “concrete objects” or “means of transport”, especially compared to concepts of medium granularity, such as “car” or “train”.

Equally, a literal string matching between single-word tags is not a feasible model how the mental tagging process could work. This renders the current tag search applications a very crude approximation of how humans would organize knowledge and memories[75].

4.8.3. Search and Content converges

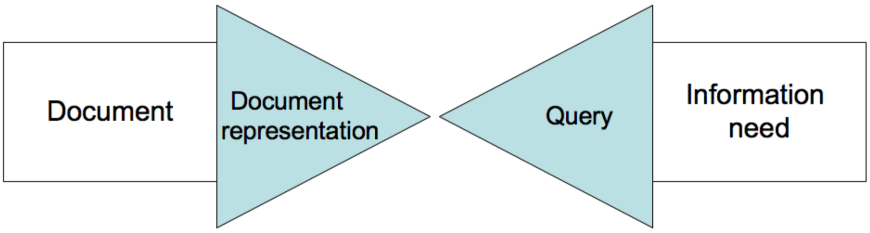
An obvious key difference between the human mind’s retrieving techniques and ordinary search interaction cycles is the separation of a search phase and a result inspection phase. Users on conventional search engines type in their query, submit it and then enter a result inspection phase that may leave them content with what they found or lead them to reformulating the search. The two phases are divided by the *submit* of the user’s query — even if the results pop up immediately after the submitting of the query, the two phases remain separated, and the query has to be fully typed in before its success or failure can be observed.

With recently evolved dynamic Web programming techniques (exemplified by AJAX), immediate inspection of result spaces becomes possible. Such an inspection routing may go beyond conventional suggest modes that simply browse through a pre-calculated list[76]. With recently available indexing technologies it is feasible to perform searches — even approximately and aware of semantic variations — while the user types (see above).

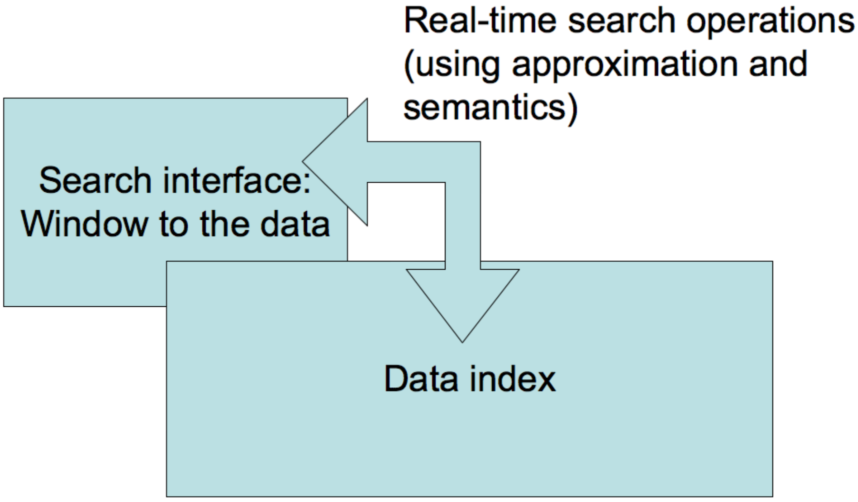
Performing the search operation and displaying results at the time the user enters the search string will in all probability have a tremendous effect on search behavior. If the system conveys to the user that the full set of possible results can be inspected while typing, searches will be redeemed from the trial-and-error cycle. This does not only hold for queries for which the user already has a clear picture of results, but also for open search queries. The dynamic inspection of the result space may also lead the user to related items not thought of when formulating the search. The two diagrams below try to set apart this new model of dynamic inspection from the classic IR model:

[75] Prefound.com announced a tag search that incorporates tag variations. See `http://home.businesswire.com/portal/site/google/index.jsp?ndmViewId=news_view &newsId=20060918005671 [Nov. 1, 2006]`.

[76] As on suggest.google.com which is a letter-per-letter browse through a list of common queries



The classic IR model


Real-time search operations (using approximation and semantics)
Search interface: Window to the data
Data index

Search and content converge: A new model of dynamic inspection

Through folding the searching and the result inspection process into one exploratory phase, there is a genuine convergence between content and search. The time needed for the cycle of formulating the search – inspecting the results – reformulating the search is drastically reduced as it becomes immediately clear when typing in the query what kind of results a query pull out. This new model will require corresponding new graphical user interfaces that has to balance between providing a preview of results and being non-intrusive for users that already know what to expect.

5. TE-Commerce interactions

In this chapter the different ways of interacting holding between E-Commerce agents are examined. Interaction is here to be understood as a very broad concept, applicable to any form of "having something to do with each other". It certainly involves some commercial-related activity taking place, though. Typical E-Commerce activities, which will be explored in greater detail below, comprise gathering of information, advertising, browsing [1], purchasing and reviewing.

In particular, it is tried to lay out

- possible modes of interaction
- the data that accumulates thereby
- who gets hold of these data and what is done with it
- how these modes are expressed verbally
- a general framework of TE pull & push modes

TE-Commerce interactions comprise anything from one-to-one relationships (one customer purchasing an item from an online store), one-to-n relationships (online stores advertising to many users) or n-to-n relationships (paid placement companies delivering advertisements to content pages).

In this chapter, it is tried to elucidate possible ways of interaction between agents of different levels by following the same lead that was taken in the previous chapter: . Starting with a sketch of all possible interactions, we then focus on the textual aspects of interactions, divided into a section that deals with modes of access ("pull" modes) and a section that deals with modes of presentation ("push" modes).

5.1. Interactions between different agents

In each of the following sections, both directions of interactions are studied, as the nature of feedback and feed"forward" differs for each agent. Setting out with 1^{st} degree agents, it is possible to demonstrate how the world of E-Commerce divides nicely into the three spheres of B2B, B2C and C2C. Crucial for the framework developed here is how data is exchanged along the interactions, who gets hold of the accumulated data and how they are used. The model of packet exchange with its request and

[1] It is worth to note in this context that browsing was once a shopping activity before becoming a common term for online access to any kind of information.

answer cycles has its origin in Net protocols such as HTTP, yet can be applied to TE-Commerce interactions in general. The diminishing importance of face-to-face on the virtual market (at least real-face to real-face; the issue of cyber identities has been touched upon before) goes hand in hand with a growing value in automatically accumulated data. With hard disk space getting cheaper every year, it is now feasible to store fingerprints of every interaction for an almost unlimited span of time.

The subsequent sections deal with interactions between higher level agents and try once again to entangle the net of relations that holds between Search Engines, Search Engine Marketing companies, Portals and Content Sites[2].

5.1.1. Along the value chain: Interactions between 1st degree agents

Interactions holding place between 1^{st} degree agents do not confine themselves to selling and buying. A common subdivision of transactions separates Business-to-Customer (b2c) transactions with retail trade at its core and Business-to-Business (b2b) transaction with manufacturing and wholesaling as core activities. With the boost in communication opportunities enabled by the Net, a new division of Customer-to-Customer (c2c) transactions has emerged on a large scale.

Apart from buying, other types of transactions are also prevalent in TE-Commerce, such as renting, ordering of services, licensing, downloading trial versions, exchange reviews or sentiments, compare prices etc. If organized sequentially along the value chain, these transactional types can be grouped in three main classes, pre-sales, sales-related, and after-sales.

In the pre-sales process, potential customers share informations, compare prices, read reviews, express sentiments. In this context, both structured information such as product feature data (for example CNet Channel's product data) as well as participative contents on review sites (such as epinions.com) play a major role. Companies, on the other hand, promote their services and products. Recently, quite a few companies also participate in social media (for example managers writing blogs). This only adds at destructing the image of a strong opponency between companies pushing their offers into the market and customers passively absorbing these. On the Net, the distinctions became much more blurred as all different shades of interactions have evolved between consumers, between companies and between consumers and companies — albeit one has to concede that commercial bulk E-mail corresponds to this image. Most of these interactions on the Net have in common, however, that they are represented textually.

The actual sales process involves activities taking place offline, apart from non-physical goods (for example software downloads or paid content) that can be exchanged online. The *click-and-mortar* notion, prevailing in the 90s to denote the necessity for E-Commerce retailers to provide both online interfaces plus offline logistics, has evolved into more differentiated ways of transactions oscillating between offline and online. Ebay users might choose to pick up their auctions if the place is nearby; Yellow Page

[2] For the German market, Stefan Karzaunikat's overview, dating from beginning of 2004, is still the most current scheme. See `http://www.suchfibel.de`.

site users might walk into the store they just located online; downloading paid content and paying with PayPal does not once leave the online world.

Regarding, after-sales the amount of accumulated interaction data allows user tailored recommendations, based on the users buying history and on the buying history of other users (collaborative filtering). Keeping track of transactions allow companies to target their customer care management activities. By harvesting user reviews (sentiment analysis, see the Outlook chapter), companies are able to spot what users like or dislike with regards to their products.

The bidirectionality of interactions on the Net is strikingly illustrated in the reciprocal ratings of buyers and sellers at eBay. Note that the ratings of buyers exist independently from a penalty system for fraudulent buyers[3], making the reciprocal ratings indeed an intentional act to tighten the community and build up a social pressure on behaving properly. As both ratings referring to buying and to selling are added up, the ratings of buyers might have influence on the success of a seller, although do not affect the chances of buying further items, as the seller cannot refuse the highest bidder. Compared to the habits of the offline world this yields a rather particular situation: Buyers rate individually and publicly the sellers (not a very common situation offline), while sellers do not only keep records of buyers, but also share these records with other sellers and other buyers, yet these information do not affect the purchase of items — still the highest bidder wins regardless of her reputation. The mixture between publicity — as ratings on eBay are openly viewable to all Net users —, peer-to-peer and a central instance with almost unlimited power of suspending users is indeed a peculiarity which could be hardly though off elsewhere than on the virtual market. Apart from the lesson that interactions on the virtual market may fundamentally differ from what is common use offline, another aspect worth highlighting is again the textual representations of interaction. The ratings themselves are delivered in short textual statements (80 letter), highlighting once again the connection between E-Commerce and crisp textual representations.

It is not only the bidirectionality of interactions which is a peculiarity of the virtual market illustrated by eBay, but also how quickly market participant can change role from seller to buyer and vice versa. Naturally, speaking of users and companies did not intended to exclude that these agents take different roles: users might also sell goods or offer services and companies, of course, do also buy products or use services. The former builds the customer-to-customer (c2c) business type, the latter the business-to-business (b2b). From the TE-Commerce point of view, the impact on vocabulary of this distinction has to be taken into account. While b2c and c2c market sites organize their content along different products (such as dvd raw media, dining tables, sweater), b2b do so parallel to branches (computer supplies, joiners, textiles). This often leads to some problems if a site of one type tries to expand their lines of business to the other types, for example b2b sites opening up for retail customers. The product types that users expect are often not included in b2b vocabulary inventories (see below, Part D).

[3] See `http://pages.ebay.com/help/policies/unpaid-item.html [Nov. 1, 2006]`.

The volume of b2b transactions is about 10 times higher than that of b2c transactions[4]. The following table presents figures of 2005 for Germany, Western Europe and the US, broken down into b2b and b2c[5]

Business model	*German market*	*Western Europe market*	*US Market*
b2b E-Commerce	289 Billion EUR	1.204 EUR	1.280 Billion EUR
b2c E-Commerce	32 Billion EUR	125 Billion EUR	US 115 Billion EUR

Apart from commercial transactions, one also has to treat interactions between users and content providers to get a full picture of interactions between first level agents. Some aspects of this, namely user intentions behind accessing content sites, will be dealt with in Part C. Here, the different site metrics will be discussed, as those have immediate commercial impact. While metrics such as impressions or unique visitors have been a commonly traded coin on the Net through advertising models that bills ad exposure, the significance of these metrics has decreased with the advent of pay-per-click models.

Impressions is a classic term in advertising that was applied to gross audience long before the Web. An impression is a single showing of an ad and are counted in thousands. The CPM (cost per mille) is the price the advertiser has to pay for thousand impressions. Impressions do not differentiate between unique and recurring visitors.

Visits or user sessions occur if a request for a page is made the first time. Recurrent requests within one timeout period such as 30 minutes belong to the same user session and do not add to the user session count. Regarding visits, there is no discrimination between brief visits and extensive inspection of a website.

In addition, unique users only count visits with an unique fingerprint, mostly IP address, user agent (i.e. browser version) or cookies[6]. Imprecisions of this metric may arise because IP addresses are assigned dynamically by many large DSL or dial-up Internet providers and because users delete their cookies.

As a conclusion, none of the site metrics presented above are fully reliable, especially if one considers the use of anonymous surfing systems (which will in general lead to an understatement of true visits, as a uniform agent serves many real users) and of automated web crawling (which will lead to an overstatement). In recent years, only very large websites have been able to promote their content based solely on impression figures[7]. Smaller websites have in general to resort to performance-based metrics, such

[4] For a detailed presentation of the German and European E-Commerce market: `http://www.bmwi.de/BMWi/Redaktion/PDF/Publikationen/monitoring-informationswirtschaft-9-faktenbericht-chartbericht,property=pdf,bereich=bmwi,sprache=de,rwb=true.pdf [Nov. 1, 2006]`.

[5] EITO 2005, online at `http://www.eito.com/download/EITO%202006%20-%20ICT%20market%20March%202006.pdf [Nov. 1, 2006]`.

[6] For details see IFABC Global Web Standard, online at `www.ifabc.org/standards.htm [Nov. 1, 2006]`.

[7] Google Adsense offers a program for domain name holders — however, these had to attract more than 750.000 page views per month (this explicit restriction was recently removed from the google.com website). See `http://www.google.co.uk/domainpark/`.

as pay-per-click or pay-per-lead.

5.1.2. Making matches: Interactions between 1st degree agents and 2nd degree agents

Feed-forward mechanisms from 1^{st} to 2^{nd} degree agents comprise users querying Search Engines, businesses advertising on SEMs, finally content operators submitting their pages to Search Engines and SEMs in order to get traffic and relevant ads to monetize it. This results in the following schematic diagram:

Feed-back mechanisms from 2^{nd} to 1^{st} degree agents occur first with search agents taking input from users in order to optimize their results (Relevance feedback). Second, businesses receive traffic (and convert a part of it to revenues) from their ads. Finally, content operators receive their share of ad revenues from SEMs. Adding these feedback channels to the schematic diagram ends up in:

The following figures illustrate the extent of interaction between 1^{st} and 2^{nd} degree agents. They are broken down by the arrows depicted in the schematic summary above:

1) Users querying Search Engines and Search Engines leading traffic to Content pages

Worldwide, about 400-500 million queries are issued per day[8]. On average, users type in 40 queries per month, yet spend only 5% of total online time at the site of a Search agent, in contrast to 41% for communication, 36% for content and 19% for commerce.

The target sites for Search Engine traffic differ remarkably between different Search Engines. For example, search clicks on Google (7.9% of all clicks) and Yahoo (6.7%) are more likely to go to Education sites than clicks on MSN (4.3%). The latter Search Engine however, sends a higher percentage of users to Business & Finance sites (8.8% of clicks) than Google or Yahoo[9].

2) Indexing of Content pages

Google's index of Web pages rose from 4 billion pages in end of 2003 to 8 billion pags end of 2004[10]. Although these figures are partly driven by marketing needs (usually they were announced just as a competitor surpassed the own index size) and should be met with some methodic scepticism (for example on the deduplicating of very similar pages or the count of pages that are not indexed but which URL is known through a link), there should be some substance to them. This being said, the average index growth per day results at about 11 million pages a day.

3) Traffic deals

[8] See `www-lsi.upc.es/ rbaeza/websearch.pdf [Nov. 1, 2006]` for figures for 2005.

[9] See `http://www.seroundtable.com/archives/003381.html [Nov. 1, 2006].`

[10] See `http://searchenginewatch.com/showPage.html?page=2156481 [Nov. 1, 2006].`

One striking peculiarity of TE-Commerce is the existence of traffic wholesalers, i.e. sites — usually for one topic — that buy traffic via advertising on portals and redirect it to smaller businesses. Examples of such aggregators are `Carsdirect.com`, `Hotels.com` or the various instances of other vertical business directories. As their outbound traffic is more qualified (users have already specified their needs) than their inbound traffic, they can charge the outbound traffic at a higher rate. The smaller businesses have the advantage of receiving traffic that they could not incite themselves by lack of budget or resources.

For the framework of TE-Commerce interactions the existence of such a business model underlines the importance of traffic quality. The quality of traffic can be measured a posteriori through conversion rates; a priori it can be characterized by the selections the user made, either through typing in queries or clicking on links[11]. In both cases, the selection is manifested textually, as a chain of anchor texts and/or query strings.

4) Business contacts created by SEM

The number of business contacts created by SEM can be deducted on two ways, either through total revenues of SEM or through the number of total clicks. Considering that Google which has roughly a share of 40% of the SEM market reported 1.6 billion USD revenues through advertisements in one quarter and has an average bid price of 1.60 USD[12], this results in the total of 28 million paid clicks for all SEMs. These clicks happen both on the site of Search Engines as well as on affiliate sites. The alternative calculation starts with 170 million queries per day with a general click rate on paid links of about 12%[13] which results in ca. 20 million contacts. The number is lower as it only includes paid clicks on Search Engines.

What transmissions are being used for advertisements to get from the vendor to the customer? This issue becomes even more important with the very specific ads that are one of the main innovations of E-Commerce. While designing a single ad for a huge audience is a most challenging task of copywriting and marketing psychology, the feeding, delivery and maintenance of thousand of different advertisements poses a more practical difficulty. As interaction method between E-Commerce players, it is included in the overview presented here Some aspects of these processes are connected with query treatment issues and will be dealt here; other aspects are demoted to the *Paid Term Space* section below.

[11] Sociodemographic properties of the traffic source — consider the difference between a teen portal and a yachting portal — play a role, too. However, these properties exhibit themselves also in queries and links leading to the traffic source.

[12] See `http://www.clickz.com/showPage.html?page=3550881 [Nov. 1, 2006]`. Note that Yahoo has a much lower average bid price, approximately 40 cent. See http://smallbusiness.yahoo.com/r-article-a-6655-m-6-sc-36-paid_search_overview-i

[13] See `http://www.seroundtable.com/archives/003381.html [Nov. 1, 2006]`.

Ad upload, management and delivery

The interactions related to online advertisements deserve an individual treatment. In this overview, only standardized and publicly available programs are taken into account. Obviously, most paid per click providers are willing to go lengths for revenue-strong advertisers and have editorial teams assisting with the set-up and maintenance of campaigns. However, a well designed online process should reveal what is immediate possible and when it is time to contact editorial assistance.

Setting up a single ad can be done in few steps on today's SEM sites such as adwords.google.com or overture. Once registered (with a validated billing address[14]), advertisers may type in in title, description and URL of their ad through HTML forms. Selecting the keywords for which the ad should appear usually finishes what the advertiser has to provide. If editorial checking is obligatory, few days have to pass before the ad may go live. Automatic checking allows an almost instantaneous placement of ads.

Even if automatic checking can detect problematic vocabulary (adult terms, offensive terms) and some breaches of guidelines, for example using superlatives, there will always be a chance to bypass the checking with confuse and irrelevant ads. However, as ads with low click-through rates are discarded soon — Google Adwords eliminates them after approximately 7–14 days — the survival of the fittest ad ensures that bypassing the filter will not have a lasting effect.

With advertising professionals handling dozens of client or one company that wants to monetize their inventory with hundreds or thousands of titles (for example bookstores or record stores), upload issues become crucial[15]. Some SEMs offer spreadsheet feeds, rendering it easy to transform inventory into ads. In general, this also requires placeholders in the teaser. For example, *buy X at ...* where *X* represents for the keyword that triggered the ad.

The following table summarizes the bid feeding options on the major SEMs:

SEM provider	*Title length*	*Ad Text length*	*Mass upload*	*Editorial rules*
Google	25	70	API	auto
Yahoo	60	250	Excel	manual
MSN	25	70	Excel and csv	auto

The standard campaign management fields for advertisers comprise of setting a daily cost limit, booking keywords, editing the ad text and setting maximum costs per click. Performance details (searches, clicks and costs generated by clicks) can be inspected for each keyword. Keywords in danger of being inactivated because of a too low clickrate are highlighted. Various grouping options are available to automate parts of the campaign setup, for example copying ad texts and titles for different keywords.

[14] Some SEMs impose a minimum turnover. For example, Yahoo search marketing charge a minimum of 25 EUR per month for any advertiser running active campaigns. Google and MSN have 5 EUR entrance fee.

[15] One could also think of sites such as ebay.com that promote almost anything that could possibly be in their index, even allowing for ads that lead to no results on eBay.

Once a SEM company accumulated a pool of advertisers, rolling out the ads on affiliated content web pages may start. Typical affiliate programs are transacted in the following way:

- content website operator registers at the SEM
- SEM spiders the website for basic checks (availability, redirect, problematic navigational behavior such as pop-ups, problematic content keywords)
- content website operators add HTML code to their page which load the ads from the SEM's server
- reporting and monitoring of the campaigns through the administration interface on the SEM's website

For affiliates that run a search slot, optimal results can only achieved if the query string is passed to the SEM. However, this is not offered in the standard interface from Google, Yahoo or MSN and is only available for larger affiliates. As an alternative, if the query string is passed via the GET-method to the search back-end, than the persistent URL allows spidering of the content and delivering targeted ads.

SEMs disapprove of any artificial increase in clicks, for example through any incentive signs around the ad. Google Adword's rules explicitly forbid *drawing any undue attention to the ads*. The affiliate's webpage must not contain phrases such as *click here, support us, visit these links*. Obviously, any incentive for clicking ads apart from their content bears the danger of delivering low quality traffic to advertisers.

Commonly, SEM programs offer also to advertisers choosing sites on which their ads must not appear (blacklists). Conversely, affiliates might choose to filter out websites from a blacklist of URLs[16]. This allows shielding of ads from being displayed on competitors websites or any other unwanted website. However, on Google Adwords filtering can only be set to URL or parts of a URL, not to keywords or topics. It is not possible to set up whitelists of preferred topics or URLs.

5.1.3. Combining results: Interactions between 2nd degree agents

Search Engines and SEMs interact tightly (so tightly that the biggest SEM providers are part of Search Engine providers), as Search Engine sites are an ideal place for placing ads delivered by SEMs. Search Engines create qualified traffic while SEM allows monetizing it through their ad pool. The traffic provided by Search Engines is filtered exclusively by the query string, in contrast to content pages where it is much harder to determine what needs brought the users to it. Furthermore, ads blend into the rest of the Search Engine's result page. If properly chosen and formatted they do not only look similar to additional search results, but also function in the same way.

[16] See `https://adwords.google.com/support/bin/answer.py?answer=13248&ctx=sibling` and `https://www.google.com/support/adsense/bin/answer.py?answer=21593&ctx=sibling [Nov. 1, 2006]`.

In a combined result page consisting of algorithmic and paid results, there is an open competition on which part delivers the most relevant results. What look like homogenous listings on the front-end can be derived from different sources in the back-end. Different logics are feasible in this context: Dividing the screen into several zones (as done with paid results and algorithmic results), ranking of candidates (when more candidate results than fit the screen are available) and finally the notion of fall-through (when too few candidates from one result set are available). Fall-throughs are a viable way to combine fine-grained results with broad coverage of queries; for example, algorithmic Web index results or Wikipedia results might serve as a fall-through for a specialized search.

The notion of combining several resources in order to set up a screen adds another facet to Term Spaces: Term Spaces from different sources, with different weights and weighting schemes may be combined to create one front-end. This again adds to the commensurability of terms in Term Spaces touched on before.

5.1.4. Restrictions and Enablement: Interactions between 2nd degree agents and 3rd degree agents

Dividing interactions between 2nd degree and 3rd degree agents top-down, one might discriminate restricting and enabling modes. 3rd degree agents in the restricting modes constrain interactions in the overall framework, those in the enabling mode facilitate interactions. This should not suppose a Manichean view of 3rd degree agents, as restricting unwanted or illegitimate interactions is as helpful for a viable framework than facilitating bona fide interactions.

Restricting modes

A common field of regulatory activities lies in supervising concentration s and monopols. In general, Search Engines have rejoiced from a considerable freedom of media regulation. In the US, the First Amendment to the constitution guarantees freedom of speech, a stronghold also for Search Engines who could claim their results to be individual expressions of opinion. German's regulatory law states that Search Engines have to name a youth protection executive (JMStV), but treat them otherwise as "Teledienste". In contrast, to "Mediendienste", these have much less obligations, basically comprising of the duty to state imprint information and removing problematic content upon notice. Being treated as neutral providers of algorithmic results, further regulation policies seemed neither feasible nor appropriate.

However, with Search Engines introducing paid results new fields of possible conflicts emerged. In Germany's state's broadcasting law ("Rundfunkstaatsvertrag"), the separation of content and advertisement is demanded. In the US, the Federal Trade Commission's letter to Search Engines in 2002 gained attention, as it enforced the visible separation of paid placements and paid inclusions from index results[17].

[17] See also www.commercialalert.org/PDFs/ftctosearchengines.pdf [Nov. 1, 2006].

Search Engine data is coveted by law enforcement agents: For many purposed of national security, crime prevention and youth protection, Query-logs are a valuable resource, especially in combination with logged IP-numbers that can be traced to individuals. The meta Search Engine Ixquick tries to make a value proposition out of this by declaring not to store any individual user data on their servers.

Regulatory measures do not necessarily come from state organs. Self-regulation is a widely discussed concept, especially on the Net[18]. In the area of Search Engines in Germany, the Bertelsmann Foundation initiated a Code of Conduct for Search Engines with rules especially on adult content, hate speech and other problematic content in the results. An example for a self-imposed code of conducts are the Adword policies established by Google. On a wide variety of topics, no advertising is in principle allowed, including the following areas:

- Aids to Pass Drug Tests
- Alcohol
- Anti and Violence
- Bulk Marketing
- Cable Descramblers and Black Boxes
- Cell Phone Jammers
- Counterfeit Designer Goods
- Dialer Programs
- Drugs and Drug Paraphernalia
- Fake Documents
- Fireworks/Pyrotechnic Devices
- Gambling
- Hacking and Cracking Sites
- Miracle Cures
- Mod Chips
- Prescription Drugs and Related Content
- Proper Names
- Prostitution

[18] See also `http://www.selfregulation.info` for a first information on this topic.

- Sexual Content (Adult)
- Solicitation of Funds
- Tobacco and Cigarettes
- Traffic Devices
- Weapons

This list contains a mixture of topics unwanted or heavily restricted in one or several markets, such as the EU or the US. While gambling is underlies state restriction in most EU countries, it was until recently far more liberal dealt with in the US[19]; alcohol on the other hand is on the whole more liberal handled in the EU than in the US. This list should not convey that no advertisements on one of its topics appears on Google; there are plenty of ads for queries such as *red wine* or *sex*, both on google.com and google.de. However, ads might be removed from Google if they promote items or services appearing on the list.

Enabling modes

Enabling modes help to create more interactions and higher the quality of interactions. An instance of enabling interactions is provided by a fair use of Search Engine Optimization. Fair use of Search Engine Optimization attention to a relevant website by adding appropriate vocabulary, trying to list the site on relevant hubs for its topic and removing technical obstacles for the Search Engine, such as texts hidden in images.

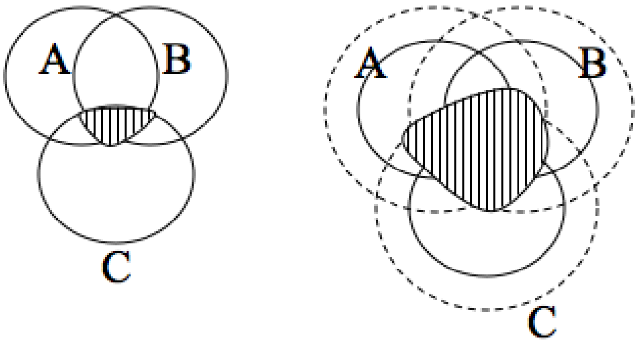
In general, term-related enhancements of search applications may help in at least four different ways:

1. Broaden the set of results to include hits that use a variant of the search term
2. Recuperate weird searches
3. Help to focus general searches on more specific topics
4. Hint the user to relevant additional searches not thought of when issuing the query[20].

Term-related enhancements allows more than just adding more hits for a search. Determining a larger set of potential hits and more precise ranking methods aware of term variation helps to produce more useful results. For example, if a query consists of three terms A,B and C, current systems return the intersection of results set for A, B and C respectively (see left figure). Clearly, one loses all relevant results that use a variant of A, B or C. Through recognizing term variants and related terms, the intersection increases largely:

[19] Unlawful Internet Gambling Enforcement Act. See `www.govtrack.us/congress/bill.xpd?bill=h109-4411` [Nov. 1, 2006].

[20] This aspect is similar to cross-selling.



Result sets for a query with the terms A, B and C.

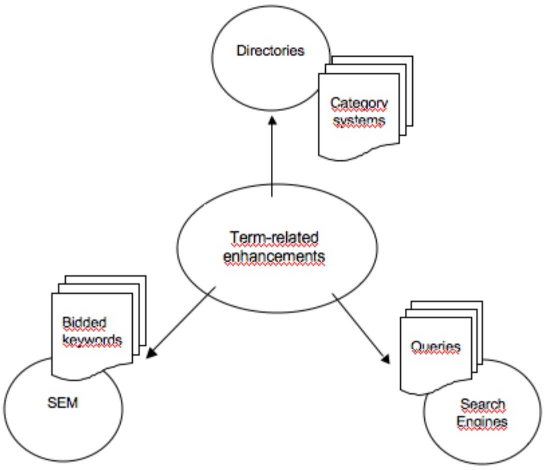
Note that just the top-ranked results of the increased intersection needs to be returned to the user. Depending on the query type, the ideal large of a query differs and might even be as small as one canonic result for one type (see below, Part C).

Among important discriminating features in search technology providers and their solutions are the following:

- standardization vs. ad-hoc solutions, i.e. the degree of customization
- pricing options: performance based, i.e. a share in the increased number of transactions (also called uplift fee) or as a license model
- form of delivery: ranging from providing static term databases to delivering packaged software to passing data on to the servers of the search technology provider

5.1.5. Schematic Summary

The scheme below positions search technology providers and their main data accumulation in relation to other agents:



In the center of this model, term-related enhancements function as 3^{rd} level agent that facilitates the matching of the individual data resources and thus increases the volume of TE-Commerce interactions.

5.2. Access modes

In the model of information driven E-Commerce, access is a key concept. By access mode we understand the type of selectional actions a user takes in order to get to E-Commerce relevant information. Selectional actions on the Web are units of user's interaction, comprising of interface device actions such as keystrokes, clicks, mouse movement and scrolling[21]. The approach presented here is restricted to a subset of selectional actions, namely those that trigger access to new information.

Considering common web technologies, typical access actions are filling out and submitting forms and clicking on links, either with textual anchors or images. Javascript also allows other events to be captured, for example moving the mouse over the area of a specified HTML element. The now widely deployed AJAX makes access actions even less recognizable by constantly feeding new information. Still, two principle modes of access can be distinguished, access by browsing and access by searching.

5.2.1. Access by Browsing

Access by browsing describes the method of delivering new content based on move, click and scroll actions of the user. Following hyperlinks is one example of access by browsing, though there are several other means of browsing. All have in common that the user does not type in textual strings conveying her selection, but chooses a visible object.

It is thus not strictly necessary that access by browsing is never performed by keyboard. However, the selectional choice must be already present when the user triggers it, which is the case for example with keyboard shortcuts.

Compared with entering a query, the selectional range of a user is much more limited when she access information by browsing. Elements that can be browsed have to compete for the limited space on the screen. Complex menu structures with encapsulated submenus have the drawback that every additional click thins out the user traffic. In addition to this, Java-script based or Flash implementations of complex menus are often not as searchable as plain html pages and are not accessible to every user.

A second crucial difference is the presence of selectional choices before the user triggers one of them. Access by searching typically leaves the user for some time in ignorance about what happens next and if any results at all are produced. Access by browsing, on the other hand, is in general accompanied by the indications that some results will follow from making the browsing selection.

[21] See [Wirth 1999].

Anchor links as descriptors

A large part of access by navigation operates through links with anchor texts. Anchor texts are the text labels of links that appears between $< A >$-Tags in HTML. Additional glossary terms in this context is the *target page* to which a link points to and the *source page* on which it appears. Internal anchor texts appear with links connecting a source page and target page of the same site. External anchor texts appear with links pointing to a page of a different site than the source page.

Navigational anchor texts, occurring mostly with internal anchor texts, do not reveal any content property of the target site. Among typical navigational anchor texts are *here*, *click here*, *next page*, *home*, *back* or indexes of first letters.

Content anchor texts, however, provide some indication to the content of the target page. In the ideal case, they state a compressed description of the target page. Through the re-occurrence of external anchor texts, a reliable picture of what a target page is about can be discerned. The reliabilty depends on how independent the sites exhibiting the links are. As Search Engines heavily exploit external anchor texts, it is alluring to manipulate them by setting up a network of interlinked sites. Different measures, including graph algorithms that detect cyclic structures may be helpful to detect manipulative anchor texts[22].

From the point of view of accumulating vocabulary, both internal and external anchor texts provide a valuable resource. As typically the size of cumulated internal anchor texts in a large crawl exceeds the external anchor texts by factor 5-10, neglecting internal anchor texts leads to a tremendous loss in corpus size.

A further benefit of internal anchor texts is their usefulness for classifying types of websites, for example business directory, company homepage or private homepage. The majority of company homepages have one of the following anchor texts: *about us*, *investor relations*, *our products*.

As a compressed description of the target page anchor texts are similar to queries. Matching queries to anchor texts, especially re-occurring external anchor texts, produces in general a highly relevant results as re-occurring anchor texts contain only meaningful and characteristic descriptors of target pages.

Site map evaluation

Based on a robust recognision of important terms on a site it is possible to print a separate site map that is ranked based on query log driven data. This new site index can be compared to the access possibilites originally offered to the user. Accessibility of a content page can be calculated by counting the selectional actions the user has to take (clicking on a link, choosing a form field, scrolling, entering text).

If restricting the access to navigational browsing, a simple metric just counts the number of clicks necessary to get to the page holding the content information. Evaluation could then be done either as a macroevaluation over all terms on a site that also appear in a query log or as a microevaluation for individual pieces of information

[22] See [Chakrabarti/van den Berg/Dom 1999].

such as products offered on a site. For example, locating a specific notebook model on samsung.de requires three clicks: Produkte — Notebook PC — Drop-down Modell auswählen. In automatically calculating the navigational distance, clicking links has to be discriminated from selecting a form field, as crawling forms is still a challenge to automated spidering.

Refining these measure should take both the graphical representation and the textual content of the selectional elements into account. The graphical representation comprises of the prominence of the link on the screen, determined by its size and position.

Examining the textual content of links that lead to a specific piece of information can be done through setting up navigational paths, i.e. the concatenation of anchor texts labeling the links that lead to this piece of information. Navigational paths can be compared both in terms of length and terms used.

DDR-RAM

`preissuchmaschine.de`

Computer > Arbeitsspeicher, CPU & Mainboards > PC-266 DDR-SDRAM

`pearl.de`

Hardware & Multimedia > Bauteile, Gehäuse, Brenner & Rohlinge > DDR > NoName

Frontlader-Waschmaschine

`preissuchmaschine.de`

Haushaltselektronik > Waschmaschine > Frontlader

`zarsen.de`

Waschen & Trocknen > Trockner > Waschmaschinen > Waschmaschinen Einbau

Miniofen

`promarkt.de`

Top-Angebote > Haushaltskleingeräte

`preissuchmaschine.de`

Haushaltselektronik > Back- & Grillgerät > Backofen > Toaster/Miniöfen

From these examples it becomes clear that it is not always easy to predict for a user which navigational path might lead to the aim. Both the length of the path and the terms used in the hierarchy vary. Although it is tempting to try to deduct synonyms and related terms by comparing the paths leading to the same product on different sites, one has to be aware of the subtle semantic changes — consider for example the quasi-synonyms *Haushaltskleingeräte* and *Haushaltselektronik* — and the edit rules of category wording — for example the usage of compounds (*Haushaltskleingeräte*) and coordinated terms (*Back- & Grillgerät*) — that occurs in the terms of thee paths.

5.2.2. Access by Query Search

One component of Internet literacy is recognizing and manipulating a query interface[23]. An indication of how deeply ingrained querying is into Internet literacy lies in the difficulty to describe what one does conceptionally when putting in a query. Several models are conceivable but not any one of them alone is sufficient: ranging from asking a question, placing a directive, describing an informational need, referring to a named entity, putting in pieces of texts memorized from the last time the page was seen etc. Access by searching is used both for getting a general overview on a topic as well as retrieving a specific information (details on a classification of search queries and user intentions are provided in Part C). Obviously, users soon develop their own strategies in order to adapt to the capabilities of a search agent.

In this adaption proces, a standard *look and feel* of search engines has evolved featuring one big query field, a submit button and a separate result page listing the results linear and in priority order [24]. Though the size and the placement of the search slot is generally less prominent on content sites (such as online newspapers), search interfaces on these sites follow the same logic. If a search agent's interface deviates from this design, it is likely that many users will not recognize its function. The downside of a commonly shared image of how a search interface should look like is that it impedes innovative interfaces. In addition, as the standard Web search engines also perform in a comparable way — for example, typing in a query *hotel near neuschwanstein at 200-250 EUR per night available tomorrow night* will not work on any search engine, *hotel neuschwanstein* will produce at least some valid results — adding new elements to the search logic will require some time until users become bolder and find adaptive strategies for the augmented capabilities.

[23] The issue of internet literacy is discussed especially from the point of view of education and school curricula. However, it affects more aspects of society, including government and work place. See the White Paper of the 21st Century Internet Literacy summit, Bertelsmann Foundation and AOL Time Warner Foundation, Berlin 2002.

[24] See Jakob Nielsen, "Mental Models For Search Are Getting Firmer" online at `http://www.useit.com/alertbox/20050509.html [Nov. 1, 2006]`.

Providing example queries around the search slot is a good measure to guide users to what type of search queries the system is supposed to handle well. However, detailed analysis of query logs reveal that many users do not pay much attention to the context of the search slot. For example, Yellow Page sites offering both a What? and a Where? slot typically have some amount of intersection between these fields that is due to users mistakenly type into the wrong search slot.

5.2.3. Combining search and browsing

Both searching and browsing entail inspection cycles in which the user moves forwards (assessing the results) and backwards (reformulating the search or following a different navigational path). These two movements correspond to the two tasks any agent providing information has to accomplish, firstly processing the users' input and transform it into a look-up command, secondly performing this look-up on the database storing the results.

In the case of Web searching, the first task consists of a query processing that analyzes the search input and transforms it into an index look-up. The second task organizes the content of crawled webpages into one or more index tables.

Compared to this, the first task in access by navigation consists of placing links and adding anchor texts. The second task is distributing the content of a site among several browseable units, including webpages, popups, paragraphs etc.

There are two filters that a user query has to pass before it is met with relevant results — the system has to recognize how the query is worded and the index has to store the result wanted. For example, if searching for a *hairdresser* in a specific location on a categroy-based Yellow Page site[25], there are two possible scenarios why the search could fail, either because the query term is not matched to the proper category or because there happens to be not hairdresser in that location.

Naturally, many sites use both strategies to get the users to the content wanted. There is major difference, however, between these two in how active users have to be in verbalizing their wants. Access by browsing provides a selection of already verbalized items, while a search slot is at first an empty line that the user has to fill by her own wording.

While access by browsing shows the user before what can be expected to be in the result set (apart from cases where the label of the navigational element and the content of the target page deviate), access by searching is at first a jump into unknown water. If it returns no hits the reason might either be that the result wanted is really not in the index or alternatively that the wording of the query was not suited for producing it. It is only through a trial-and-error process that the reason might be determined. The uncertainty why a search fails is even aggravated if a search application answers similar searches with disparate results as most Web searches do today (see Part C).

As a consequence, evaluating the quality of a search agent has to make sure that it does not evaluate the content of the index. Zero hits for a search might be the

[25] Different search logics on Yellow Page sites are discussed in detail in Part D.

consequence of a genuine gap in the available data which cannot be attributed to the performance of the search agent. Zero hits because of lack of records, vs. failure of query-recognition. If a result is in fact not in the index, this information has its worth even without presenting suggestion on where to look at elsewhere. In this context, it is the consistent handling of similar searches that is the indication that the first filter — recognizing the intent of the users' search — performs well and that zero or few results are really due to gaps in the index and there is no need to further try very similar searches.

A solution could lie in combining searching and browsing by creating link to results while the user starts to type a query. Such a suggest mode has to be capable of query variation principles — orthographic, morphologic, syntactic, semantic and head/container-related (see Part C) — in order to get the best out of both worlds. *Shrowsing*, the mixture of searching and browsing, would allow dynamic inspection of the available index data as the user types [26].

5.3. Presentation and Delivery Modes

Apart from the physical exchange of goods and money in E-Commerce, transactions are also accompanied by various textual presentation and delivery modes. The conventional concept of documents as main textual container has to be questioned and augmented with alternative concepts.

5.3.1. Documents revisited

One of the historic remnants that still play a role when talking about search on the Web is the unit of documents. There are at least three different aspects of what defines a document: Firstly, documents as units described by their syntax and meta data including semantic meta data; secondly, documents as textually bound units; thirdly, documents as units for documentation and archiving purposes.

Following the first sense — typically, employed in Information Retrieval — a document is a unit if information characterized by the properties ruling it([27]. The document syntax refers to the structure in which its content is stored. Metadata information contains about the document data, such as who created it, when it was created or what the subject matter of the document is. The latter part of metadata is the semantic metadata of a document that characterizes how its content has to be read.

The second concept — with its origin largely in text linguistics — of documents stresses their textuality. It is observable through cohesion and coherence within a document, created by a network of intra- and extra-linguistic references interconnecting the parts of a document. The size of a textual document might span from complete rhetorical units (such as articles) to shorter forms (such as entries in a dictionary).

[26] A somewhat similar concept called opportunistic exploration is introduced by [Bryan/Gershman 2000]. It was not feasible then, however, to perform real-time fault-tolerant searches on very large indexes while the user enters the query string.

[27] See [Baeza-Yates/Ribeiro-Neto 1999].

However, one of the main properties of textuality — the possibility to determine a text's boundaries is challenged by new presentational and logical styles of documents on the Web. Dynamic reloading and changing of a document, for example through AJAX, blurs the distinction inside - outside a document much drastically than even the Hypertext visionaries dreamed of[28].

In a third sense — stemming largely from library and documentation science —, documents are viewed as units of documentation and archiving. Here, a document is a piece of information that needs to be stored and made accessible for future look-up[29] The size and content of such a documentary unit can be variable, as long as it is technically storable and worth to be archived.

All three model take the document as the main container of information. Regardless of what properties of this container are focused, documents are stable in the sense that if a information is in them, it remains in there. However, as many E-Commerce applications do not operate on any form of precompiled information units, but rather on dynamically created result sets, all three concepts seem insufficient. In addition, typical E-Commerce result types, exemplified by database records or listings of named entities, such as people or prices, do not obey to any model of document. As a consequence, alternative models to the concept of documents have to be introduced.

5.3.2. Database result and record pages

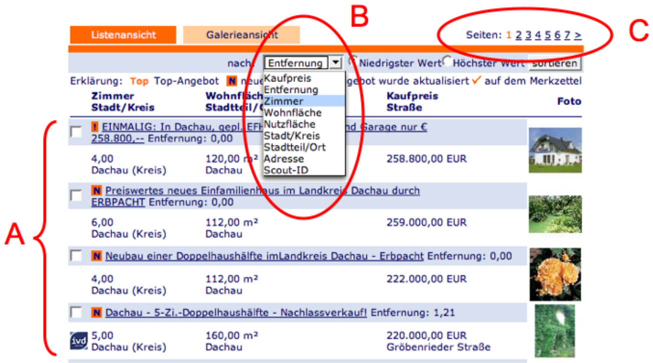
A typical TE-Commerce application has a database back-end, for example a database of products, prices or shops where to buy them. The prevalent model for displaying the contents of the database on the front-end is the combination of result and record pages.

The result page is a sorted list storing links to the records and preview information. Depending on the fields the data provide, different sorting mechanisms are conceivable. seeing sorted listings, which hold links to the actual records. This repetitive feature of inspecting results of the same kind brought life to different visual metaphors: book, book shelf, newspaper, building, lens, guided tour, pile of documents, galaxy/universe, aquarium[30]. In this screenshot of a typical result page three main visual elements are highlighted: A) The listings in a table structure, B) the sorting criteria of listings, C) the page navigator, usually with first/previous/next/last page buttons.

[28] See [Landow 1997].

[29] See [Gaus 2003].

[30] See [Mann 2002].



Erklärung: Top Top-Angebot N neue… …gebot wurde aktualisiert ✓ auf dem Merkzettel

Zimmer Stadt/Kreis	Wohnfläch… Stadtteil/…	Kaufpreis Straße
☐ ! EINMALIG: In Dachau, gepl. EFH… …nd Garage nur € 258.800,-- Entfernung: 0,00		
4,00 Dachau (Kreis)	120,00 m² Dachau	258.800,00 EUR
☐ N Preiswertes neues Einfamilienhaus im Landkreis Dachau durch ERBPACHT Entfernung: 0,00		
6,00 Dachau (Kreis)	112,00 m² Dachau	259.000,00 EUR
☐ N Neubau einer Doppelhaushälfte imLandkreis Dachau - Erbpacht Entfernung: 0,00		
4,00 Dachau (Kreis)	112,00 m² Dachau	222.000,00 EUR
☐ N Dachau - 5-Zi.-Doppelhaushälfte - Nachlassverkauf! Entfernung: 1,21		
ivd 5,00 Dachau (Kreis)	160,00 m² Dachau	220.000,00 EUR Gröbenrieder Straße

Result page of a database-driven Web site

Wrapping the content of hierarchical or relational databases into HTML-code leads to repetitive HTML code patterns which can be exploited to extract the database content from the HTML code.

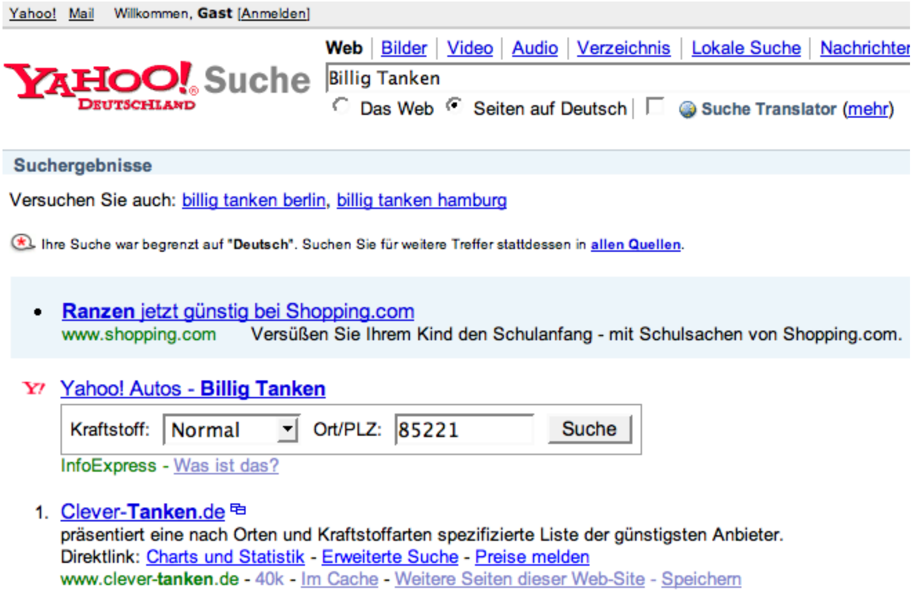
5.3.3. Snippets

As many needs in TE-Commerce require to scan through many possible offers — for example, searching for a hotel in a big city —, it is necessary to allow the user a quick inspection into results. Web Search Engines generally use a short preview snippet from the result document, including the title, the URL and a context of the query keywords. Previews should higher the perceived relevance of results and give a first impression of the result page's suitability for the search.

There are several other TE-Commerce applications where short interaction cycles call for a concise representation of results. Presenting products with features for example benefits from selecting and highlighting the most decisive features instead of delivering the full data sheet.

Various snippet sizes and property selection strategies are conceivable. In the case of Search Engines, for example, different query types would ideally product different kind of snippets. Results to almanac queries (see Part C) could then be presented on the basis of text style features (prosa, listing, register etc.). Results to queries for which one valid answer exists (for example asking for a telephone number of a person) only need to present the particular answer and information about the place where it was found.

A typical case of snippets are the short-cuts used on Web Search Engines, i.e. additional results displayed on top of the index hits for some selected searches. Querying the weather in a location on Yahoo, for example, produces a snippet that contains information from Yahoo's weather channel. Some short-cut results resemble the index hits in their lay-out, others deviate from it, for example by offering a separate search slot:


Yahoo! Mail Willkommen, Gast [Anmelden]
Web | Bilder | Video | Audio | Verzeichnis | Lokale Suche | Nachrichter
YAHOO! DEUTSCHLAND Suche
○ Das Web ● Seiten auf Deutsch | □ Suche Translator (mehr)
Suche

Billig Tanken

Suchergebnisse

Versuchen Sie auch: billig tanken berlin, billig tanken hamburg

Ihre Suche war begrenzt auf "Deutsch". Suchen Sie für weitere Treffer stattdessen in allen Quellen.

- Ranzen jetzt günstig bei Shopping.com
  www.shopping.com Versüßen Sie Ihrem Kind den Schulanfang - mit Schulsachen von Shopping.com.

Yahoo! Autos - Billig Tanken

Kraftstoff: Normal Ort/PLZ: 85221

InfoExpress - Was ist das?

1. Clever-Tanken.de
   präsentiert eine nach Orten und Kraftstoffarten spezifizierte Liste der günstigsten Anbieter.
   Direktlink: Charts und Statistik - Erweiterte Suche - Preise melden
   www.clever-tanken.de - 40k - Im Cache - Weitere Seiten dieser Web-Site - Speichern

A shortcut on yahoo.com

5.3.4. Presenting paid results

Paid results deserve a separate treatment in the context of Internet search agents, as conventions on how to place them, recognize them and react to them are not as stabilized as for offline advertisements. Many users regard search agents as objective devices and are not aware of commercial results[31]. It is not only on the side of the user that a lack of accustomization can be registered: Advertisers and content providers also have to adapt to the peculiarities of paid content, especially on the site of search agents.

Just as there are many different forms in which users view a web page, a variety of advertisement (in a broad sense, including all form of paid insertions on the space of a web page) display mechanisms has evolved since the early days of the Net. In this section, both the physical appearence of ads, i.e. banners, pop-ups or sponsored links, and the logic behind their delivery is going to be discussed.

One discriminating feature between ad formats is how they blend in or stand out from the surrounding content. Blended forms of advertising might appear in list contextes — for example paid inclusions in Search Enging results — or in block contextes, as can be often seen on online newspapers. Apart from a small indication, they look just like the surrounding content and are similarly worded. Gradually, there are forms of advertisments that are more easily discernible, yet do not interrupt the layout of the surrounding content, for example banners on the page margins. On the extreme pole, formats such as pop-ups or layer ads, especially when coming as rich media including sound, interfere with users perceiving the page content. These formats are not only

[31] See [Machill/Welp 2003].

often blamed as intrusive, users have also taken evasive actions in form of various ad blockers[32]

While from a stand-point of media policy the separation of index results and paid results is a necessary prerequisite, a qualitative commercial result should not need to snuck in among algorithmic results. Says Dan Thies, an independent SEO, “Speaking as an advertiser, I would prefer that my ad is known as an ad because it is more likely to reach the right audience. I wouldn’t want my ad snuck in”[33].

Balancing separation of ads and a homogenous layout is not always easy. Intrusive ads, especially on sober sites such as Search Engines, are likely to be ignored and in the worst case negatively affect the visits to a page[34]. On the other hand, some amount of separation is required, not only from a regulatory stand-point, but also because conversion rates will suffer if users without commercial intents are passed on. One component of the solution surely lies in the relevance of results, both algorithmic and paid listings. Signaling what kind of interactions the user can expect from the the site (being informed, sharing thoughts, purchasing etc.) adds to the relevance needed in this context.

Note that in the following, only sponsored links are taken into consideration. Other forms of SEM activities such as paid inclusion have lost their importance in recent years[35].

5.3.5. Partially filled out form

For a broad variety of E-commerce transactions, a relatively stable set of features exist that further defines or specify the transacted products or services. The transaction *buying a flight ticket* has obligatory features such as departure and destination airport and flight data. The transaction cannot be conducted without specifying these features, although the mode of transmitting these specifications might be offline. Other features are optional, as seating on a window or aisle seat in the case of buying an airplane ticket. In vertical Search Engines, such features are regularly presented as forms, in which the optional features usually appear only after the user selected an *advanced search*.

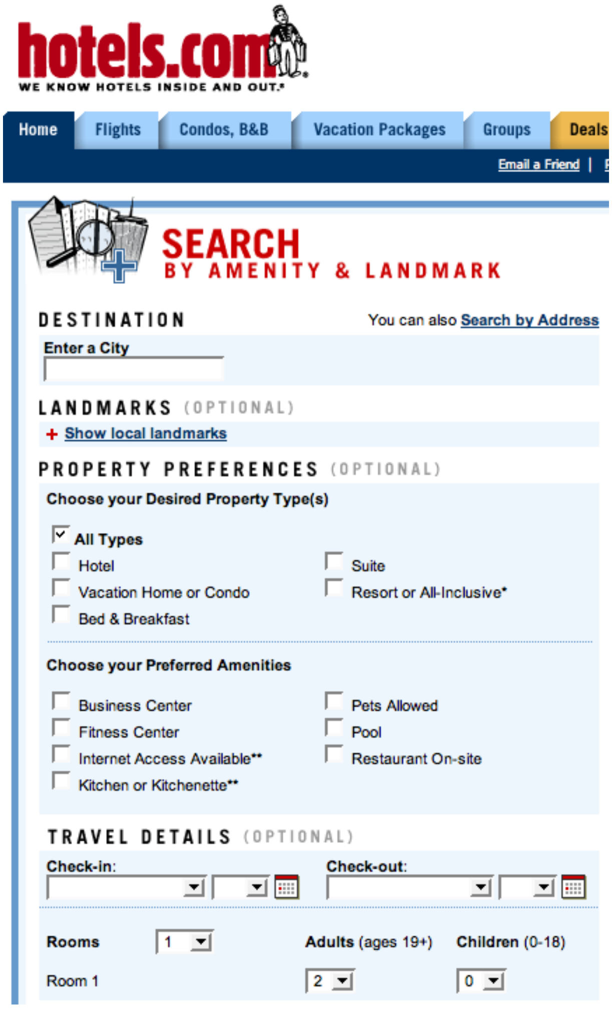
The two main components of any search provider, data acquisition and data search interface and result presentation are affected by the intrinsic features. Data acquisition — whether automatic by crawling Web pages or via submitting data to the search provider — has to include additional fields for the features that should be made searchable. The interface and presentation need to accommodate features, ideally in a usable way, for example by sorting features into simple and advanced search.

[32] In fact, one of the main propositions of the Firefox and Opera browsers are their capabilities to zoom out these ads.

[33] At `www.seobook.com/archives/000958.shtml [Nov. 1, 2006]`.

[34] Apparently, many advertisers try different routes and splash rich-media layer ads on content pages.

[35] MSN dropped paid inclusion in 2004 (see `http://www.clickz.com/showPage.html?page=337592 [Nov. 1, 2006]`), making Yahoo the only large Web Search Engine with a paid inclusion scheme.



SEARCH BY AMENITY & LANDMARK

DESTINATION

You can also Search by Address

Enter a City

LANDMARKS (OPTIONAL)

+ Show local landmarks

PROPERTY PREFERENCES (OPTIONAL)

Choose your Desired Property Type(s)

☑ All Types

□ Hotel	□ Suite
□ Vacation Home or Condo	□ Resort or All-Inclusive*
□ Bed & Breakfast	

Choose your Preferred Amenities

□ Business Center	□ Pets Allowed
□ Fitness Center	□ Pool
□ Internet Access Available**	□ Restaurant On-site
□ Kitchen or Kitchenette**	

TRAVEL DETAILS (OPTIONAL)

Check-in:

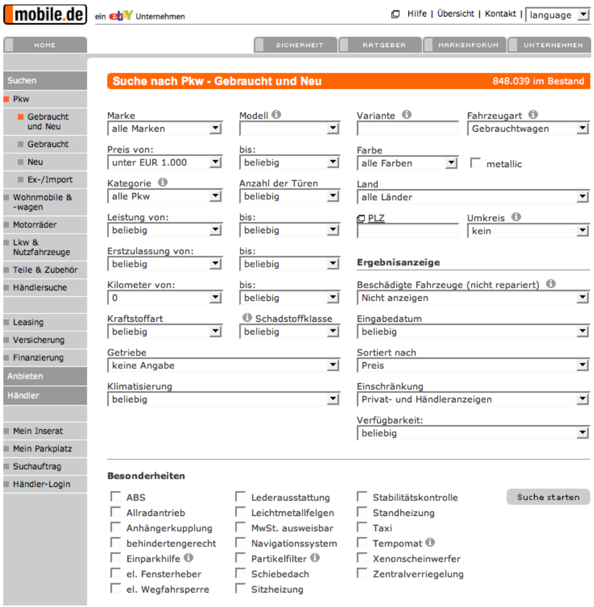
Check-out:

Rooms 1

Adults (ages 19+) Children (0-18)

Room 1 2 0

Hotels.com — advanced search



Suche nach Pkw - Gebraucht und Neu 848.039 im Bestand

Marke: alle Marken
Modell:
Variante:
Fahrzeugart: Gebrauchtwagen
Preis von: unter EUR 1.000
bis: beliebig
Farbe: alle Farben
□ metallic
Kategorie: alle Pkw
Anzahl der Türen: beliebig
Land: alle Länder
Leistung von: beliebig
bis: beliebig
PLZ:
Umkreis: kein
Erstzulassung von: beliebig
bis: beliebig
Kilometer von: 0
bis: beliebig
Kraftstoffart: beliebig
Schadstoffklasse: beliebig
Getriebe: keine Angabe
Klimatisierung: beliebig

Ergebnisanzeige

Beschädigte Fahrzeuge (nicht repariert): Nicht anzeigen
Eingabedatum: beliebig
Sortiert nach: Preis
Einschränkung: Privat- und Händleranzeigen
Verfügbarkeit: beliebig

Besonderheiten

□ ABS
□ Allradantrieb
□ Anhängerkupplung
□ behindertengerecht
□ Einparkhilfe
□ el. Fensterheber
□ el. Wegfahrsperre
□ Lederausstattung
□ Leichtmetallfelgen
□ MwSt. ausweisbar
□ Navigationssystem
□ Partikelfilter
□ Schiebedach
□ Sitzheizung
□ Stabilitätskontrolle
□ Standheizung
□ Taxi
□ Tempomat
□ Xenonscheinwerfer
□ Zentralverriegelung

Mobile.de — a complex form

There are several challenges in setting up a search interface for features. It is tiresome for users to be presented with dozens of checkboxes for individual features, especially if it is not clear how well the searchable records are enriched with these features. In most cases, the database will not contain feature values for each and every entry. If a textual search is implemented in parallel, it is more often than not unclear how textual search and filling out the form interact, especially as many feature values can be expressed in multiple ways.

These challenges lead to the notion of a query recognition process that leads to partially filled out forms. The features already specified in the textual query — modulo all different ways of specifying the feature value — would then correspond to form fields already preselected. Various interfaces build upon this are conceivable, for example presenting the partially filled out form with some entries already checked. Another interface solution could make use of the still empty form fields for a query drill-down mechanism[36]. This would form another component of an intensional query answering, yet requires integrating data acquisition components (for example focussed crawlers), search interfaces, query processing and result presentation into a new design. In Part C, queries containing feature values are examined in further detail.

The digital environment in which all these agents move allows for data of a different quality: Complete factual historic data which features ar encapsulated in the term "log". The content characteristics of "logs" will be examined unter the heading Term Spaces which is the main concept of TE-Commerce.

[36] It is even conceivable to use actual queries resolved to other features as representants.

6. E-Commerce Term Spaces

Introducing the notion of E-Commerce Term Spaces erects the main pillar to the framework of TE-commerce. Wherever terms are aggregated, Term Spaces open up. Such aggregates build up either over time — for example search query logs — or simultaneously, for example as with meta keywords extracted from web sites or a category system on a YP site.

The notion of Term Spaces bears similarities to corpora, albeit with same shifts in accent[1].

One of these shifts is the introduction of a spatial metaphoric layer which will be discussed shortly below; a further marked difference is the non-textuality of Term Spaces. A list of YP categories has no textual cohesion and hardly bears any textual coherence.

It is not surprising that these aggregates reveal properties specific to the way they have been gathered. A large list of query strings looks obviously different than a large list of meta keywords, and this difference can be measured, just as corpuses can be compared. A first division of Term Spaces segregates open Term Spaces such as Query Logs that are constantly in flux and have no intrinsic limitation to what content they might contain from closed Term Spaces such as a YP heading system that are relatively stable over time and homogenous. Until the advent of Web 2.0 and tag browsing (for example with tag clouds), it had been mostly closed Term Spaces had been displayed on the screen. Obviously it is not a trifle task to get open and closed Term Spaces to align — although this is exactly the challenge for many Yellow Page site providers that only have category information for the bulk of companies they list.

Open term spaces in TE-Commerce are created through any feeding slot, most commonly a textual form field on a web page, although other feedings slots, for example on mobile devices, are conceivable as well. Clearly, there is a relationship between how a feeding slot is labeled and the what users type in — on a Yellow page site the *Where?* on one form field opens up a geographical names term space, while the *What?* field opens up a term space of business types and specialties. However, not all users are diligent when typing into form fields and their expectations will be influenced by more than the immediate context around the search field. Even if almost identically labeled (for example the search field on Web Search Engines), the resulting term spaces often differ greatly, as many external factors such as users' sociographic background or interests influence what gets typed in.

A further aspect of term spaces anticipates what will be laid out in further detail

[1] Confer also the title of a paper by Olena Medelyan, "Why Not Use Query Logs As Corpora", online at `www.cs.waikato.ac.nz/ olena/ publications/esslli2004_mining_querylogs.pdf` `[Nov. 1, 2006]`.

in part B: as almost all terms vary on different levels, each representation is in fact surrounded by equivalent or closely related representations. This creates a new layer of term spaces: the *corona* of terms used for the same purpose, including orthographical, morphological, syntactical and semantical variations.

Summing up the main points concerning Term Spaces, one ends with four important notions that need to be explored in detail:

- Saturation and Divergence of Term Spaces
- Weirdness and Commensurability of Term Spaces
- Internal structuring of Term Spaces
- Term variations as Term Spaces

Setting out with comparing the concept of Term Spaces laid out here with seemingly similar concept, namely the vector space model in Information Retrieval, operations on Term Spaces are presented that help to explore the four issues listed above.

6.1. Beyond IR spaces

Speaking of Term Spaces evokes the most prominent model in Information Retrieval, the vector space model. Though there are certain similarities to the concept of term spaces, there are also important differences that will be laid out in this section.

The spatial metaphoric that is at the foundation of both models has enormous effects on what is to be understood as a term and what term handling is supposed to achieve. In nuce, bringing geometry into the world of terms establishes a general commensurability of terms. If every term can be placed somewhere and every operation on terms remains in this space, the world of terms can be surveyed and mapped[2]. Translating the sphere of language into numerics allows to apply mathematic approaches to terms, for example calculating the similarity of terms as the difference of angles.

The main difference between IR spaces and the notion of Term Spaces developed here can be captioned by the contrast pair geometry vs. geography. While the geometric IR models translate terms at a very early stage into numbers (after a usually very limited extent of preprocessing), the geographic Term Spaces invite for linguistic exploration. The exploration metaphor should convey, just as was laid out in the Introduction, the iterative detailing of observation. Similar to zooming in from far above to a detailed level at maps.google.com in order to get a feeling how a stretch of land looks like, the exploration of Term Spaces entail both observational work on a global and a very local level, yet always observing the actual facts, not a numeric representation of them.

The notion of multidimensionality is substantially different in the vector space model and the Term Space model proposed here. Vector spaces are multidimensional because

[2] The book *Geometry and Meaning* sets off with a historical note regarding Descartes who introduced the notion of coordinates, therefore allowing to place every object into a homogenous system based on differences. See online at `http://infomap.stanford.edu/book/ [Nov. 1, 2006]`.

of the relative mutual independence of individual terms in a document or document collection. This is connected to the understanding of documents as bag of words in standard IR approaches [3]. In contrast, the multidimensionality of the Term Space model follows from the several types of relations holding between terms.

These relations follow ultimately from the Term Spaces itself, in the form of observing usage patterns, especially distributional patterns. In Term Spaces, these observations should not operate on the level of strings, but on terms which includes a notion of term variation. For example, an E-commerce taxonomy with its various relations (narrower term, synonym, related term, instance-of, part-of etc) can contribute to a Term Space, but eventually the a priori relations have to be validated by observational data.

Reducing the dimensionality of Term Spaces differs substantially from the same operation on IR vector space models. rojections of Term Spaces on a single relation. For example, a frequency list is a projection of a term space that only preserves the unary relation "an occurence of term X was observed".

The notion of Term Spaces entails commensurability and exploration both on the level of macro term aggregates as well as of individual terms' local environments. The abstract concept of Term Spaces is concretized by term collections, which are at the most basic level just list of terms. Usually, however, term collections have an additional structure to them, which will be referred to from now on as the lines of a term collection. For example, a meta keyword repository (meta keyword lines from a crawl) consists of meta keyword lines that themselves contain terms. A query log consists of lines holding query strings which themselves might contain more than one term. Thereby, a hierarchy of two levels characterizes a prototypical term collection. Both levels can either hold sorted or unsorted elements. For example, in a meta keyword repository the order of lines and the order or terms within a line does usually not matter. A special kind of term collection is the frequency list. Here, each line has an additional column that indicates the frequency. The two-level hierarchy of term collection is a further important differences from corpora which might contain many different hierarchy levels of textual content (books, chapters, paragraphs etc).

6.2. Overview of the resources used

Here, a short overview of the corpora used in the experiments below is presented with the origin, date and size of the corpora being listed.

The main resource are several large query logs which were made available ranging from different time spans, size and type of search agents.

The first two large query logs that have been studied are from Altavista and cover each one month worth of searches in 1998[4] — further called AV1 for June and AV2

[3] See [Baeza-Yates/Ribeiro-Neto 1999]; Modern IR approaches such as Latent Semantic Indexing acknowledge the inherent dependencies between terms, yet work exclusively on the numerical representations.

[4] See also [Silverstein/Henzinger/Marais/Moricz 1998] for a study of a similar log.

for July 1998 — and one month worldwide Yahoo queries, passed through to Overture in 2002 — further called YO. These logs were available in the form of frequency lists. AV1 and AV2 have 169 million and 111 million query types, respectively; YO about 766 million query types[5] Both are predominantly English. A German log was gathered through Espotting which received at that time all queries from yahoo.de. This log, called YG below, contains 22 million query types from searches of the first three months in 2003.

Yellow Pages logs stem from two large German providers. The first Log (TG) from telegate contains all queries issued between April and October 2006 on one provider, the other log (YG) contains all queries from two months in 2005.

A specialist query log, henceforth abbreviated as SQ, was provided by Pangora. It contains queries on shopping sites from four months (May to August) in 2003.

Finally, Espotting provided the list of all German paid keywords in their system as of 2004 (henceforth abbreviated as ESP), plus information on the top 50.000 searched-for and clicked terms. Together with this, they also shared the top bid prices per term. This allows to calculate the revenue generated by term, as it follows from multiplying the bid price with the click volume[6].

Apart from query logs and paid keyword repositories, a multitude of E-Commerce resources both offline (such as telephone data CDs, among this a complete dump of German business-related telephone entries from DeTeMedien [2004]) and online (Yellow Pages sites' and business directories' categorization systems; product catalogs; various other gazeteers) have been examined and extracted. Concerning standard language prose corpora, both the German and English wikipedia have been used, as these are available as a complete database dump[7].

The two AltaVista logs were kept divided in order to provide a sample of two closely related logs. As can be seen, these two are indeed very similar to each other, thereby strengthening the hypothesis that these figures are indeed a signature of a log that discriminates it from other logs.

A large collection of meta keywords gathered by crawling 500.000 German business homepages provides an additional resource that is very rich in TE-Commerce relevant c vocabulary (MKW-DE).

6.3. Counting terms: Tokenization and Lemmatization

The motto of *divide and conquer* will lead the way not only through the subsequent sections, but stands as main principle of TE-Commerce term spaces. Starting with dividing term spaces into single terms, it will become apparent how even simple operations such as tokenization depend on a model of term variance.

[5] For practical purposes, a subset of YO was used that went down to frequency 3. This subset YO-3, consists of 142 million query types.

[6] Obviously, some share then goes to the content provider, but as the ratio is usually constant, the multiplication of bid price with bid volume allows a relative ordering of what terms are worth.

[7] At `download.wikipedia.org`.

Before one starts to worry about the sheer number of query types, some simple methods can be applied to detect and fold identic queries. On a very basic level identity is here to be understood as preservering meaning, syntax and wording of a query. Under this assumption, queries represented by strings which only differ through orthographic variance has to be considered as identic. Among such variation on a basic level:

- Usage of non-alphanumeric chars, such as ? ! % ”
- Usage of diacritics and transcriptions
- Casing conventions (upper- or lowercased characters)
- Spacing conventions (writing words together with spaces, hyphens or without spaces)

Two question will be treated with respect to this. Firstly, to what extent can non-alphanumeric chars, casing and spacing be safely discarded without losing information or leading to mismatches. Secondly, what extent of compression is achievable through folding term collections by applying a normalization with regards to non-alphanumeric chars, casing and spacing.

6.3.1. Non-alphanumeric chars

Non-alphanumeric chars comprise all chars outside the range from $AtoZ$ and $0to9$. It is obvious that the following findings only make sense in Western languages. Furthermore, the examinations made subsequently are restricted to English and German.

One way to treat non-alphanumeric chars in terms is simply discarding them. In contrast to chars with a delimiting property or punctuation chars — such as hyphens — that will be treated in a the section below on spacing, discarding skip chars means simply removing them. For example, the plus (+) is quite frequently appearing in query logs, and it is not always clear what kind of operator the user expected when entering it. As today’s generic Web search engines are all based on an AND-connection between query terms, using the plus is in many cases redundant. It may thus seem feasible to remove all instances of it. However, some bona fide terms are written with non-alphanumeric chars and deserve a special treatment, see below.

There are two main reasons why alphabetic chars outside the range of standard Latin letters deserve a special treatment. The input interfaces are not for all cultures and all computer configurations 100% reliable today outside the range of 7-bit Ascii chars. In addition, index data acquired through Web crawls often displays inconsistent or even invalid encodings. Although utf-8 helped to alleviate this situation greatly, still every thinkable and unthinkable combination of header and encoding can be found on the Web.

At least if addressing an English market it is tempting to restrict the range of chars to to 7-bit ascii chars and either remove or map all other chars. One common replacement is to map diacritics either to the standard letter or diphthongs (ä → a, ä → ae). While these rules will only be applied in a limited number of instances in English-speaking countries, in German queries umlaute occur in every 13th term in general. Some confusion examples between bona fide terms can arise when automatically replacing all such occurrences:

Conducting both the removal of skip chars and the replacement of diacritics led was done on YG, the German Web query log. Below is a list of top frequent terms that include non-alphanumeric chars. Removing these chars perturbates the term and may lead to confusion with other words lacking the distinctive meaning:

- 1&1
- t.a.t.u.
- d&w
- H&M
- harman/kardon
- c&c generals
- 0: 0
- a&p
- c&a
- AC/DC
- k&m
- c++
- essen & trinken
- c't
- ver.di
- j.lo
- villeroy&boch
- l'tur
- such&find

For example, if removing the dot in *ver.di* the name of this German labor union becomes conflated with the Italian composer. If removing the pluses of *C++*, the two distinct programming languages *C++* and *C* become mixed up[8]. See Part C for a figures on the folding ratio achieved through removing these terms.

6.3.2. Casing

The search interfaces of today's largest Web searches are without exception case-insensitive. Putting a query in lower case, upper case or any mixture does not change the number or selection of results produced. It is not possible to distinguish in a search between different casings and most users would be puzzled if a search device reacted case-sensitive to their queries.

In the process of acquiring and analyzing term collections, however, it may well make sense to leave the casing as it conveys information about proper names, at least in English and — to a lesser extent — in German. An inquiry into the effectiveness and reliability of this method for the detection of proper names is done in Part B, Lexical units.

The process of collecting case variants can be done by once passing through each line, and adding it to the values of an associative array which keys are the lines in lowercased form. Lines which differ only by their casing are thus kept in the same place. If this approach is ruled out by the size of the term collection, a related method puts out a two column table, with the lowercased form of the line in the left column and the original form in the right column. Afterwards, the table is sorted and aggregated, given that after the sorting all the casing variants are kept in one block.

One of the applications of gathering casing variant is to repair casing. For purposes of displaying terms to the user, finding the proper casing to a term is often a helpful routine. On the base level, it can be achieved by looking for the most frequent instance of casing for a term. However, this only works for texts that are edited according to orthographic rules such as newswire. It does not yield any results if the creation of the examined text does not follow such rules. For example, eBay occurs in the German YG log (see above) with different casings as follows:

[8] This effect can be seen in practice on seekport.de [as of Nov. 1, 2006]. The algorithmic results for the queries *c programmieren* and *c++ programmieren* are identical.

97804	ebay
3813	Ebay
1739	eBay
1245	EBAY
146	EBay
70	eBAY
32	ebaY
16	ebAY
4	eBAy
4	EBAy
3	EbAY
1	EbaY

Here, it is striking that the by far most frequent variant is the all-lowercased form. The correct one appears only in third position. If the query log is examined in full, almost all conceivable casing variants can be found, some obviously not typed in intentionally (for example *EBAy* with the last letter lower cased).

Looking for the most frequent occurrence of a casing variant has to go beyond single words, given that the casing of some multi-word units differs from the casing of its single word segments. An example of such an behavior is *New York* with *new* written in lowercase if occurring isolated. In addition, some ambiguous words have different casings for their different meanings, for instance *apple* (the fruit) vs. *Apple* (the brand).

6.3.3. Spacing, Hyphenation and other forms of Delimiting

In much the same manner than described above for casing variants, also spacing variants can be gathered. Here, the normalized form is achieved by removing all spaces, hyphens and other delimiters. For example *star-wars iii* becomes *starwarsiii* in this normalization and thus equals the normalization of *star wars-iii*.

While the occurrence of variants in spaces and hyphens may in some cases be the effect of mistyping — much like any other char — it is apparently a sign of intentional use if these variants appear about the same number of times. Some examples from YG illustrate cases where their is no predominant usage of spacing or hyphenation, i.e. the variants in spaces and hyphens occur approximately the same number of times, indicated in square brackets:

suchmaschinen top ranking [347]	suchmaschinen topranking [348]
antiagingsubstanz [171]	antiaging substanz [171]
musterbewerbung [733]	muster bewerbung [729]
zimmer reservierung [172]	zimmerreservierung [172]
e-cruiting [190]	ecruiting [191]
oldtimer-Börsen [188]	oldtimer Börsen [189]
sony dcr-pc 120 [836]	sony dcr-pc120 [844]
aerolloyd [337]	aero lloyd [340]

Examples of three variants that occur about the same number of times:

barbiepuppen [27]	barbie puppen [31]	barbie-puppen [27]
citychat [200]	city chat [181]	city-chat [215]
design hotels [15]	design-hotels [17]	designhotels [15]
firewirekarte [9]	firewire karte [9]	firewire-karte

How many instances of hyphenation can be traced back to two words, how many to words written in one

While detecting spacing variants is not very a challenging task, it is much harder to consistently normalize lines with regards to spacing. A baseline algorithm replace global strings through the most frequent spacing variant. However, if the term is contained in a longer phrase, it is not guaranteed that a consistent replacement takes place. For example, if the isolated occurrence of *note book* is replaced by *notebook*, it is not guaranteed that the line *samsung note book* will be correspondingly treated, as *samsung notebook* might not be in the term collection or at least not with a higher frequency. Given that the tail of the term collection consists of many longer terms there will inevitably occur inconsistencies in the replacement mechanism if only complete lines are taken into account.

If the term replacements are integrated into a transducer, it is possible to run through the complete term collection and replace all occurrences in longer units. However, conflicting replacements might occur that can lead in the worst case to cyclic references, i.e. a chain of replacements $T_0 \rightarrow T_1 \rightarrow T_n \rightarrow T_0$. Consider the following replacements:

> *starwars* → *star wars* (assuming that the variant written in two words occured more often than as one word)
>
> *star wars fanshop* → *starwars fanshop* (assuming that in this case, the right variant occured more often)

Here, the effect of the second replacement produces the starting condition to apply the first replacement again which then reverses the effect of the second replacement. While detecting such cyclic references can be done easily by applying the replacement rules on the left hand side of the the same replacement rules, it still remains an issue what replacement rule to choose from in a set of conflicting rules. An even larger issue that cannot be solved by simply counting the number of occurrences for spacing variants is the consistent treatment of related words and word forms. For example, if the singular form *note book* is replaced by the variant without spacing, its plural form should also be replaced in the same manner.

Finally, a special case occurs if there is no isolated occurrence of the spacing variant that needs to be normalized. Consider for example the following two lines

> find wal mart stores
> walmart shops

and suppost that *wal mart* and *walmart* do not appear as isolated terms (this is not inconceivable for terms from the frequency tail of the term collection). Here, finding the replacement rule would require looking at tuples of words and detect whether they also appear written as one word. However, this will drastically increase the number of conflicting replacement rules.

As a preliminary conclusion it can be noted that while it is feasible to detect the numerous variations based on spacing and hyphens, inferring a consistent spacing conventions requires more than just counting frequencies of spacing variants. It is therefore necessary to distinguish the scope of application for term handling routines. If the term handling is used to broaden the recall in search applications, the spacing normalization is a viable way. Note that such a normalization can hardly be done on-the-fly. While removing spacing and hyphenation is of course easy, finding the right break in a string consisting of two words written together requires to keep track of terms. For other applications, such as a keyword normalization process that require a consistent handling of spacing and hyphenation, edit rules have to be formulated which require manual intervention. It is not enough to decide on whether to remove or pertain all spacings. Besides genuine variance on spacing, there are also clear cases of spacing errors or at least very unusual spacing habits in query log data that call for a repair process (for example *on line* instead of *online* or separate words written as one, such as *berlinhotel*).

6.4. Multi-word unit recognition

The single most tempting separator of prose text in western alphabetic systems is the blank[9]. It is tempting because white spaces are easily observable both for humans and for machines. However, in many cases dividing at white spaces breaks up units that belong together. There is no synchronic ground that allows to split up *New York* into two units. Writing it in two words fits into analogous patterns of place names (especially names for places that were once colonies). It is a orthographic convention, ingrained deeply enough to make the string *Newyork* look startling. Apart from the surface conventions, it should be treated as one unit and prevent any comparison to free occurrences of *new* and *york*. However, not all cases are equally clear. It is necessary to discriminate between different types of MWUs and set up methods to detect them and deal with them.

6.4.1. Definitions and Purposes of MWU detection

The recognition of multi-word units (MWU) is a crucial prerequisite for any further term space exploration. A multi-word unit is a lexical item consisting of more than one whitespace or hyphen separated word[10]. Several linguistic phenomena are responsible

[9] This hints at the long tradition of the opponency word-based vs. sentence-based linguistics, see also [**?**]. As a recent counter-position to sentence-based linguistics, see also the study [Stainton 2006].

[10] There are numerous names for this phenomenon — collocation, non-compositional expression, n-gram, sticky phrase, lexical atom etc —, but these are in general either only covering a part of

for MWUs[11]:

- Named entities, such as *New York* or *Caroline Beil*
- Compounds, such as *usb stick* or *junior high school*, either fully motivate, partly or fully demotivated
- Syntactic freezes, such as *kick the bucket*[12]
- Frozen modifiers (*strong tea*) and support verbs (*spiele zocken*)

Various degrees of semantic compositionality have to be separated in the treatment and definition of MWUs. One group of MWUs cannot be decomposed in any way, for example *New York* or *area codes.* For those, the blank should really be treated as a part of the complete descriptor. It does not bear any delimiting value. Its constituents do not contribute their normal meaning to the meaning of the MWU. A further group has at least one component that preserves its meaning in a free distribution. In *usb stick* for example, *usb* contributes in the same manner to the meaning of the compound than it does in other compounds or in isolated occurrence. The second component, *stick* needs to be shielded because it has little to do with the meaning of *stick* in other compounds (*walking stick*) or as isolated term. Finally, a third group contain elements that build up compounds, but do not exhibit a different meaning when being combined. For example, in *usb cable* bith the meaning of *usb* and of *cable* are preserved. These units do not need any shielding at all, but it makes sense to keep track of them.

Common tests for MWUs use their non-compositional semantic [13] and syntactic encapsulation of their parts, which prevents modification, pronominalization and coordination of their components.

Among genuine multi-word units in a query log are to a large part named entities (*New York, Bill Clinton*). Several techniques to extract such units will be presented, using either linguistic, statistical or both approaches. Consecutively, an approach based on candidate term processing and verification methods is presented.

Operational tests comprise associative metrics, distributional metrics, linguistic features, heuristics and the use of gazeteers. The specific information contained in a query log (for example phrase operators) and their potential benefits to the MWU detection process are also discussed. In the processing pipeline, MWUs are going to be protected (for example by inserting underscores instead of whitespaces) before the subsequent steps follow.

Summing up the problematic issues that are related to MWUs:

MWUs or go beyond MWUs in including groups of lexical items that occur regularly with each other. See also [Evert/Heid/Lezius 2000].

[11] An overview is provided by [Blanco/Guenthner 2004].

[12] Cf. the papers of the conference Collocations and Idioms 1: The First Nordic Conference on Syntactic Freezes, see online at `http://cc.joensuu.fi/linguistics/idioms2006/index.shtml [Nov. 1, 2006]`.

[13] [?] describes what he calls collocational expressions as “sequences of words whose unambiguous meaning cannot be derived from that of their components, and which therefore require specific entries in the dictionary”.

- Co-occurring forms need not be MWUs, and some MWU occur only infrequently

- It is not always clear when to split up a supposed MWU into smaller MWUs and when to join supposed MWU into a bigger unit.

- MWU often have variant forms, among these also reduction forms

6.4.2. Two words and larger words-MWUs

The metrics and heuristics presented below all work basically as a filter that takes all bigrams as input and delivers MWU candidates as output. As in reality MWUs are not restricted to bigrams, a routine that allows to build up larger units.

In general such a routine will build upon bigram MWU candidates and enlarge them by looking at their adjacent words. A simplistic first approach could make us overlapping MWU candidates. If for a sequence of three words A, B, C both A, B and B, C have been extracted as candidate MWUs, then A, B, C could be considered also to be a MWU candidate. For example, if the MWU detection delivers *New York* and *York Knicks* as candidate MWUs then the sequence *New York Knicks* could also be considered to be a MWU. Naturally, some filter for A, B, C needs to be implemented, as otherwise many errors will occur. For example, *red hot* and *hot pursuit* would be wrongly concatenated to a larger MWU if no filtering is in process.

A second approach also build up MWUs iteratively. However, it recalculates the MWU detection once a MWU is detected by treating it as one word. If *chilli pepper* is detected as a MWU candidate in a first pass, it is shielded for a second pass and treated as one word, just as if it blank were replaced by an alphabet char.

Both approaches are problematic in the respect that they require that parts of larger MWUs significantly co-occur and pass the hurdle of being accepted as bigram MWU. This hurdle might prevent that valid MWUs are recognized. For example, the bigrams *deutschland sucht* and *den superstar* are both not necessarily good MWUs — their concatenation, however, is.

One great advantage of using a query log is that the terms are already extracted from their context. Instead of testing bigrams, trigrams and larger n-grams on a prose text, it is feasible to restrict the MWU detection to word sequences that at least once appear as full queries. Therefore, the most convenient way to grasp larger MWUs when a query log is available is to test only complete query lines consisting of more than one word.

Reversely to the process of adding words, it is also necessary to establish a decomposition procedure that breaks down MWUs if they consist of other already known MWUs. For example, the candidate for a MWU *microsoft windows usb stick* should be broken up into *microsoft windows* and *usb stick.* In general, a MWU candidate that can be decomposed into already validated MWU should be rejected[14].

[14] See also [Martinez-Santiago/Montejo-Raez/Urena-Lopez/Diaz-Galiano 2003].

6.4.3. Baseline: Frequency counting

The baseline heuristic for extracting multi-word units from a list is simply counting the number of occurrences of term tupels and set a threshold above which these tupels are considered to be terms. A line consisting of more than one word is broken up into pairs of adjacent words, triples of words and all n-tupel including the complete input line. All these MWU candidates are then stored and their frequencies added up. If a MWU candidate passes a given threshold, then they are accepted as MWU.

By simply counting the token frequency, these word sequences containing more than one word are the top frequent:

22191 in der
13545 was ist
13489 in deutschland
12197 in berlin
11028 windows xp
9617 of the
8728 bilder von
7355 new york
6889 windows 2000
5470 in münchen
5350 in den
5236 in hamburg
5189 and download
4976 an der
4818 für die
4787 in and
4724 and in
4657 and in and
4651 hotels in
4496 in the

This is certainly not a convincing result. If counting full query lines instead of counting subqueries (i.e. sequences of words that appear within a query line), the results become much better:

online casinos
last minute
gay pics
gay cam
sex shop
telefonsex livecam
private krankenversicherung
gay porno
hardcore sex

telefonsex mit livecam
online spiele
las vegas
online games
black jack
internet kasinos
online kasino

In this lists, not only the improvement in cleanness is striking, but also the variations of MWUs (for example *online kasino – online casinos*).

Even from this short list the need for a mechanism that decomposes and adds words on the left or right hand side to candidate MWUs become apparent, for example in the lines *last minute reisen* vs. *last minute* or *herr der ringe* vs. *der herr der ringe*. The former example illustrates two different degrees of stickiness in MWUs, given that *last minute* cannot be meaningfully separated (last minute bookings do not normally happen in the last possible 60 seconds) and *last minute reisen* is a frequent transparent compound. The later example shows a typical behavior of MWUs in that it a reduced form of a MWU occurs.

Refined frequency tests can be based on the notion that the frequency of a MWU has to be considerably higher than what could be excepted from the frequencies of its part if the parts appear fully independently. If $f(AB) > f(a)$ and $f(AB) > f(b)$ the word combination in question is almost always a MWU if its frequency is high enough.

Another effect of the syntactic freezing that compounds undergo is that their order of their components is largely fixed. This should allow for an additional MWU test metric, based on comparing $f(AB)$ and $f(BA)$. If one ordering of terms is much more preferred over the other, this is a indication of a MWU. This method is especially helpful in query logs, as they do not in general follow the standard syntactic patterns of English and use AB in much the same fashion than BA if AB is not lexicalized. The phrasing operator (“AB”) in a query log may also give hints on MWUs.

The following list shows the frequency AB (left column) and the individual frequencies for A and for B as they appear in isolated positions (the numbers after the Q in square brackets). In the line *barnes and noble* for example, the frequency of ABC is higher than that of every part in isolated position which underlines its status as genuine MWU. In the last two lines, one part of the two word combinations has an higher frequency than the two word sequence:

263815 [search = Q949277] [engine = Q51968]
262177 [mapquest = Q1781317] [com = Q345370]
261610 [barnes = Q24972] [and = Q19024] [noble = Q5116]
261054 [at = Q45881] [t = Q116367]
260913 [corporate = Q9789] [kit = Q7336]
260055 [swiss = Q10290] [cheese = Q43244]
258562 [capital = Q12879] [one = Q11649]
257513 [hgtv = Q378225] [com = Q345370]

255659 [big = Q32444] [tits = Q343774]
255078 [ja = Q6524] [rule = Q1362]
254733 [birthday = Q109428] [cards = Q503981]
254156 [us = Q31480] [travel = Q1943798]

A further observations that can be made from these numbers is that in semantically transparent units (such as *birthday cards* or *us travel*), the term with the higher frequency contains the core meaning of the unit. For example *swiss cheese* is still primarily a cheese (relating a search for *swiss cheese* to cheese-related websites is often a more relevant result than relating it to a general Swiss-related website). The same holds for *birthday cards*, *us travel* and also *search engines*, where the head word appears on the left hand side.

This last observation will be deepened in Part C by applying the notion of heads and containers to queries.

6.4.4. Association metrics

For the detection of MWUs, association metrics prove to be very helpful. Association metrics can be based on several models, for example a statistic or an information theory model. In all cases, the different association metrics can be calculated for a pair of words (a,b) based on four figures: the number of occurrences of a, b, $aandb$ and the total number of words (N) in the sample examined. From these four figures, all fields of the contingency table can be deduced.

Note that association metrics do only rank bigrams — it is still necessary to set a threshold to what bigrams are accepted as MWUs. With respects to evaluating for the different association metrics, this means that either the threshold is also part of what is evaluated or that those cases are taken into account when a valid MWU is ranked below an invalid one.

Regardless of what association metrics — Dice coefficient, Log-likelihood, Mutual Information, Chi-Square or Fisher's exact test — bigrams for which the elements never occur outside the bigram will always pass the metrics as very high ranked candidates for MWUs. Some metrics depend on the total frequency of the bigram as well, but their main difference is how they treat bigrams with low numbers in the contingency table. This may lead to results as being top ranked by the association metrics which are valid MWUs, but come rather as a surprise.

To illustrate this point, here is the list of top-ranked bigrams resulted by Mutual Information, as tested on YG:

jeet kune
malice mizer
breast expansion
stages pratiques
jolene blalock
nux vomica

dragan stojanovski
waris dirie
karlovy vary
estee lauder
grim fandango
vel satis
darth vader
severus snape
jeremias gotthelf
barbra streisand
south africa
idar oberstein
tonya harding
soutien scolaire
hülya avsar
divine divinity

While a sufficiently large corpus allows to detect many valid MWUs through associative metrics, there is no systematic discrimination between words that just happen to occur together — for example common collocations like *cheap flights* — and true MWUs. Given that all metrics are just variations on the numbers in the contingency table, it is clear that if two pairs behave very similar with regards to frequencies, they will be assigned the same value regardless of the association metrics. It is possible to extract all pairs have very similar values for $f(a)$, $f(b)$ and $f(ab)$. Going through this list will very likely yield all different pair of terms — this effect is demonstrated in Part B, "Semantics Matches".

6.4.5. Linguistic filters

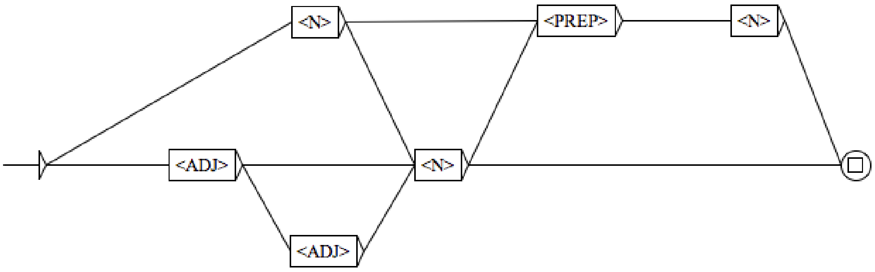
Linguistic filters can be used in the MWU detection process as a constraint on what is allowed to appear within the MWU and what is allowed to appear adjacent to a MWU. As a subcase of the first case, filtering can be applied on the words appearing at the beginning or at the end of a MWU candidate.

In many approaches using a linguistic filter, a set of stopwords is deployed that must not occur at the beginning and at the end of a valid MWU. Among these stopwords are usually conjunctions, articles, particles, prepositions, pronouns and some high frequent adjectives and adverbs[15].

Just by using such a filter and the baseline algorithm that counts all occurrences of sub-phrases in a line (see above), the top ranking results are almost without exception valid MWUs. Based on the same YG log used above and filtering out only closed part-of-speech classes, the following MWU candidates are delivered as the top frequent ones:

[15] See also details of the KEA algorithm, [Witten/Paynter/Frank/Gutwin/Nevill-Manning 1999].

If an electronic dictionary with broad coverage is available, MWUs can also be filtered by specifying which combinations of parts-of-speech build up MWUs.



A simple Unitex graph with English MWU patterns

The main advantage of a linguistic filter is to ensure that all extracted units are not only known as textual strings, but also in their morphologic behavior, because their parts have been found in the electronic dictionary. This allows to discern MWU word forms such as singular/plurals, but also to generate these word forms.

6.4.6. Repetitive metrics

Looking at the definition of MWUs, it stands to reason that the distributional patterns of MWUs may serve as a sufficient indication. If a MWU candidate appears within the similar contexts than single-word terms, a strong indication to its unithood is provided. For example, the MWU *Las Vegas* appears in the following contexts (or containers, as will be explained in Part C) in the YO Log:

643947 las vegas
216520 las vegas hotels
96247 las vegas hotel
69370 las vegas shows
62892 las vegas casino
56122 las vegas hotel discount
51867 travel to las vegas
46841 las vegas show ticket
46625 las vegas show
46056 las vegas accommodation

Comparing this to the distribution of *hotel* makes it obvious, that *Las Vegas* behaves like many single-word terms do:

86836 new york hotels
86022 paris hotels
85632 london hotels
74693 discount hotels
65386 hilton hotels

60173 orlando hotels
58364 vancouver hotels
216520 las vegas hotels
56094 lasvegas
54268 vegas hotel

Although one can also find *vegas hotel* in this distribution (see last line), it would be highly unusual, if the most frequent context containing *hotel* was *las X hotels* instead of *X hotels*.

6.4.7. Adding gazetteers

Adding gazetteers to the MWU extraction process makes sense if the accuracy of the gazetteer is guaranteed and if it helps to cover cases that have not already been take care of. Such resources are not abundant in number, however there are ways to set them up using simple heuristics. The resources tested here are different language versions of Wikipedia and list of product titles as they appear on Web price comparison (shopping.com, billiger.de) or evaluation sites (epinions.com, ciao.de). Another benefit of gazetteers of this kind is that they provide also variant forms of MWUs, such as *Bill Clinton vs. William Jefferson Blythe III vs. William Jefferson Clinton*[16]

By intersecting the article titles from the German and English Wikipedia it can be expected to grasp at least all prominent named entities such as persons, organizations, place names and products. Another filtering step can be administered by checking the query log frequency and only pertain those article titles that appear at least once. This helps to remove Wikipedia-typical article titles such as *Liste aller niederländischen Gemeinden AL*.

6.5. Term space operations

6.5.1. Indexing

Indexing is necessary in order to allow operations such as determining co-occurrence or extracting repetitive patterns in a reasonable time even on large term space data. A recognition of term units is considered as prerequisite. In addition to this, determining co-occurrences or repetitive patterns renders a grouping of terms into larger units necessary. Such windows of co-occurrence or repetition are exemplified by meta keyword lines, sentences from corpora or query strings. This two-fold division of the complete term space into term groups and those into individual terms results in a representation of the term space as a sorted set of all the sorted sets of terms.

The indexing process should allow to obtain all lines in which a term occurs, regardless of the size of the term collection. The largest term collections examined here are in general too large to be kept in memory, requiring some sort of indexing.

[16] See `http://en.wikipedia.org/wiki/Bill_Clinton [Nov. 1, 2006]`.

For the purposes needed here, a binary tree is a feasible data structure[17]. The baseline approach of just adding positions in the original file to an index of terms soon becomes infeasible, as the top frequent terms have enormously large strings as values.

A remedy is to single out those terms that appear in too many positions. Through setting a threshold t it can be parametrized how many occurrences are needed before a term gets removed from the main index and inserted in a top index.

The positions of lines containing these terms are printed in a separated file (the top-list), with all the lines containing one top-frequent term grouped together. For each of these top-frequent terms the starting position of its block and the length of it are stored in a separate index[18].

The look-up procedure thus looks like like below:

```
if term is in top index then
    $p \leftarrow$ start position
    $l \leftarrow$ length
    $S \leftarrow$ Retrieve from top-list(p,l)
    for $S$ do
        Retrieve from main index.
else
    if term is in main index then
        $S \leftarrow$ vector of start positions
        for $S$ do
            Retrieve from main index.
    else
        fail.
```

Procedure `Look-Up in index structure`

For example, if the term *hotel* has to be looked up, the process would start with querying the top index. Assuming that it would find an entry for *hotel* (consisting of the starting position and the length of a block), the block of the top-list is retrieved. This block contains all positions of lines that contain *hotel* in the original data file.

Although the top index is separated from the main index, ultimately all terms produce a list of positions for lines in the original data. This makes it possible to perform a quick unison and intersection of these positions, as they refer to the same original data file. Intersecting the positions of two terms can be done by remembering all positions for the first term and then filtering the positions for the second term by this set (for example through a hash).

The main advantage of this approach is that it provides a fast and relatively scalable indexing method that only uses standard and lightweight database modules. Both the indexing and the retrieving of terms can be realized with only few lines of code. A

[17] Available for example through the Berkeley DB suite.

[18] Technically, the same binary tree can be used. It is only needed to make sure that entries referring to positions in the original data are kept separate from entries referring to position and length of a block in the top-list.

disadvantage of the approach that prevents the usage of extremely large resources is that the top-list becomes very large. It might even become bigger than the original list, as some terms appear in more than one line. A remedy is to split the index into term prefixes — for example one index for a, one index for b and so on. As the term prefix is immediately discernible in a term, the look-up time is only slightly larger.

Optimization can be done by fine-tuning the threshold t that separates the TOP-index from the main index. Increasing t will lead to a bigger binary tree, but a smaller TOP-index, decreasing t will lead to smaller binary tree and a larger TOP-index. Lower values for t will also speed up the creation of the index, as the lines for the TOP-index are just flushed out in plain text on the hard-drive.

The aim of the indexing process used here is to provide a fast and scalable access to large term collections that have a frequency list format. Assuming that the term collection is divided into two columns, a column holding the frequency and the other one the observed unit (consisting of one or more terms), the index should primarily be fit for retrieving all occurrences of a given term from the collection. As the term collections can have an enormous size, it is necessary that this step does not require a sequential pass through the complete list. Therefore, some sort of indexing is obligate. Apart from retrieving lines through terms, the indexing should also allow for fast determining of co-occurrences, not only the co-occurrences for one term (the terms that appear along with it in the same lines), but also for list of terms.

6.5.2. Index operations

Once the index is built as described above, two basic operations can be conducted. One is to look-up all positions for a given index term, the second operation retrieves the content stored at these positions. The processes switch between the original flat text file and the two index tables. Three applications of these basic operations will be explained in more detail, namely co-occurrence, context listing and context instantiations.

Co-occurrence starts with a given term or a set of terms. It looks for all terms that occur together with the start term or terms in a given window of co-occurrence. In the term collection model described above the window of co-occurrence is usually one line, for example one line with meta keywords or a query. For a term t the creation of a co-occurrence starts with looking up all offset positions in the index, retrieve the lines starting at these positions and tokenizing them. In the case of co-occurrences for more than one term, i.e. looking for terms that co-occur with a set of terms

The context operation extracts all contexts for a given term. For example, a context for *hotel* could be *cheap _ new=york*. The context operation starts by looking-up all positions of lines in which t occurs. These lines are read in and t is substituted by a regular expression through a placeholder. The resulting list of contexts can then be ranked by frequency or other metrics.

The instance operation is needed for getting all insertions for a given context. For example, the context *cheap _ new=york* will have insertions such as *flight* or *hotel*. The instance operation starts by splitting the context into terms (omitting the underscore) and then obtaining the positions in which all of the terms in the context line appear (i.e. performing an intersection on the positions). Once these positions are determines, lines are looked-up and the instances can be extracted by a regular expression that specifies what kind of instances are allowed (single words, multi words, words with other delimiters etc).

6.5.3. Folding Term Spaces

Assuming a term collection consisting of lines which have one or more terms in it. What is the minimal set of terms so that every line has at least one term that is in this set? In other words, how can a complete coverage of lines be achieved with a minimal number of terms? Two solutions to this algorithmic problem are presented here, one an exact solution meaning that the resulting set can be proven to be a minimal set (naturally, there are often more than one of such minimal sets), the other one an easy and fast approximation.

Finding such a minimal coverage by terms or morphemes helps to fold the Term Space. If only meaningful units are allowed as elements in this minimal set — what are later called Sense Morphemes — finding this minimal set does indeed imply that at least something is known about all lines in the term collection. This process will be called *CoreMaker*.

For example, the two minimal covering set for these lines

k m n
o p
l n
l
k j

are $\{l, k, o\}$ and $\{l, k, p\}$.

This problem can be described as traversing through an directed graph. In a brute force approach the lines have to be completely worked through before deciding on the minimal set of terms. When passing through the lines, there are as many alternatives to move on as there are terms in the next line. By this, a large tree structure builds up that represent paths through the lines. When arriving at the end line, the different paths are examined and ranked by the number of unique terms they contain. The paths with the lowest number form the minimal covering set.

A refinement of this approach looks first for isolated terms, i.e. terms that make up one line. All these terms have to be in the minimal set. All lines which contain one of these terms can be safely discarded, as these lines are covered in any case.

Then, sort the terms within each line alphabetically and sort the lines according to their first term also alphabetically. This ensures that from a line that starts with the term t to the end of the list, no terms s that are prior in alphabet to t can occur. Such a term cannot occur in the line starting with t, as all other terms of the line come after t in the alphabetic ordering. All further lines start either with t or a term that comes after t in the alphabetic ordering, therefore they cannot contain any term that comes prior to t.

After sorting the lines and removing all terms in isolated positions and all lines that contain those the example lines look like follows

j k
k m n
o p

When traversing through these lines from top to bottom, a path build up by adding terms that are necessary to cover the lines from the bottom to the current line. For example, in the first line, both the path j and the path k are possible. At the second line, the paths j m, j n, j k, k m, k n and k result. If moving on to the third line, it can be observed that no term that comes before o in the alphabetic ordering can occur in the rest of the lines. This allows to prune paths that only differ by terms below the current threshold o. The part of the path that contains terms below the current threshold will be called the path prefix. In the example, all lines only differ by terms prior to o. It is only necessary to keep the shortest of these paths, because all subsequent lines will not change anything in the path prefix. There is no chance for the other paths to outrun this path, in this case k. Therefore, the two resulting paths are l k o and l k p (the terms in isolated position have to be added to the paths)

Calculating all possible paths for large lists becomes cumbersome soon, even if the path pruning is used as described above. Furthermore, for many applications it is not necessary to list all minimal covering sets. They do not even have to be strictly minimal — the quality of the covering is much more determined by the stop-terms that are discarded when looking for the minimal covering sets.

Based on that, a simplified algorithm just starts — after the terms appearing in isolated positions are removed — by the term with the highest number of lines in which it occurs. The lines in which it appears are removed. This procedure is recursively applied to the list, until no lines remain. If two terms appear in the same number of lines, one term is arbitrarily chosen.

The resulting set of terms can then be ranked by the number of lines in which they appear. In conjunction with the removal of stop terms, this allows to extract the building blocks out of term lists and deliver them in a ranked order.

6.5.4. Dividing term spaces

One of the main goals when exploring Term Spaces is to divide them in such a way that either tasks on Term Spaces are facilitated or new insights on them can be obtained.

A useful operation is the division of a Term Space into different frequency spectra. An intuitive approach is to bucket all terms into very, medium and low frequent terms. Given that more than half of all unique terms do occur only once, it is clear that these buckets cannot be of equal size with regards to the number of unique terms they hold.

One division principle is to divide the frequency of the most-often occurring term by three and sort the remaining terms into these buckets. However, this will leave the bucket with very-frequent-terms almost empty compared with the other terms.

A more natural division principle divides the full frequency range by a logarithmic scale. For example, if the highest frequent terms appear about 10^6 times, the first bucket could contain frequencies ranging from 10^0 to 10^2, the second bucket those between 10^2 and 10^4 and the third bucket the rest. Applied to YG, this would lead to the following distribution of types and tokens into the three buckets:

Terms can not only be divided by frequency ranges, but also based on other numeric values, Examples are the *weirdness* of a term, i.e. the ratio of occurrences in texts of a specific domain versus that in standard texts, or simply the length of terms.

Besides sorting a Term Space into different frequency spectra, another important division principle is the separation between simple and complex units (complexity may mean different things in different Term Spaces, though, ranging for example from non-monomorphemous words to query lines containing more than one phrase). Finally, Term Spaces can be divided into a known vs. unknown terms in which case "known" could mean a term is listed in a lexicon or can be decomposed into known units.

6.6. Comparing and combining term spaces

6.6.1. Micro- and Macro-comparison of Term Spaces

The concept of Term Spaces, as was laid out, can be applied both on a micro and on a macro level. On the micro level, individual terms and their surroundings are taken into account. Following this line, a micro-comparison is the comparison of two individual term's positioning within a Term Space. This includes comparing the distribution in which two terms appear and the terms' variations.

For this purpose, a notion of context sets for a term t needs to be introduced. A member of a context set C for a term t is a line (or its equivalent for a given Term Space) that contains t and has t replaced by a placeholder. For example, in the line *hotel new york*, hotel has the context *_ new york*. All the different contexts in which a term appears form the set of contexts for this term. Such a set might consist of a flat list of contexts, but might also contain weighted contexts that discriminate between the number of occurrences.

One way to compare C_1 with C_2 follows the logic of the Dice coefficient. It counts the number of shared contexts (the intersection of elements in C_1 and C_2) with the number of all contexts (the union of elements in C_1 and C_2). This can easily be extended to weighted lists of contexts. A baseline micro-comparison of weighted contexts could work by counting context tokens instead of types. For example, considering the following terms and their contexts

For detecting common variations, it is necessary to introduce a measure that penalizes uneven a skewed distribution. For example, assuming that a mechanism for spotting permutations in lines is implemented (see below, Part B, "Morphological-syntactical matches"). It delivers back lines with variants with respect to the ordering of terms and indicates the frequency of variants in square brackets, such as

arzt hals nasen ohren [1] – hals nasen ohren arzt [309]

or

heizung und sanitär [88] – sanitär und heizung [24]

Although the sum of frequencies for the two variants is lower in the second line, they variants are more equally distributed. This makes the second line more interesting when looking at commonly used variants. On the other hand, the total frequency of all variants taken together is also important

The following formula serves the need for penalizing if two values are it is scalable, as multiplying the frequencies f_a and f_b with a constant factor n does not change the value for p.

$p = \frac{\sqrt{f_a f_b}}{(f_a + f_b)}$

For a macro comparison between two Term Spaces, three different approaches are conceivable. Firstly, the Dice-coefficient (or any other similarity measure such as Jacquard coefficient, Cosine similarity, Pointwise Mutual Information, Kullback-Leibler etc) calculated upon the number of times a term appears in both Term Spaces and the number of times it occurs in each Term Space individually tells something about the similarity of the distribution of one term. Through building the average for all term that appear either in one or in both Term Spaces, a value for the similarity/dissimilarity of these Term Spaces can be deducted.

A second approach lies in comparing the logarithmic frequency values for terms. One easy way to detect changes in the frequency distribution is the difference between logarithmic frequency values. Again, an average value can be build up by calculating the logarithmic differences for all terms.

Finally, looking for the most frequent terms of each Term Space that are not in the intersection is a qualitative way of inspecting the differences between two Term Spaces. By adding up the frequencies of these terms and comparing them to the total frequency of all terms (a Dice-coefficient, this time on a macro level), a value for the similarity/dissimilarity results.

6.6.2. Quality of Term Spaces

In many cases it is necessary to assess the quality of a Term Space, especially if the exact process with which it was originally accumulated is not known. For example, there are many different ways how a query log can be created. If some kind of preprocessing or filtering had taken place before a query occurrence was counted, artifacts may occur in the log and, even more grave, a general skew in the term distribution might follow. A large part of detecting *fishiness* can be done solely by looking at the frequency distribution. By comparing Term Spaces, advanced detection can be done that includes the terms themselves not only the pure frequencies.

The CoreMaker algorithm introduced above can re-engineer the filtering elements, if the accumulated terms underwent a sort of `grep`-filtering.

Apart from filtered term collections, there are also differences in how suitable raw term collections are for a specific purpose. In general, bigger is better for term acquisition and testing, but there also considerable differences in cleanness. For setting up a list of correctly written terms (see Part B, “Orthographic matches”), certainly a professionally edited resource is the first choice.

Typical indications of the quality of a term collection are a sharp decline in frequency distribution, many junk terms and the deviance from the usual list of high and middle frequency terms.

If the number of tokens decreases very quickly over the frequency ranks of the term collection, it is usually a sign of poor Term Space quality. Another indication is the amount of junk terms — i.e. terms that consist of non-alphabetic chars, have the same char three or more times in a row or exhibit other features discriminating them from standard terms. Finally, if a term collection does not have a substantial intersection in high and medium terms with other large term collections, a certain amount of suspicion is justified.

6.7. Structuring Term Spaces

The processes discussed above operate within one Term Space or between several Term Spaces, for example dividing a Term Space by looking for repetitive patterns or combining two Term Spaces. The following paragraphs go beyond a mere re-organization of a Term Space by adding a structure to it.

6.7.1. Complex vs. simple units and the issue of pre- and postcoordinations

One of the core properties of Term Spaces lies in repetitive patterns they generally exhibit. One kind of repetition is the combining of simple units to larger units. Depending on the distributional restrictions, few simple unites can combine to many complex ones. For example, consider a list of garments and a list of materials. Through combining these lists (”wool sweater”, *denim jacket*, *silk blouse*) many new units can be created, although not all combinations will make sense (”angora shoes”). Taken the other way round, if the combinatory principle is detected behind the specific garments listed above, describing them can be done in a much more compressed way[19].

Through discovering combinatorial principles, it is not only the reduction of descriptive length that follows as an advantage, but also discerning meaningful building blocks of complex units. Listing hundreds of combinations such as *Italian wool sweater* or *French silk blouse* may not only redundant compared to listing the three components

[19] See the principle of Minimum Description Length, introduced online at `www.mdl-research.org/`.

country, material and product separately, it also loses an important insight into the meaning of these complex units. If a new complex term of the same kind comes in, it can be analyzed following known combinatorial principles and be related to already known units, both simple and complex. For example, consider the new complex unit *Turkish silk blouse* that may now not only be related to the simple units *Turkish*, *silk* and *blouse* but also to complex units such as all the other silk garment terms already seen.

To what extent should complex terms then be stored anyway? This question is related to the opponent pair pre-coordination vs. post-coordination. In library science, these describes indexing systems that either explicitly contain complex index terms (pre-coordination) or let the user combine index terms in order to build up complex units (post-coordination). Obviously, pre-coordinated indexes are much larger and more redundant, while post-coordinated indexes rely on the user to make the right connections[20].

One advantage of pre-coordinated indexes lies in the control over complex units. Especially the most frequently used ones often exhibit subtle deviations from the principle of compositionality in meaning and usage.

Transferring the index principles to the world of Term Spaces could result in two findings. Firstly, discerning combinatorial principles is helpful in order to structure a Term Space. Secondly, keeping track of already seen items should include complex units, as even if they are fully reconstructable in meaning from their parts, their popularity is in general not predictable from the simple units it is built up. As was laid out above when treating Term Space cleansing procedures, the popularity of a term often gives a valuable indication of possible idiosyncractic behavior with regards to meaning and usage.

6.7.2. Classifying, Categorization, Clustering and Annotating Term Spaces

The semantic answer to the "divide and conquer" cry is manifold. One reason for the multitude of answers is that similar concepts are named and defined differently. In the following, it is tried to single out genuine divergent models of semantic-related subdivisions of Term Spaces, bearing in mind that the names given to these approaches can without doubt be discussed.

Nevertheless, it is hoped to develop a stringent framework of semantic dividing operations on Term Spaces, based upon the different logic of applying these operations to real Term spaces instead of formalisms. This is not to say that a formal view on these models and their application logic would not converge, yet this convergence is not laid out here.

While the terms categorization and classification are often used either interchangeable or without stringent discrimination, it is useful to distinguish between them. Adding to these basic principles of grouping and dividing are clustering and annotating.

[20] [Foskett 1982].

Categorization

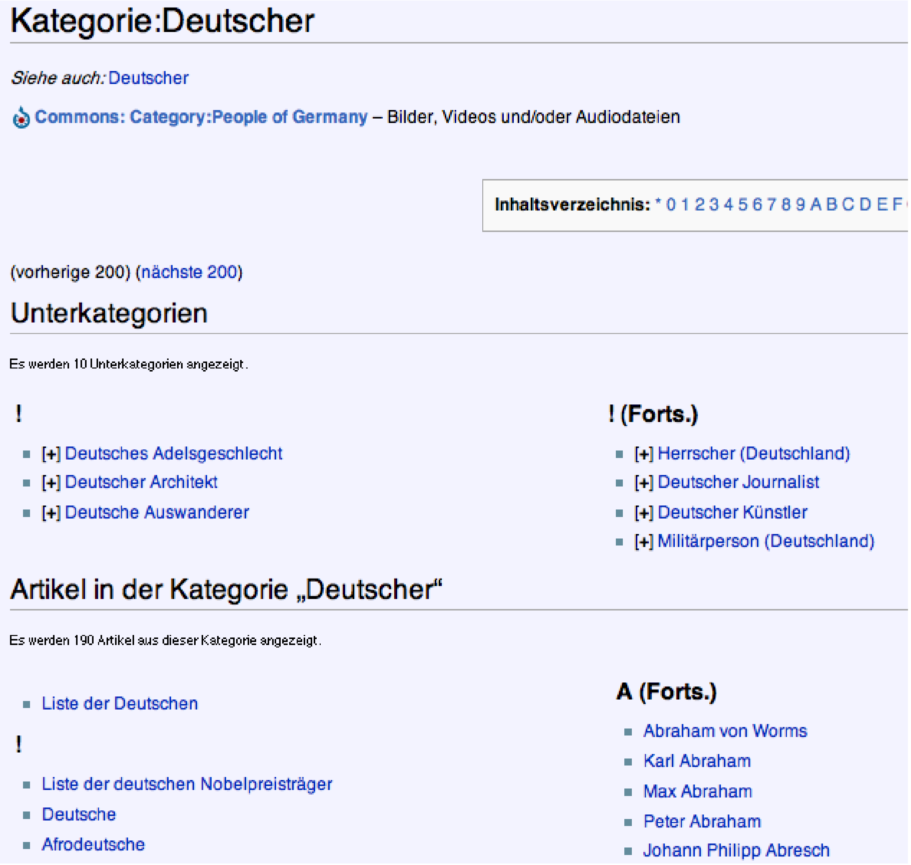
Categorizations is the process of grouping items based on similarities holding between them. It is a process deeply ingrained in human cognition and as such ubiquitous in cognitive activities. Indeed language in its function to give name to things is an act of categorization[21] As a information reduction process, categorizing facilitates communicating and handling observational data.

This does not only hold true on the level of general cognitive processes, but also applies to the handling of terms. Categorization is one way to reduce the complexity of Term spaces. While categorization has the benefit of operating on fewer units, it also bears the danger of disregarding differences under the label of a category. For example, if a Yellow Pages site allows to search for pet shops (i.e. a category of businesses), it is not possible to differentiate between different subtypes of pet shops (exotic animals, pet supplies, reptiles etc).

Categorization is usually non-exclusive, meaning that one unit may belong to more than one category. Unless classes, categories are potentially overlapping groupings of items. It is not necessarily possible to enumerate the instances of a category. This does not exclude a hierarchical organization of categories. Subcategories have higher restrictions on the similarity between the items they hold — for example, Pekinese dogs are more similar to each other than dogs in general are to each other. Supercategories have accordingly lower restrictions on the similarity between their instances.

A typical example of categorization are the Wikipedia categories. Typically, an article on a person has at least categories for gender, birth year, ethnicity and profession. Obviously, such a categorization cannot be easily integrated into a hierarchy, as the categories reflect independent facets of the person. Although Wikipedia also allows to put subcategories under categories, this feature is neither used extensively nor consistently. The following screenshot shows

[21] Aristoteles' Categoriae, the seminal text for categorizations, links the ten categories closely to the grammatical structures of Ancient Greek.



Kategorie:Deutscher

Siehe auch: Deutscher

Commons: Category:People of Germany – Bilder, Videos und/oder Audiodateien

Inhaltsverzeichnis: * 0 1 2 3 4 5 6 7 8 9 A B C D E F

(vorherige 200) (nächste 200)

Unterkategorien

Es werden 10 Unterkategorien angezeigt.

!

- [+] Deutsches Adelsgeschlecht
- [+] Deutscher Architekt
- [+] Deutsche Auswanderer

! (Forts.)

- [+] Herrscher (Deutschland)
- [+] Deutscher Journalist
- [+] Deutscher Künstler
- [+] Militärperson (Deutschland)

Artikel in der Kategorie „Deutscher“

Es werden 190 Artikel aus dieser Kategorie angezeigt.

- Liste der Deutschen

!

- Liste der deutschen Nobelpreisträger
- Deutsche
- Afrodeutsche

A (Forts.)

- Abraham von Worms
- Karl Abraham
- Max Abraham
- Peter Abraham
- Johann Philipp Abresch

A screenshot of `wikipedia.de`: Categories for *Deutsche*.

For example, the category *Deutsche* lists hundreds of persons as immediate members of the category, but also articles on groups of Germans (for example *Auslandsdeutsche, Aussiedler, 100 Köpfe von morgen*) and in addition subcategories (again groups of Germans, *Deutsches Adelsgeschlecht, Deutscher Architekt, Deutsche Auswanderer*). None of these associations are per se inaccurate, yet they do not build up a consistent hierarchy. If the membership of a unit in a category is not always clearly assignable, the building of hierarchies cannot rely solely on the principle of inclusion, i.e. supercategories include all instances of their subcategories.

Classifying

In TE-Commerce, a classification is a non-overlapping partitioning of a domain. It is defined by the semantic of the classes (class intension) and members (class extension). Classes are build a priori, they do not follow from the observed data but rather from general principles that are then applied on the data.

This makes classifying a viable way of grouping items if it is already known before hand what different features and values these items exhibit. An example of classification is the drill-down based on document types (html, pdf, doc, Excel etc) that is provided by Search Engines such as `clusty.com` or `exalead.com`. Here, the membership relation is known a priori and can be applied without creating uncertainty.

The extensions of a class can in general be enumerated and their is in general no uncertainty whether a unit belongs to a class or not. In detecting classes as they appear within language, some amount of fuzziness cannot be avoided. Contrasted with taxonomy classes, for example in biology, classed grounded in language can only be observed in a much more uncertain way, through testing the acceptance of a context in which only instances of the class can be inserted. For example, testing the class *clothing* could make use of a context such as "today, he is wearing X" where all possible insertions for X should ideally make up a class of similar objects. In fact, this context also allows insertions that are clearly no clothing articles, for example *a smile* or *a monocle*. By posing more contexts, however, the set of instances that can appear in all of them thins out. This requires to introduce a sort of weighting of how many and which test contexts have to be passed before a class is assigned. Moreover, it is not always clear what a requires that a term instantiation in a context is accepted — can this decided by introspection, by asking speakers or by observing corpora data? Through these two mechanisms — uncertainty what it means to fulfill a context requirement and uncertainty how many and which context requirement need to be met, fuzziness sneaks in testing a linguistically motivated act of classification.

One way to gain safe ground in classifying terms is to start top-down. Following this lead, applying the concept of classification to TE-Commerce could start by dividing the top semantic types appearing in the textual representations of commercial activities, such as human descriptors (for example professions or job titles), institutions, natural concreta and artifacts (for example product types), places (for example vendor types), activities (producing, transacting, communicating etc.), forms or states[22]. On a more granular level, the building of semantic classes resemble the content of a product and service taxonomy.

The upper bound on granularity that can be achieved through the method described above is reached when underlying concepts cannot be discriminated through contexts. Different taxons might be indifferent to contexts — consider for example, all species of a sparrow. This subdivision is not modeled within the lexicon, at least not those of the standard language.

Classifying terms requires a considerable effort, given that the classes not only have to be disjunct and covering, but also that finding contexts that validate the assigning of a term into a class can be demanding as well. In the further experiments, the use of classes is restricted to the high level classes denoting semantic types. Among these types, it is not only natural concreta and artifacts (products) and activities (services) that are of main importance, but also the human descriptors.

[22] A classification of German simple nouns into more than 400 different semantic classes is described in [Langer 1996].

Clustering

Clustering is an operation that tries to discern similar items in a space. It can be made operational through a twofold maximization process, maximizing the similarity between items within in the clusters and maximizing the distance between individual clusters.

The notion of clustering is tightly connected to a spatial view and thereby to geometric models of terms and terms in relation to each other (see above). In accordance to the two maximizing processes that may take place in clustering, one strategy is to find a cluster centroid and determine a radius around it that contains the elements of the cluster while a second strategy cuts through potential clusters. In both cases, the process of finding clusters is either terminated through a predefined number of clusters that have to be created or through a threshold that defines when no-more items can be added or removed to a cluster[23].

Without a recognition of term variants, it is not guaranteed, however, that equivalent term variants are put into one cluster[24]. It might happen that they create two different clusters[25]. In this case, the values of clustering for TE-Commerce purposes is lowered.

The main application of clusters in TE-Commerce lies in the refinement of searches through a restriction of the search. As usually only one cluster can be selected for enhancing the search, currently deployed clustering systems do only provide a narrowing down of searches. This is not always satisfying the information needs of the users and it might even have a detrimental effect if the clusters are not disjunct[26].

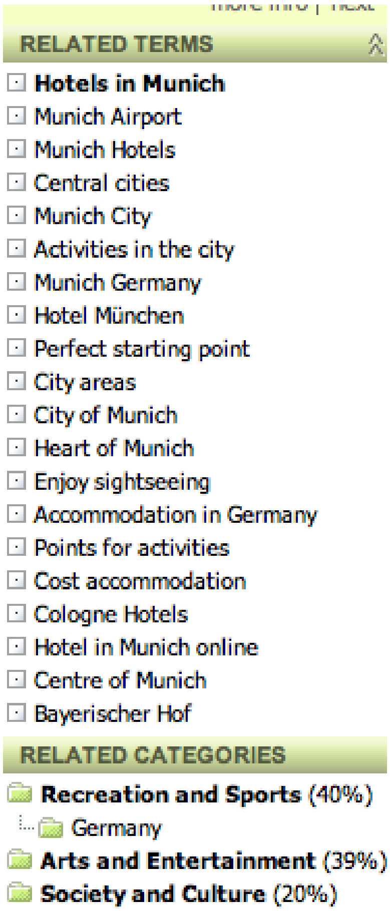
To illustrate these shortcomings, here is an example of clustering in a search applications (Vivisimo/clusty.com):

[23] Cluto, a freely available software for clustering, for example requires to set the number of results beforehand. See `glaros.dtc.umn.edu/gkhome/views/cluto [Nov. 1, 2006]`.

[24] Using WordNet to overcome the limitations of the bag-of-word approach is reported to have positive effects in document clustering. See [Hotho/Staab/Stumme 2003].

[25] See [Chakrabarti 2003].

[26] See [Lewandowski 2005], Ch. 10.4.



more info | next

RELATED TERMS

□ **Hotels in Munich**
□ Munich Airport
□ Munich Hotels
□ Central cities
□ Munich City
□ Activities in the city
□ Munich Germany
□ Hotel München
□ Perfect starting point
□ City areas
□ City of Munich
□ Heart of Munich
□ Enjoy sightseeing
□ Accommodation in Germany
□ Points for activities
□ Cost accommodation
□ Cologne Hotels
□ Hotel in Munich online
□ Centre of Munich
□ Bayerischer Hof

RELATED CATEGORIES

Recreation and Sports (40%)
Germany
Arts and Entertainment (39%)
Society and Culture (20%)

Clusty.com results for *hotel in munich*.

These results are at best a mild inspirational source. There is no consistent structuring provided by these clusters, neither with regards to coherence within clusters nor with regards to clusters being distinct. A third issue is that the naming of clusters does not always reflect a super-ordinate term for the members of the cluster.

Annotating

A light-weight approach to relate terms with meta information is annotating (sometimes also called tagging, though this should not be confused with user-based tagging the context of Web 2.0, see above). Annotations insert the meta information within the textual string of the annotated data. For examples, annotating named entities in a queries would transform the query string *bill clinton biography* into *[bill clinton = _person] biography*.

When examining human language, the meta language has to take elements from the examined language. In annotating, the correspondence between meta and object language is acknowledged through the flattening of structural discrimination between the both.

Putting all information into the textual string allows to set up efficient parsers using finite state methods. Instead of just applying a lexicon on a string input, a transducer can operate on already recognized units and produce higher level matches. This cascaded method of applying transducers makes use of fail-proof base-level matches before moving on to uncertain rules[27]. The MWU procedures describe above work similar, given that there the already known MWUs are shielded from higher-level matches.

Annotating has its drawback if larger structures are taken into account. While it is suitable for application on a local level (such as n-grams), the overhead needed for annotating structures spanning over long distances is considerable. Any grouping of non-adjacent units needs some kind of additional index counter: In the example above, the alphabetic index A indicates the grouping of unit *Bill Clinton* and *president*:

> In 1992, [Bill Clinton=person A] was elected as [US president=title A]

However, long ranging dependencies are by no means central to the approach laid out here. TE-Commerce is based on terms and the syntagmatical and paradigmatical dependencies between them, making annotations a viable methods for enriching Term spaces with information.

[27] See [Roche/Schabes 1997].

Part B.

The lexicon and grammar of E-Commerce

7. E-Commerce Term Management

A viable term management system for E-Commerce terms has to lay out at first suitable classes to which E-Commerce relevant terminology can be assigned. In a second step, database design and In the following sections, different classes, specific to E-Commerce terminology, are introduced and examined both from the perspective of traditional electronic lexicography and from the perspective of their usage and occurrence in E-Commerce data. Each class is introduced by explanatory notes of its intension and examples — both from German and English — demonstrating its extension in the world of E-Commerce. Special focus is given to separation problems between these different classes. A related issue is the internal distribution of these classes, a question which connects this chapter with the notion of Term Spaces.

Obviously, each class has some prototypic representants and also less clearly demarcated cases. The prototype representants provide explanatory material allowing to convey what a given class is supposed to entail — in fact, introducing the E-Commerce classes will in general resort to state examples, instead of laying out definitions. This topic had already been touched upon and will be continued in the concluding chapters of this book. While finding analogies is considered here as superior to applying in a non-operative way definitory clauses, there are, however, cases when the analogies need to be stretched in order to come to a classificatory conclusion. This does not refer to lexical ambiguous cases that feature one literal expression referring to more than one concept — such as *nursery* for children and for garden nursing —, but to cases which are intrinsically hard to decide. These do not entail cases that are hard to resolve because they require specialist knowledge (but once this knowledge is acquired, become easy to decide). A genuine example, however, is given by the lexicalization process proprietary product names undergo that become the common name for the kind of product, even if sold by other companies. These Tivo-cases reveal that the actual usage of lexical units is to be taken most seriously. There is no short-cut to bypass meticiously observing it: Even if *Tivo* appears in trademark lists, one cannot inference it is solely perceived and used as a proprietary term. Apparently, these issues border to the legal domain in which the status, protection or removal of proprietary terms is decided.

The frequency distribution of instances of a class adds another structure to the class, not necessarily concomitant with the prototype vs. peripheral division. The large number of elements that occur very rarely may hold convincing prototypical examples of the class as well as the much smaller number of high frequent instances. While frequency is one of the most important heuristic measures to determine the quality and usefulness of a term, it is not the sole answer to the question of what makes up a good term.

In this chapter, this question is going to be addressed via laying out the possible atoms of TE-Commerce and how they relate with each other. Starting with the introducing the different classes, the combinations of their instances in real-life terms, namely YP inventories and product catalogs, are examined. This chapter concludes with a sketch of a TE-Commerce database system that helps to keep track on the seen items.

7.1. The classes of E-Commerce term units

In this section, the different classes of E-Commerce terms are introduced, starting with the most predominant (albeit not most numerously instantiated) class of generic product and service types. Whereas these terms belong largely to the realm of the standard lexicon of a language, the subsequently presented classes appear almost exclusively within E-Commerce-related texts. From this it follows that standard lexical resources are in general not covering for E-Commerce.

For the purpose of gathering and lexicalizing E-Commerce terms, the top 63.630 frequently asked queries in SQ have been manually inspected and classified as described below[1]. These searches were taken from the intersection of four months of queries, i.e. they recurred in each month of the four month period. The results of this analysis will also be presented with regards to query log anaylsis (see Part C, Query System).

7.1.1. Product and Service types

Naturally, the first class that springs to mind when one is taking a look at the textual dimension of E-Commerce is the class of product and service types. In general, instances of this class are lexicalized and non-proprietary terms. They are integrated into the lexical system of a language, including various syntagmatic and paradigmatic relations and morphologically conditioned realizations.

As prototypic examples of product types could rank terms such as *computers*, *lily bulbs* or *office chairs* (just to demonstrate the diversity of product types). Their predominant characteristics is a relatively clearly defined concept that attaches with them and a firm integration into standard lexicon. They usually populate a middle level of granularity [2]. For example, *electronic devices* represent a very general concept, while *dvd recorder 160mb harddrive* represents through its syntagmatic modifiers a very specific concept. Both occur much less frequent than *dvd recorder* of which the former term is a hyperonym and the latter a hyponym.

Very general concepts have, however, some importance as upper level terms in E-Commerce hierarchical categorizations. Thus they will be dealt again later in this chapter, in the context of E-Commerce ontological resources (see below, Part D) . On the other end of the continuum, very specific concepts are not only expressed by adding modifiers to a generic product type (*dvd recorder with hard-drive*), but also often by

[1] This work was performed by student assistants under the supervision of the author.

[2] See [Cruse 1986].

referring to a concrete product, such as *Sony RDR-HX 725* (see below). Naturally, these proper names exhibit different distribution and variance patterns than generic terms.

The distinction between product and service types can be explained on several levels. The underlying model for a product type P is a process in which A sells P to B with the result that A no longer possesses P, but B does. In contrast to this, service types are characterized by an activity taking place between A and B. On a linguistic level, product types are nominal types (including proper nouns), services span verbal and nominal types. An example for a nominal service type is *download*, as a download does not imply a change of possession took place.

Product and service types behave similar in that they can both be related to generic E-Commerce features, especially price, quality and availability (see also below, Product features).

From a word formation point of view, the occurrences of compounds in product types (e.g. *digitalkamera, mp3 player*) and acronyms (e.g. *pda, dvd*) is striking. For TE-Commerce purposes, it has to be made sure that the electronic lexicon contain both relevant morphologically complex entries and short forms. Even from the small list presented above it becomes clear that many E-Commerce-related terms used in German are foreign words (such as *notebook, mp3 player*) and/or do not exhibit standard German morphology. Generic lexical resources will not in general be sufficient to deal with these phenomena.

A common linguistic phenomenon is forcing a reading by a context. The same occurs in TE-Commerce contexts with terms that are forced into a E-Commerce relevant reading through the context they appear in. Often reduced forms are triggered to a E-Commerce reading by their context, for example *mp3* often refers to *mp3 download* or *mp3 players*. Detailed examples will be discussed in the chapter on the paid keywords space (see below, Part D).

Product and service terms are dominant in shopping log queries. If taking a look at the top 63.630 frequently asked queries in SQ, a manual inspection yielded that 31.680 contain product types. Out of this number, contained a product type together with a brand or a feature while the rest (21.735) reflects unique product types that made up full queries.

Three samples from the top, medium and low frequency spectrum of this list give an impression of what product types can be found in

HIGH FREQUENCY

handycover
auspuff
antibabypille
herrenschuhe
plasmafernseher
stempel
druckerpatrone

lcd-fernseher
holzspalter
gehrock
kran
fussballschuhe
wandhalter
tischdekoration
milch
verkleidung
tischwäsche
brotbackmaschine
autoreifen
allesschneider
tischleuchten
notebook zubeör
postauto
tiere
windlichter
wimperntusche
cb-funk
faschingskostüme
sat schüssel
high-heels
container
kinderuhren
senf
mokassin
gartenliege
überwürfe
vitrinen
malbuch

MIDDLE FREQUENCY

teppichböden
laufbänder
wachs
spanplatten
auflage
kantholz
tauchsport
tft lcd
kontrabass

stehpult
art schuhe
fluege
freizeithose
motorradjacken
dsl kabel
füllhalter
damenröcke
bastelartikel
elektrische zahnbürsten
feuchtigkeitskiller
kapuzenshirt
drucker patrone
kissenbezüge
hühner
schiebegardinen
herrenschuh
lederjeans
schuhe pumps
muskelstimulator
minikühlschrank
federkernmatratzen
angora
damen uhren
medikamenten
kompletträder
armani uhr
tourenski
duftkerzen
espressotasse
hängevitrine
gesichtsbräuner
gästebett
balkongelander
salbei
dämmstoffe
balkontüren

LOW FREQUENCY

raeucherofen
sektionaltore
yogi
recorder

schwenkarm
gurte
rollläden
haar entferner
raffgardinen
abtropfgitter
jagdzubehör
feuerwehrbekleidung
klavierbank
damenanzug
trekkingstiefel
gurtband
truhencouchtisch
spritzpistolen
kinderanzug
pu erh tee
baskenmütze
federmäppchen
kreiselpumpe
freischwinger
entspannungsmusik
papiertüten
mobilheime
tv-wagen
treppenlift
schlagzeuge
jahreslöffel
mini skateboard
küchenrollenhalter

It is not not always clear whether an item listed here is indeed a product type or a product type combined with a feature. One indication that it is indeed a product type is the frequency of usage. Another indication is given by a non-compositional semantic. Considering *Baskenmütze*, it is not possible to fully deduct its meaning from the meaning of its components. In the case of *Plasmafernseher*, it is the high frequency that suggests to treat it as separate product type, although its semantic follows a very productive pattern (consider *Elektroauto* and *Halogenlampe*). Test patterns such as *no conventional X (e.g. tv), but a YX (e.g. plasma tv)* — compared with ** no conventional sweater, but a red sweater* — are a method to discern constitutive features.

7.1.2. Brands and Makers

It is often not cleared where to draw the line between brands and company names, especially as company names often appear in a reduced form with indicating the legal status (*Siemens AG – Siemens*).From the perspective of users, the difference should not in general not play a decisive role. The legal relations between brands and companies will be in many cases not of interest to users — moreover, these relations might change over time. The TE-Commerce lexica should not include facts such as that *Mediamarkt* and *Saturn* are two chains that belong to the same holding, just as it does not contain information on how many shares in *Volkswagen* are owned by *Porsche*.

Besides listing all these brands and makers, an additional valuable information is the domain for each brand and maker. For example, *Porsche* is connected to the automotive domain, while *Microsoft* is connected to the IT-sphere. These associations can be stored in the lexicon in oder to provide a default entry for the domain. If coming across a parse result such as *software made by volkswagen*, it can be marked as suspicious, to allow a human to check it again. For the purpose of clustering Web documents or meta keyword lines, the appearance of a brand in a domain and other products of the domain can fortify the association to this particular domain.

Brand names can be obtained either by harvesting existing resources such as product catalogs for terms that appear with specific classes of product types, for example only home electronic product types. Legal repositories of brands and trademarks are available[3].

A list of 20.000 brand names sorted by frequency:

- vw
- bmw
- audi
- honda
- ford
- nokia
- renault
- peugeot
- mazda
- toyota
- sony
- nissan
- chrysler

[3] For example, the Romarin CD that contains WIPO's database of international trademarks. See `www.wipo.org/madrid/en/romarin/ [Nov. 1, 2006]`.

- mitsubishi
- samsung
- yamaha
- canon
- ikea
- alfa romeo
- panasonic
- medion
- esprit
- toshiba
- motorola
- hyundai

The names of companies, especially smaller companies, deserve a special treatment, given that often the name indicates the business type. That allows firstly to cluster company names in order to obtain more vocabulary for a given domain. For example, here is the word frequency list (insignificant words such as legal titles have been removed) for the category *French restaurants*:

A similar list can be obtained by clustering name parts of a large database of companies to their categories, in this case four-digit standard SIC codes[4]. This list shows company name parts (single words) that occurred more than four times more frequently for one SIC code than for any other SIC code. It is listed by total frequency of the company name parts:

CONSTRUCTION – 1521 General Contractors-Single-Family Houses
HAIR – 7231 Beauty Shops
INSURANCE – 6411 Insurance Agents, Brokers, and Service
RESTAURANT – 5812 Eating Places
REALTY – 6531 Real Estate Agents and Managers
SALON – 7231 Beauty Shops
CHURCH – 8661 Religious Organizations
PLUMBING – 1711 Plumbing, Heating and Air-Conditioning
BEAUTY – 7231 Beauty Shops
ELECTRIC – 1731 Electrical Work
AGENCY – 6411 Insurance Agents, Brokers, and Service

[4] This experiment was conducted using the database of 10 million US records from 1997. See above, Part A, Term Spaces.

SCHOOL – 8211 Elementary and Secondary Schools
APARTMENTS – 6513 Operators or Apartment Buildings
HEATING – 1711 Plumbing, Heating and Air-Conditioning
PAINTING – 1721 Painting and Paper Hanging
CLEANERS – 7212 Garment Pressing, and Agents for Laundries and Drycleaners

This procedures does not only show what company names are most prominent for a SIC category, but provides at the same time very good representatives of these category.

Experiments with a set of over 10 million US businesses from 1997 showed that at least 25-30% of company names contain a part of that indicate the line of business, at least finely granulated enough for SIC codes.

Another issue in connection with business names are the variant forms in which they appear. One typical scenario is that the legal name and the name by which a business is most commonly known differ. A test based on business name variants from two directories (one being dmoz, the other one a proprietary US business index) that could be linked through their common website yielded that only 510 appeared in the same form in both directories, while 2.626 appeared in divergent forms. With a simple local grammar it was possible to cut down this number to 1.150 non-matches. The outlines of this grammar which is based on a segmentation of business names into a body and a suffix (such as legal form) is presented below. The square brackets indicate what process was necessary to transform the variant on the left hand side to the variant on the right hand side:

Plascore Inc. → Plascore, Inc. [additional separator]

Kemlite Company → Kemlite [suffix reduction]

ACE - Applied Composites Engineering → Applied Composites Engineering, Inc. [acronym introduction; suffix expansion]

Ceradyne Inc. → Ceradyne Thermo Materials Corp. [body expansion; suffix substitution]

Uneeda Enterprises → Uneeda Enterprizes, Inc. [body orthographic variant, suffix expansion]

Radiac Abrasives → Radiac Abrasives, Inc. [additional separator; suffix expansion]

Crystal Associates, Inc. → Crystal Assocs., Inc. [body abbreviation]

AFG Industries, Inc. → A F G Industries, Inc. [body acronym spacing variation]

Weyerhaeuser Corporation → Weyerhaeuser, Containerboard Packaging & Recycling [body expansion; suffix reduction]

Metso Paper, Inc., → Metso Paper Service Center [body expansion; suffix reduction]

Nanophase → Nanophase Technologies Corp. [body expansion; suffix expansion]

Crain's Cleveland Business → Crain Communications, Inc. [body substitution; suffix substitution; additional separator]

ACT → Advertising-Communications Times [body expansion of abbreviation]

Midlands Business Journal → M B J Corp. [body abbreviation; suffix expansion]

7.1.3. Proprietary Product Designators

Proprietary product designators are the names that companies give to their products. These are in general made-up words, sometimes using letter-digit combinations. In the latter case, the numerals often represent subtypes of a product line, for example Samsung X-05 vs. Samsung X-15. In this case, the proprietary product designators has a decomposable structure — in the example above, it makes sense to group all Samsung X models because they belong to one product line. In many other cases, the proprietary product designators are just opaque strings that have to be treated as a unit.

Proprietary product designators are strongly correlated to brands. This allows to use association metrics to them. They can then be filtered by a regular expression that allows non-lexical words and digits to be part of the proprietary product designators.

- toyota supra
- microsoft windows
- honda civic
- adobe acrobat reader
- microsoft word
- microsoft internet explorer
- ford mustang
- new balance 991
- dodge viper

7.1.4. Vendors

Vendor types denote types of commercial organizations that provide products or services. In general, these organizational types occur in the same sentential and phrasal frames as physical objects, for example in the following sentence frames:

X drives to the
In the
At the site of the

The prototypic vendor type is a place where a customer can purchase goods or ask for services. However, the understanding of vendor type used here does not necessarily involve a physical meeting of customer and seller, but might be for example make use of online transactions instead.

Vendor types have to be discriminated from proprietary vendor names (for example *Mediamarkt, Saturn, Kaufland, Tesco*) and metaphoric vendor names. The latter group refer to those parts in company names that indicate the line of business but do so in using a metaphoric term, such as *Blumeneck* for a flower shop or *Getränkequelle.*

A list of frequent German vendors:

- Metzgerei
- KFZ-Meisterwerkstatt
- Malerbetrieb
- Bauunternehmen
- Restaurant
- Gaststätte
- Fleischerei
- Unternehmensberatung
- Bäckerei
- Schreinerei
- Bauunternehmung
- Werbeagentur
- Massagepraxis
- Tankstelle
- Pension

- Fahrschule
- Architekturbüro
- Naturheilpraxis
- Tischlerei
- Ingenieurbüro
- Fuhrunternehmen
- Rechtsanwaltskanzlei
- Autohaus
- Haarstudio
- Pizzeria

Typically, vendor names are orthogonal, meaning that they can combine with many different branch heads. This list of vendor types is sorted by their specificity for a domain, starting from very general and moving to more specific vendor types:

- Filiale
- Geschäftsstelle
- Büro
- Agentur
- Studio
- Kanzlei
- Praxis

Heuristics to find a vendor type include morphological clues. Examples are the suffix *-erei* (noun agents ending on *-er*) or right parts of compounds that indicate a rather metaphorical vendor type such as *-haus, -stelle.*

Occupational titles often serve the same function than vendor types. These appear generally in the same contexts. Discriminative contexts for the two classes can be based on the different semantic classes to which its instances belong (humans vs. places). For example *X wanted to become a* is a discriminative context for an occupational titles. No vendor term can enter this context.

7.1.5. Additional terms (problems, challenges, shortcomings / need for action)

Many keywords are pointing to YP categories but do not fall under any of the classes presented above. For example, consider the term *acne*. It can be associated to several YP categories (such as dermatologist or cosmetics), but it is not a product or a service.

Other examples are terms such as *Heizoelpreise* or *computerabsturz*. These do not represent services or products per se, but serve as a trigger for these. Answering them is often underdeveloped in today's search agents. A special case are terms such as *handynummer* or *hausnummer*, as a query containing these terms is a strict White Pages-query. It requires an exact answer and cannot be answered for example through a listing of businesses in a category.

7.2. Product features

Product features specify different manifestations of a product type. For example, a television set can be specified by the features *NTSC/PAL capable* or *30 inch flat panel*. Product features in TE-Commerce usually refer to broader groups than individual product instances. If one individual television set has a scratch on its case, this property does not rank as a product feature. However, a property *used condition* would rank as a product feature.

General product features apply to a broad variety of product types. Among those general product features are price, size, color, weight and condition.

Other features and feature combinations are distinctive for a product type or a group of product types. For example, the feature combination material-of and size attributes such as S, M, L or XL are characteristic of garments. If creating a taxonomy of product types, it is good advice to take heed of features. Product types that are put in the same branch of the taxonomy should share the same set of features. To facilitate the labour of adding feature options, inheritance mechanisms can be used that specify the feature options for a subtype through its supertype.

Product features are combinations of a feature name and a feature value. Feature values might be of a binary type for example the feature *teletext* of a TV set), be chosen from a closed set of feature values (for example the format of dvd recorders such as +R, -R, RW etc) or be expressed by numerals, optionally with units. The latter case is exemplified by general properties of products such as price or size.

Both feature names and feature values exhibit term variance. For example, the feature *56 inch screen size* may also appear as *56-inch diagonal screen size*, *Visible Diagonal Screen Size: 56in*, *56-inch (diagonal)* and in many more variants.

7.3. Branches and Yellow Page headings

Through examining about 30 German Yellow Pages, business directories and product search sites and extracting their category systems, it was possible to create an inventory of more than 200.000 branches and product categories.

This list was assembled solely from sites that present their categories in a flat list. This was done because the lines were taken out of their context. In hierarchically organized category systems, it happens quite regularly that a category descriptor is just the discriminating feature of its level and cannot stand alone.

For example, a hierarchically organized branch list might contain lines such as

Ärzte
Nieren
Herz
Magen

Obviously, it is not possible to take these lines out of their hierarchical context if they still should reflect what they meant in it.

7.3.1. Atomic Branch Descriptors

Extracting the atomic branch descriptors from this list of branch descriptors can be done by looking for typical right segments in compounds with the branch head as the left segment. These segments comprise *-branche, -gewerbe ,-sektor, -firma* etc. In a second step, second level branch heads, such as *Zukunft* or *Haupt* have to be pruned.

The following list shows the top resulting branch heads, sorted by the number of right segments with which they appeared:

268	computer
259	medien
259	bau
241	film
240	software
240	musik
233	kunst
233	kultur
233	auto
231	sport
225	internet
223	projekt
222	netz
218	umwelt
218	foto
217	video
217	system
215	kommunikation
214	information
214	haus

7.3.2. Complex Branch Descriptors

Complex branch descriptors come in two different fashions. One is complex branch descriptors in hierarchical lists, the other one in flat lists. Complex branch entries from flat lists will be treated first.

A typical form of complex branch descriptors is a coordinate structure. The coordination can be spelled out differently, for example using *und*, *u.*, *&* or a slash. A morphological phenomenon in this context is the appearance of elliptic coordinations in which one repeating segment is left out.

Different types of elliptic coordinations have to be treated separately. One type is leaving out the first segment of the second element, for example *Autoreifen und -felgen*. Another type is to leave out the second segment of the first element, for example *Haus- und Gartenarbeit*. In few cases, both types are mixed (*Holzbe- und -verarbeitung*). Even if no reliable compound analysis is available, it is feasible to detect the proper reconstruction of these elliptic coordinations. This can be done through using a large frequency list with a fast look-up methods — a hash table or a binary tree —, shifting the segments position by position and monitoring the maximum frequency count.

In the most basic case, the complex branch descriptor is just the sum of the atomic branch descriptors. Through lexicalization, though, the meaning of the coordinated from might differ from an aggregate of its elements.

There are different methods to build up a complex term from atomic branch descriptors. One way is to append a business facet, for example accessories, raw materials, products etc., to the atomic branch:

keramische rohstoffe
keramische erzeugnisse
zoologische bedarfsartikel

A detailed list of such business facets is provided under the title "orthogonal terms" below (see "Semantic matches").

7.4. Vocabulary maintenance and control

The different kind of terms and parts of terms introduced above have to be stored, inspected and optimized if they are going to be used or re-used in applications. For this purpose, a database (DB) setup — including interface and querying mechanisms — has to be designed and implemented. With reusability as one of the core assets in the TE-handling proposed in this paper, this setup has to allow efficient storage and inspection of terms, yet also needs to be flexible enough to incorporate newly detected or emerged terms.

While the units stored in this database appear and are accessible in the form of terms, they may refer both to language-external entities such as products, services, brands, vendors etc, and to language-internal entities such as morphemes[5]. Units described in the term database are not solely stored as isolated elements, but enter relations which also have to be encoded. Unary relations assign properties to terms, for example their average worth as paid keywords. Binary relations hold between two units, for example between synonymous morphemes or cross-selling products.

In addition, a control process has to established that ensures a high quality of the stored vocabulary. While it cannot replace human editing, it may aid the editor as it points out suspicious looking units and statistical outliers. Several control processes, covering both unary and binary relationships will be introduced.

7.4.1. Term Maintenance DB structure

The core term repository in the TE-Commerce framework is the Term maintenance DB. It needs to fulfill the following requirements:

- Integrity of the data structure and content
- Deployment in multi-user environments
- Providing simple and non-intrusive lookup-procedures

Some explanatory notes on each of these requirements:

Integrity of data structure and content has several aspects. Firstly, all entries have to obey syntactic requirements. For example, a field that requires a value (such as the normalized term form, see below) cannot be left empty. Fields that require numerical entries may not store alphabetical chars. Secondly, all changes to the data content need to be protocoled and need to be reversible (undo functionality). Thirdly, the data needs to be stored safely, both in terms of backup strategies (such as RAID hard drive clusters) and intellectual property theft. The latter can be achieved through salted entries, i.e. entries that are altered or made-up in a way that any re-occurrence of it outside the DB is a clear indication it was taken from the DB, encryption, access restrictions and access protocols.

Deployment in multi-user environments is required as the amount of work needs to be distributed both in a horizontal way among collaborators and in a vertical way to different stages of QA and auditing.

This includes a stack of “to-do terms” that allows different users to pick those terms for which they feel fit, based on their domain knowledge. Monitoring and reporting components are a valuable addition. Standard Wiki-software[6] usually ship with the necessary prerequisites for these components. By looking at the version history, it is possible to extract the amount of work that was done individual users. It is also

[5] See Entity-relation model in Database setup, [Chen 1976].

[6] Such as Media Wiki, the software behind wikipedia.

possible to create Wiki pages through a CGI script, allowing complex interactions between the different components starting from one workbench.

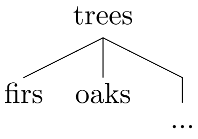
One component of a simple and non-intrusive lookup-procedures should be a fault-tolerant dynamic search interface, ideally with a Suggest functionality (as presented above in Part A, Interactions). Inspection routines should contain checks for frequency, concordances and intersecting the terms with other lexica.

The format of the lexicon lists can follow the DELA format which lists base form, inflected form and inflectional attributes for words. This has to be enriched with database maintaining columns,

7.4.2. Representing and storing term hierarchies

Several commonly used methods for representing and storing hierarchical data are presented. Following what was said above in Part A about the two-level hierarchical structures of term collections, it is sought here to achieve a representation of hierarchies that fits into individual lines and allows to take the lines out of context (for example by sorting them).

In the following section, the following mini sample taxonomy is represented in different ways:



It is therefore necessary to transform representations such as the widely used XML format. This should not deny the worth of structured data formats. However, as was laid out above, it is especially the content that matters in TE-Commerce. In XML, one way to represent hierarchical structures is to use nested tags:

```
<type ID=”1” name=”trees”>
<type child-of=”1” ID=”2” name=”firs”>
...
</type>
<type child-of=”1” ID=”3” name=”oaks”>
...
</type>
...
</type>
```

Here, the <type>-elements are nested, allowing to represent a hierarchy[7]. An attribute *child-of* links the children type to their parent. Polyhierarchies can be represented if the attribute *child-of* accepts lists of IDs.

A simplified representation of nested tags that works only for monohierarchical relations are indented lists. These represent the level of hierarchy by indentation of node labels. The lines may no be taken out of context if the hierarchy is to be preserved.

Representing hierarchies in a way so that individual lines might be taken out of context can be achieved by using relational tables[8]. Usually, the term content of the hierarchy is separated from the structural content (see also Part D, Applying taxonomies in TE-Commerce). The term content is a list of relations between IDs and descriptors. Structural relations are presented in tuples of IDs. The tables for the different relations (such as broader term, narrower term, used-for etc) can either be split up in several files or put in one file that lists the type of relation as well as the IDs of the relata.

A more verbal way of hierarchy representation is a hierarchical numbering scheme. This scheme is mostly known through table of contents. It divides the layers of hierarchy through dots and numbers the nodes at each level:

```
1.
1.1.
1.1.1.
1.1.2.
1.1.3.
1.2.
2.
```

The lines here can be taken out of context. Another advantage is that broader and narrower terms can be extracted very easily. All terms with a number that is a true prefix of the number of a given term are broader terms to it. All terms with a number of which the given term's number is a true prefix are its narrower terms.

Finally, storing all hierarchical information of a term within one line allows to integrate it into other repositories without keeping the other lines or losing any information about the taxonomical context. Such a representation takes of course more space than the methods described above. However, it allows to accumulate repositories with hierarchical information from different sources into one large list. It can be achieved by adding the navigational path to a term, i.e. the concatenation of the higher nodes in the hierarchy. A line could thus look as follows:

mountain oak > oaks > trees

These lines can be sorted by their suffixes to keep terms of the same branch in the hierarchy together.

[7] This behavior needs to be specified in the accompanying DTD.

[8] The format in which WordNet's databases for Prolog are provided.

A recursive programming technique allows to set up this representation from the relational tables. Assuming a list of IDs that store a hierarchical relation such as immediate parent:

$ID1, ID2$ ($ID2$ identifies the node which is an immediate parent of the node identified by $ID1$)

In a first pass, all top level terms are identified by looking for those IDs that never appear on the left hand side. These top level terms are associated with a navigational path, consisting of the (virtual) root node[9]. Then, the list is recursively processed and for those pairs ($ID1, ID2$) that have a $ID2$ with a navigational path already associated, add the term identified by $ID2$ to this path at the beginning and associate it with $ID1$. Leave the other pairs for a next pass. This is done until no pairs are left.

7.5. Negative terms

Setting up a term maintenance system also requires a component that handles rejected candidate terms. The *garbage collector* of a term maintenance system ideally discriminates between those candidate terms that have been rejected under the requirements of a specific task (for example looking for terms that belong to one category and rejecting those that do not fit for this category) from those candidate terms that are universally rejected. To illustrate these garbage lists: If the term maintenance system is supposed to manage a database of product and service types, such universally rejected terms could be for examples geographical names. Storing these terms in a garbage list guarantees they do not turn up again if importing new term collections. If this garbage list is implemented in such a way that the an item in it does never show up when inspecting term collections, it has to be made sure that no potential bona fide term sneaks into this list. A remedy to this issue are regular inspection cycles that especially focus on those term candidates in the garbage list that were most often prevented to show up.

Usually, it are either very spurious term candidates or overly general terms that are rejected in a broad range of term management tasks. As only latter group has candidate terms in it that appear recurrently, the garbage list will consist mostly of these. Typical examples of such overly general terms in TE-Commerce are generic properties of companies, such as *24h open*, *credit cards accepted',* parking lots availabe". These terms do not tell much about the line of business companies follow.

For purposes of assigning keywords to businesses' websites according to their line of business it is necessary to remove these generic properties. Among very frequently occurring terms that have to be removed are:

542 click here
525 rights reserved

[9] The technique works the same way whether there is one ID at the very top of the hierarchy or more than one.

520	email address
516	further information
507	privacy policy
506	high quality
504	head office
501	telephone number
500	industrial estate
491	company name
490	customer service
487	post code
484	postal address
484	further details
483	enquiry form
480	web design
480	privacy statement
478	general information

A last segment of negative term candidates are adult and other offensive terms. While these terms play a considerable role in E-Commerce, it is crucial to segregate them from non-adult terms in order to prevent any unwanted exposure of adult content. The bulk of adult vocabulary can be gathered rather easily, given that in general if a meta keyword line contains an adult term, it will contain more than one and also very frequent ones (such as *sex*). Starting with few seed terms, it is thus possible to gather a large of adult terms that can then be shielded.

Although the distribution of adult terms has once again a long tail of seldom used terms, these terms in general co-occur with frequently used adult terms. Moreover, they often contain morphemes that occur within other adult terms as well. However, one issue that cannot be sufficiently solved with this approach are person names in the adult industry, as the number of these names outrank that of generic adult terms and cannot be detected by any morphologic indication[10]. In this respect, it is only through gathering and updating large lists of names that one can hope to detect this part of adult vocabulary.

[10] With few exceptions, such as made-up names that end in two x (e.g. Lisa Sparxx).

8. Orthographic Matches

This chapter is a brief treatise on orthographic term variation and its relevance for TE-Commerce. Considering that these variations can be observed, examined and even treated independently from all other variation principles, they build a base level of the different variation principles.

Normalizing on a semantic level without first normalizing orthographically is certainly not feasible[1]. Relating *laptop*, *notebook*,*labtop*,*note book* with each other is clearly redundant. Instead of relating four items with each other, it is sufficient to relate the normalized forms, laptop and notebook, and factor out the spelling out of these forms. Handling orthographic variation hence has to precede all further variation principles.

In the following section the orthographic matches and variations are presented and discussed through examples, figures on the extent of this variation and highlighting challenging issues in its handling.

Several types of orthographic variation can be differentiated through the origin of the modified spelling. One reason for orthographic variation lies in rivaling spelling conventions, either if canonic or just widely adopted. Another reason is the limited presentational spaces (truncations) that are especially important for printed matter. Finally non-inteded misspellings, i.e. typos, have to be taken into account.

While all of these issues can be resolved by the same principle of edit-distance, the fine-tuning of an orthographic approximate matching requires acknowledging their specifics.

8.1. Spelling conventions

Spelling conventions can be responsible for orthographic variation if they are not universally shared by writers. This happens especially if a change in orthographic norms take place, as for example in German-speaking countries since 1996 ("Rechtschreibreform"). The co-existence of spelling conventions will continue for some time, even if only one convention is declared to be correct. In many cases, however, the new orthographic rule system is rather more permissive than the old one, allowing several variants in many cases where only one was correct previously.

Another factor that contributes to systematic orthographic variations are limited keyboard capabilities, for example keyboards that do not support entering a ßor umlaute or at least impose more keystrokes to enter those letters than for the standard

[1] That is not to say that exactly this is done, albeit very uncontrolled, in some unsupervised semantic matching approaches.

range of A-Z letters. The amount of variation which is due to these systematic variations has already been examined (see above, Part A, Termspaces). In the same chapter, the similar type of variation that is due to the use of spaces and hyphens has also been covered.

8.2. Truncations

Truncation of terms as one way of how a term can be orthographically perturbed occurs in TE-Commerce applications mainly because of two reasons.

One source of truncated terms are formats with a fixed maximum length for a field. Such limitations are now largely historic, but still there are large data repositories that contain fixed length fields with truncated entries. For example, in printed telephone book the space for a single entry is limited. In the case of paid listings, advertisers usually are invoiced on a line basis. This results in a wide usage of abbreviations such as truncated forms or acronyms in all fields, especially in the paid add-on fields that are billed by lines or chars.

Sometimes, truncations might be due to the search logic that allows a wildcard search. For example, the TG Yellow Pages search log shows traces that a wildcard search was implemented on the search site. The frequency of truncated forms is much higher than could be expected if looking at the GY search log:

GY
95.795 Zahnarzt
11 Zahnarz
7 Zahnar
7 Zahna
Ratio of truncated form to non-truncated form 1 : 3800

TG
28.036 Zahnarzt
16 Zahnarz
18 Zahnar
6 Zahna
Ratio of truncated form to non-truncated form 1 : 700

GY
57.783 Rechtsanwalt
5 Rechtsanwal
0 Rechtsanwa
12 Rechtsanw
Ratio of truncated form to non-truncated form 1 : 3400

TG
8.621Rechtsanwalt

3 Rechtsanwa
6 rechtsanwa
13 Rechtsanw
Ratio of truncated form to non-truncated form 1 : 410

Even though the wildcard functionality is not communicated on Telegate's search site (`www.11880.com`) it is observable that at least some users found out about it. The numbers for the truncated forms are so few, though, that this is not a clear-cut case. However, apart from the ratio of truncated form to non-truncated form that is one magnitude higher in TG, it is also individual gaps such as the zero count for *Rechtsanwa* in GY that corroborate the hypothesis that the wildcard search indeed affected the user habits.

8.3. Intended Misspellings

Some words, especially foreign words, are frequently misspelled. Although only one variant is correctly spelled, other variations are similar frequent. For example, *accommodation* (correct spelling) appears 191 million times on google.com[2] and *accomodation* still 25 million times.

A further examples of misspellings that can be found with comparable frequencies in the YO query log are *geneology (184392) – genealogy (348442) – geneaology (21539).*

Many intended misspellings have the same or at least a similar phonetic expression than the correct version. Thus, they can be captured via graphem-phonem matching algorithms such as SoundEx. For certain applications, it might pay off to list the most common intended misspellings as aliases in order to save the resources of approximate matching to unsystematic errors.

8.4. Keystroke errors

Repairing keystroke errors does not only address computer novices who are not accustomed to typing on a keyboard. In fact, one of the results from lab experiments (see above, Part A) is that typographical errors occur more often for expert users. As they still perform faster searches than novice users, they either rely on spelling correction mechanisms or are able to correct their input quicker.

The basic errors that can occur when typing in text are omitting a key, inserting a spurious key, double-clicking on a key, hitting the key for the next letter before typing the current letter and finally hitting an adjacent key. These error types resemble of course the standard basic edit operations.

A possible modification of edit distance that could be considered is to penalize hitting of adjacent keys lesser than the general substitution of a letter. It can be demonstrated that indeed the adjacent keys on usual keyboard layouts (either QWERTY or QWERTZ) are more frequent than other letter substitutions.

[2] As seen on 1st of October 2006.

By a small script, pairs of words were extracted from a large German Yellow Pages query log, so that they differ solely by the substitution of one letter through a different letter. Thus all pairs differ by an edit distance of one. Then the frequencies for misspelled entries through substituting with an adjacent key are compared to the frequencies of entries that underwent a different substitution:

restaurant
3 testaurant
! 7 rastaurant
20 reataurant
8 resraurant
! 1 reszaurant
! 7 resteurant
! 17 restourant
3 restairant
31 restautant
! 14 restaurent
! 12 restaurand
10 restaurang
! 1 restaurank
3 restauranr
! 3 restaurans
! 2 restauranz

The entries marked with an exclamation mark cannot be traced back to adjacent keys on the keyword. A substantial part (out of) of misspellings, however, obviously stems from hitting the wrong adjacent key. This ratio is even heightened if those misspellings are factored out that seem to be due to intended orthographic variants (such as *restourant*), which could be recuperated by phonetic search.

It seems worth testing how the integration of the notion of phonetic equivalence and keyboard proximity enhances edit-distance based approximate search solutions.

8.5. An approximative matching framework

While current Search Engines let some typos go unnoticed, problems arise, however, when the hammering distance between two bona fide words is too small (for example: *gold, hold*). As bona fide words do also include rare words and proprietary names, there is no single formula to get hold of them. Neither a strict frequency threshold is sufficent to separate misspelled items from bona fide ones nor does an electronic lexicon suffices without comprehensive lists of proper names, including brand-new product names. Choosing corpora of carefully edited texts might help, but in every kind of text with a sufficient size there are some typographical errors in it.

In general, both the average word length and (somehow correlated to it) the range of available characters determine how much can be achieved through edit distance methods. The longer the average words are and the less characters are in the alphabetic system (the less information one character bears), the bigger are the opportunities for recuperating misspellings. To some extent, this can be done by pre-calculating errors systematically — for example, by creating an automaton based character insertions or deletions. In addition, the most common misspellings can be incorporated into the lexicon.

8.5.1. Approximate matching options

At a basic level, the parameters of approximative matching are the threshold on similarity and the weighting of deviance[3]. As was illustrated above for the example of adjacent keys, it often makes sense to discriminate between different spelling deviations in terms of different degrees of punishment[4]. On a more advanced level, the ranking of possible approximate matches can be calculated by additional values such as plausibility (for example, boosting the more popular terms).

External ranking systems can help to order suggestions by plausibility. For example, if looking up city names, the number of residents is a possible boost for a suggestion.

A final parameter on the ranking of possible approximate matches is a break that sets a maximum distance in quality between matches that are displayed. If the quality is measured by a value from 0 to 255 (the latter indicating a perfect match), a break of 10 means that the maximum difference between two approximate matches that are displayed is below 10.

Other options affect the set-up of the index for the approximate matching. The index can be created on a word-per-word basis or alternatively on a per-line basis. If the index is created on a word-per-word basis, it his requires to split up an input string at word delimiting chars. This allows to handle perturbations in the ordering of words.

On a per-line basis, the space char is treated just as any other alphabet char. While this set-up requires that the ordering of words must not change, it is then possible to detect additional spacings within words.

In many approximate search solutions, deviant forms in the suffix of terms are less penalized than those in the prefix of terms. For example, the Celebros-powered search on `www.ice.com` allows to search with *rinh* for *ring* without any need for interaction. It asks the user whether to correct the search into *ring* if the query *rung* is entered, which has the same edit distance than the previous query but a shorter common prefix. Finally, a query for *ting* does not produce any results at all.

[3] See also the documentation accompanying Exorbyte's MatchMaker software.

[4] Another exemplary application is post-correction of OCR results — in this case it is often useful to punish less the typical errors in OCR such as erroneously recognizing a ligature.

8.5.2. How to discern correct terms

Helpful resources for the purpose of setting up a set of list of bona fide terms are electronic lexica, lists of named entities — in the context of TE-Commerce, this should include lists of products, brands, vendors etc. as described above — and frequency information. Frequency information are ideally be taken from curated corpora such as newspaper.

The usage of frequency information for approximate matching is twofold. Firstly, it can be assumed that a term appearing frequently is in general a bona fide term. Secondly, if two terms are closely related via edit distance and both appear frequently, an approximate matching between is probably erroneous.

To avoid conflating correct distinct terms it is also feasible to resort to co-occurrence information. Genuine misspellings often occur together in meta keyword lines (in this case, used intentionally, of course), while unrelated items with a low edit distance will do so significantly less often.

It is thus possible to detect the following pairs of terms in which both are bona fide terms:

hacken – haken
tuch – touch
sitze – size
düngen – dünen
tuning – tunning
raucher – rauscher
stollen – stolen
kneipen – kneippen
taucher – tauscher
hasen – hassen

8.5.3. Examples of spelling variation

The following examples illustrate spelling variation of a generic vendor term (*Nagelstudio*) that were found in the TG-YP query log.

additional letter

- nagelsstudio
- nagelsdtudio

substituting a letter

- nagelstudia

- nagelstidio
- magelstudio
- nägelstudio

omitting one letter

- nagelsudio
- naglstudio
- ngelstudio
- nagelstdio
- naelstudio
- nagelstudi

omitting two letters

- nagelstudo

Below are additional examples of a spelling variation on the term *Lufthansa*, including intended misspelling (*Lufthanser* - a phonetic equivalence) and typos (*Lufthanza*) in YG:

- luftansa
- lufthanza
- luthansa
- lufhansa
- lufthanse
- luft hansa
- lufthanser
- luftgansa
- lufthasa
- luftthansa

- lufttansa
- lusthansa
- luftahansa
- lifthansa
- lufthunsa

9. Morphological-syntactical matches

9.1. Morphemes, inflection, derivation and compounding

9.1.1. Morphemes as proprium of terms

Morphological-syntactical variations contribute greatly to term variation in general. Indeed, there are arguments as to why morphological-syntactical variation belong to the *proprium* of terms. As was laid out in the previous chapter, the spelling-out of terms is handled by general orthographic principles and does in general not belong to the level of individual terms. Typographical errors or alternative spellings may obstruct the recognition of a term, but they do not change the nature of this term. Semantic variations of terms are by and large much more disputable, as the concrete lexeme in use tends to differentiate its usage in separation from other lexemes. Many speakers will challenge the synonymy of, say, notebook and laptop or seats and chairs. They might conceptualize these terms as co-hyponyms or as belonging to different registers. There are only few cases of generally accepted synonyms. In contrast, the equivalence of a morphological-syntactical variation such as *services of consultants – consulting services* can hardly be disputed.

Moreover, the concept of handling term variation used here also is based on morphemes as a central unit of term constitution. Despite it its quality as a constituting unit in term building, Sense-Morphemes follow a modified logic than morphemes in the item-arrangement model[1]. Sense-Morphemes are not used to explain word formation and are not supposed to convey a grammatical meaning — there is no Sense-Morpheme for the meaning {PLURAL}. Their place of application lies in detecting and describing variation of terms and how people recognize a term as a variation of something seen before.

On a cognitive level, Sense-Morphemes work as short-cuts to term recognition. Through spotting the recurrent morpheme in two forms such as *cleaning personal* and *cleaners*, their relatedness can be immediately inferred. The cognitive commitment that is needed for following this association is comparatively low, while it is much higher if synonyms using other morphemes are involved. For example, *braid* and *lace* might be synonymous, but it requires certainly more involvement in detecting their relatedness than to detect the relatedness in a pair such as *lacing* and *lace.*

[1] See [Hockett 1954].

9.1.2. Introducing Sense-Morphemes

Sense-Morphemes are to be understood as combinatorial units that carry the main aspects of a term's meaning. The usage of the terminus *morphemes* is inspired by the works of Zellig Harris, namely by his calculus of sentence transformations and the identification of boundaries through peaks in the successor variety[2]. The notion of distributional properties as being the main characteristics of a linguistic unit reverberates in the Head / Container calculus described below[3]. In analogy to how Harris introduced a kernel of simple sentences that can undergo unary and binary transformations that preserve the meaning, the different kind of TE-Commerce term variations will be introduced. Moreover, in a similar fashion in which lexical morphemes are used as a building block for elementary sentences and their variations — such as permuting the morphemes or introducing grammatical morphemes — Sense-Morphemes are the building blocks of terms in TE-Commerce.

The two key properties of a Sense-Morpheme are its robustness in meaning and its variability in expression. The first aspect refers to the quality of Sense-Morphemes in carrying a meaning that is not substantially affected by the context in which it appears. For example, the Sense-Morpheme *[medizin]* preserves its meaning in all of the following occurrences in longer terms:

- sportmedizin
- medizintechnik
- tiermedizin
- medizinstudium
- medizinlexikon
- zahnmedizin
- arbeitsmedizin
- medizinisches wörterbuch
- reisemedizin
- notfallmedizin
- medizinbedarf
- alternativmedizin
- gerichtsmedizin
- medizinprodukte

[2] See [Harris 1951], [Harris 1955].
[3] See [Goldsmith 2001].

- medizinisches lexikon
- chinesische medizin

This does not imply that *medizin* is the most important constituent in all of these occurrences — for example, *medizinisches wörtberbuch* is foremost a *wörterbuch*(dictionary). Yet in all examples, medizin preserves the meaning of isolated occurrences in larger terms. As the avid reader might remark, this is certainly not true for all compounds, or at least not to the same degree. A *Medizinball* has only vague associations with Medizin. Aspects of how compounding affects the meaning of one or more of the compound constituents are discussed below in this chapter.

The second aspect of Sense-Morphemes, their variability in expression, allows to grasp many different apparitions of a term through one Sense-Morpheme.

Genuine Sense-Morphemes are not necessarily one-word-terms or parts of words. Many Sense-Morphemes that can be found consist of more than one word.

Given that Sense-Morphemes build immediately convincing relationships, they make it possible to demarcate a group of related terms and filter them for any outliers. Conversely, if a user is presented with a term collection that should reflect the basic concepts of a category, missing Sense-Morphemes will be almost immediately visible. For example, if the category *Jagd* is going to be enriched, a missing Sense-Morpheme such as *Wild* will be immediately recognized.

9.1.3. Inflectional morphology

The most basic level of morphologic variation is flexion, and plural-singular variants form the most frequently used part of flexion. Although electronic lexica are available that cover either via explicit listing or morphological analysis almost all forms that are written correctly, finding singular - plural variants of terms needs in some cases special consideration. Three different challenges are detailed in the following sections: problems of lexical ambiguity, restrictions of numerus forms and the issue of MWUs.

E-Commerce terms are often not listed in an electronic dictionary, for example terms such as *iPod.* In these cases, their inflectional morphology usually follows the most productive patterns in a language.

Paradigmatic restrictions on full forms

Even if the electronic dictionary lists the E-Commerce relevant terms, it is not guaranteed that they can be simply expanded to all full forms of the paradigma. Several challenges have to be taken into account. Firstly, some full forms do not occur because of paradigmatic restrictions. Consider for example the pair *führung – führungen.* While the base form can convey both senses guided tours and management, the plural form does not. This prevents a complex term such as *Marktorientierte Unternehmensführung* to appear in the plural.

Another issue orginates in seldom used singular or plural forms, for example *büro – bureaux* or *tuben (pl.) – tuba (sg.).* Without checking the frequency distributions of the generated or analyzed units, these errors will lower the value of applying an electronic dictionary.

A last challenge is posed by complex noun phrases. In theory, these should be inflected by modifying the head and leaving the container. However, again paradigmatic restriction may occur that prohibit the inflection of the head only. Consider for example, *Meldestelle des Einwohnermeldeamtes.* Its correct plural has both parts pluralized: *Meldestellen der Einwohnermeldeämter.* In other cases, it is not even clear what the correct plural could be, for example for the complex term *Schuh in Übergröße.*

9.1.4. Affixations

One of the most prolific forms of term variation is affixation. Adding bound morpheme to forms is responsible for a large part of variance. Its importance for TE-Commerce largely stems from the capability of affixation to change categories. For example, adding a suffix *-er* to produce an agent noun is a common way in TE-Commerce to derive a profession from an activity (”print” *rightarrow printer*). As was laid out, TE-Commerce as a field is defined by the participating agents and their interactions with each other.

Affixation can be controlled by part-of-speeches of base form and resulting form. For this purpose, a database of more than 250 common German affixed have been accumulated and described in the following manner (here for the example *heit*)[4]:

-heit
art: SUF
kat: N(fem)

Through defining the part of speech of base and derived form and checking them through the electronic dictionary, the number of erroneous segmentations can be kept low:

- A+e → V: A-1+heit (fadheit)
- A+heit → V: A+heit (fiesheit)
- V (Part II) → V+heit (besonnenheit)
- N → N+heit (christenheit)
- NUM → NUM+heit (dreiheit)

.

Two interesting special cases shall be highlighted in this context. Firstly, major changes in the base form of an affix that occurs with Latin bases. These are captured by applying a regular expression to the lexicon and conflating pairs that are related

[4] This list has been assembled using both [Fleischer/Barz 1995] and the IDS-Grammis framework, see [Donalies 2002].

module the change in the base form (for example *X-z-ieren* vs. *X-k-ation*) and than filtering spurious entries. Examples of this include

publizieren → publikation
reduzieren → reduktion
erodieren → erosion
konvertieren → konversion
qualifizieren → qualifikation

Secondly, genuine German morphological changes to the base form also need to be taken care of in the system. For example, consider these lines with an *e* omitted in the derivated form:

wechsel → wechsler
moebel → moeblierung
basel → baslerisch
pinsel → pinsler

For most practical purposes of TE-Commerce, it is sufficient to precalculate allomorphes such as *wechsel/wechsl* for a limited number of patterns.

Finally, a negative list of non-valid segmentations is needed to avoid an analysis such as *astern = a-stern* or *regieren = re-gieren* which is not detected by the part-of-speech restrictions in the affix definitions[5]. By working through a frequency list of segmented forms, it was possible to set up a negative list of about 1.000 non-valid segmentations that allows for an acceptable precision with regards to real-life TE-Commerce data.

9.1.5. Compounding

German displays a large amount of productivity, especially through compounding. A considerable part of newly created compounds stem from commercial relevant fields, with advertising and product naming as an important source of neologisms.

A test of the current decompounding capabilities of search engines can be done by trying to retrieve a specific page by a part of a compound that appears on this page. The test page was `www.pilzbuch.de/site/buch/pilz081.html` which contains the compound *Satansröhrling* in the page text. The page code does not contain *satan* or *röhrling* as isolated words. The phrase search “Maschennetz, kugelig, später dickbauchig, 5 -12cm hoch” produces this page as the only hit (as tested on Nov. 1, 2006). The test queries add a part of the compound or an inflected form of this part to the phrase. If the page is still retrieved, the search engine uses decompouding:

[5] Arguably, it would be possible to detect such errors by introducing a feature LATIN / NON-LATIN to base morphemes. However, this feature is not available for the majority of lexicon entries.

Results:	Google	Yahoo (DE)	MSN
Page in Index	+	+	+
Found by phrase and satan	-	+	-
Found by phrase and satans	-	-	-
Found by phrase and röhrling	-	+	-
Found by phrase and röhrlinge	-	-	-

It is only Yahoo (and moreover, only the German language version of Yahoo) that is currently able to split the compound[6].

Smaller Search Engines in the German market (`Ask.de, Web.de, Lycos.de` or `seekport.de`) do not have this particular page in their index, but additional tests showed their incapability of decompounding.

Some of these compounds have lexical ambiguous heads, which makes it difficult to create the plural form for a given singular. As a consequence of Zipf's distributional law, not all inflected forms to a citation form will be found in the corpus, regardless of its size. In practice, though one can look-up the correct plural for *Rockband* — *Rockbands* (and not *Rockbände or Rockbänder*), there will be some bands without an observable plural form.

The major part of the lexical ambiguities does not need to do any harm in applications, as one of the variants is much less common and can the other can then be safely ignored. One way to find out these distributional patterns is to switch between right and left extension and cluster the results. Given that *Rock* is in the music cluster (by sake of *musik* and all other similar terms that combine with *rock* to build a compound, it is clear that the correct plural is *Rockbands.*

9.2. Permutations

In general, modern Web search engines treat the combination of query words as Boolean AND-operators. Furthermore, apart from phrasing, the order of words does not play any role in determining the search result: The query *new york hotel* and *hotel new york* yields largely the same result [7]. While the phrasing operator allows to group words into non-breakable units — allowing to shield multi-world units such as *New York* – there are cases, however, when the order of units needs to be sanctioned. Examples of such bracketing phenomena are *junior high school* that means a different thing than *high school junior*, because the head of the phrase differs.

> avi umwandeln in mpeg [14] → mpeg in avi umwandeln [24] → avi in mpeg umwandeln [50]
>
> audio cd cover [189] → cd audio cover [26] → cd cover audio [25]

[6] This is not done via a substring operation, because neither adding *röhrlin* or *satansr* to the phrase delivers any results.

[7] A notable exception are the flash-ins. On `www.google.com` the query *new hotel york* is indeed interpreted by the Local shortcut as a search for a new hotel in York, as seen on Nov. 1, 2006.

london cheap accommodation [10] → cheap accommodation london [193] → cheap london accommodation [103]

messe ambiente frankfurt [12]→ messe frankfurt ambiente [23] → ambiente messe frankfurt [40]

It is interesting to note that even in the last example which relates to a named entity, permutations can be observed without disturbing the identification of this particular named entity.

There are some other cases,however, where the ordering of terms is almost equally distributed but bears a different meaning. In these examples it can be assumed that the look-up direction is divergently specified:

wörterbuch deutsch englisch online [11] → wörterbuch englisch

deutsch online [12] → deutsch englisch wöterbuch online [8]

englisch deutsch wörterbuch online [17] → online wörterbuch

englisch deutsch [27] → wörterbuch online englisch deutsch [18] → online wörterbuch deutsch englisch [30]

What occurs here is that an underlying predicate *translate* has two arguments in ordered slots that are of the same type, in this case languages. Similar cases are travel itineraries for which the arguments "arrival place" and "destination place" are also of the same kind. Pragmatically, it often makes not that much of a difference if the order of arguments for such predicates is preserved, as for example one flight route usually applies to flights in both directions.

9.3. Grammatical insertions

Grammatical insertions refer to the usage of syntactic fillers, mostly prepositions, in terms. Variants of this type are exemplified by pairs such as *hotel new york* vs. *hotel in new york* or *chromosome abnormalities* vs *abnormalities in chromsomes*[8]. In general, insertions or deletions of syntactic fillers does not change the term meaning, but is rather a matter of term stylistics.

In the world of generic TE-Commerce with its skew towards compact term representations, the usage of such grammatical insertions is rather limited. It plays a decisive role, though, in specialist domains.

A case study of the usage of the preposition *in* within the YG query log in approx. 5.000 lines revealed that in 383.859 cases it was used to in a topological sense, and in 196.384 cases used in a different distribution (either temporal or figurative). While it often appears too demanding to automatically grasp the intentional content of such a preposition in an individual query (on today's Web Search Engines, they are removed except in phrasal searches), it is feasible to use the query log to extract information

[8] See [Jacquemin 2001].

through such syntactic constructs. For example, the terms that occur after *in* and also frequently isolated in queries are very likely geographical names.

9.4. Reduction forms

A common issue that hassles search and text processing in general is the existence of reduction forms. Two types of reduction forms have to be discriminated. The first type refers to a reduction on the level of expression while preserving in an unambiguous way the meaning of the full form. This happens for example with company names that have their legal title as full form, but are commonly referred to by a reduced form (*Microsoft Corp. rightarrow Microsoft* or *MS*). The second type of reductions include the notion of an implicit reading. For example, take the query *berufsunfähigkeit.* Obviously it was not issued because someone wants to become invalid, but as a reduction form for information on or insurances for *berufsunfähigkeit.*

Other cases where a reduction form has an implicit facet are for example queries that just contain a named entity (*Britney Spears*) and that probably can be expanded to the full form through adding *information on* or *resources related to* (see also Part C, "Query Types and how to recognize them")[9].

Reduction forms of this kind are very numerous in the top frequent searches. One way to treat them is to look at their expansions in the term collection. Details on this procedure are discussed below (see Part C, "Head/Container principle for search queries"). For example, the frequent expansions of *berufsunfähigkeit* — including expansions in the form of compounds such as *berufsunfähigkeitsversicherung* or *berufsunfähigkeitsrente* — indicate that the assumed reading that was presented above has indeed substance.

A morphologic method to detect the first type of reduction forms is to segment compounds consisting of two and three parts and relate triples A, B, C to pairs A, C. The examples below show unambiguous reduction forms of such ternary units:

trickfigur – trickfilmfigur
privatsitz – privatwohnsitz
treppenhilfe – treppensteighilfe
steuerfonds – steuersparfonds
betriebsdefizite betriebskostendefizite
straßenzeitung – straßenverkaufszeitung
billigarbeiter – billiglohnarbeiter
grüngestaltung – grünflächengestaltung
gemüsegeschäft – gemüsefachgeschäft

The last two examples show that more than one expanded form to a reduction form may exist:

[9] A special case is provided by reduction forms such as *Picasso* for a *painting by Picasso.*

radprofis – radrennprofis / radsportprofis
notdienst – notarztdienst / notrufdienst

10. Semantic matches

Term variations on a semantic level are apparently not only the hardest to catch, but it is also be the hardest to meet unanimous agreement. For some, the very notion of terms that look different but mean the same is disputable, as was pointed out above in the section about term handling in Web 2.0. The chances of finding truly substitutable expressions for the same concept with the same distributional properties are indeed low. In many cases, what is accepted by some speakers as genuine synonyms, is rejected by others, not at least those with a professional stake in the domain of the synonym candidates. For example, *laptop* and *notebook* are used by many speakers in an almost fully interchangeable way — as can be proven by looking at their distribution in query logs —, but are sometimes considered to be distinct product types, especially among manufacturers (usually marketing *notebook* as a product with an even higher portability)[1].

In the approach followed here, semantic variations form one part of multiple term variation patters, putting it into line with morphologic, orthographic and syntactic term variations. Handling semantic variation requires finding a viable path between the subtle differentiations of semantic relations that can be made on a per-example basis and what can be extracted in an operational way from textual resources, including lexica and thesauri. It will be laid out that neither pure statistical methods nor heuristics nor application of lexical resources are alone sufficient to deliver precise and covering results for capturing semantic variations in the context of TE-Commerce.

Starting out with presenting a short overview on lexical semantic with a focus on classifying semantic relations, the suitability of these theoretical propositions for TE-Commerce is discussed subsequently. Various open repositories with instances of suitable relations (especially WordNet) will be discussed. A treatment of different approaches to create, refine and clean such repositories follows, including syntagmatic as well as paradigmatic patterns. Special attention will be given to the notion of distribution and co-occurrence (first-level and higher-level).

Semantic variations on the level of terms are a rather sensitive field to move in and their handling requires more precautions than morphologic or orthographic variations. While morphologic or orthographic mismatches (for example matching *walking stick* to *stickiness*) are at least comprehensible in terms of what they tried to match to a semantic mismatch bears the danger of total incomprehensibility.

The main needs that have to be covered by a semantic matching component in TE-Commerce applications:

[1] One test is to submit a query "notebook vs laptop", or any synonym candidates, on Search Engines. This generally produces discussions on the status of the pair and also voices that deny the synonymic relation.

- Incorporate the few almost universally accepted synonyms with a considerable frequency in TE-Commerce resources

- List related terms as a more relaxed relation compared to synonymy which can be used as a fall-through

- Incorporate lexical function, especially related to orthogonal terms

- Allow a drill-down and drill-up through the generality relation

10.1. Semantic relationships

Two principle approaches of describing lexical semantic relationships exist: The first approach (that was advocated in Generative Semantics) tries to establish semantic primitives and models relations through combining these primitives[2]; the second approach dismisses primitives and works with lexems and how they contribute to each other's meaning instead[3].

Apart from describing semantic relationships on a lexical level, it can also be sought to shed light on the meaning of a word by listing related words, an approach followed for example in tagging. Note that in practice, the difference between operating on words and on lexemes is that the representatives of the latter are morphologically normalized (for example, all in singular) and filtered for non-lexicon words.

While the two lexical approaches discriminate between different relations that characterize what lexemes have in common in terms of meaning, the latter approach typically only has one relation, namely a broadly defined *is related to.* Operationally, this relation can be extracted from distributional patterns — considering that synonyms should occur within the similar contexts — and syntagmatic patterns (for example, *X, also called Y* or co-occurrance in meta keyword).

10.1.1. Lexical semantic relations

In order to illustrate what is conventionally understood by lexical semantic relations, the seven main relations discriminated by Cruse in his standard textbook are examined as an exemplary set[4]. This set relations comprises hyponymy, antonomy, (cognitive) synonymy, taxonomy, meronymy, plesionymy and paronymy. The first three of these relations are intuitively understandable and their usage is not restricted to the sphere of linguistics. The latter four relations are of a more technical nature, as they draw a distinction between different subkinds of a broad concept of *relatedness.* Nevertheless, from a viewpoint of operational testing it is necessary to make a distinction between different types of relatedness, as only then a testing through specific distributional patterns is made possible.

[2] See [Jackendoff 1990].

[3] [Cruse 1986] calls this contribution to the meaning of other lexemes the *semantic trait* of a lexem.

[4] See [Cruse 1986].

Taxonomy can be tested by patterns such as *X is a kind/type of Y*, meronymy through patterns such as *X is a part-of Y* and plesionymy as a rest class through patterns *X is similar in meaning to Y*. In addition to these, paronomy holds between one item and a derived item of a different category, for example *paint-painter*.

These relations can be further subdivided, for example meronymy into the following three subtypes: member-of (*ship – fleet*), material -of (*clay – pottery*) and finally part-of proper (*finger – hand*). Typical patterns for relata of the two first subrelations would be *many X make up a Y* (member-of) and *Y is made out of X* (material-of).

Testing the relations through the distributional patterns of the relata instead of defining them on an ontological level has two main advantages. Firstly, it rules out that a relation is included that is not commonly accepted by speakers only on the grounds that a domain-specific taxonomy relates them to each other. Secondly, it helps to avoid bridging too many levels when stating a relationship. For example, if meronym and holonym are on levels of granularity that are too far away, such a relation should be avoided. To say that a *nut* is a mernonym of *ship* renders the meronymy relation almost useless in practical applications, although ships are almost without exception built with nuts — making it a correct meronym relation from an ontological viewpoint.

An even finer set of lexical semantic relation appears in the context of noun modifiers[5]. The list below illustrates the subtle differences that can be stated in the relation between modifier and head. This model can be transfered from its original application, i.e. classifying compounds, to classifying relations between two separate terms:

- *Agent*: compound is performed by modifier – student protest, band concert
- *Beneficiary*: modifier benefits from compound – student price, charitable donation
- *Cause* modifier causes compound – exam anxiety, overdue fine
- *Container*: modifier contains compound – printer tray, flood water, story idea
- *Content*: modifier is contained in compound – paper tray, eviction notice, oil pan
- *Destination*: modifier is destination of compound – game bus, exit route, entrance stairs
- *Equative*: modifier is also head – composer arranger, player coach
- Instrument: modifier is used in compound – electron microscope, diesel engine, laser printer
- Located: modifier is located at compound – building site, home town, solar system

[5] The following relations are taken from [Barker/Szpakowicz 1998]. One should also note the reduced set of only four basic relations that [Fanselow 1981] has worked out for German noun compounds

- Location: modifier is the location of compound – lab printer, internal combustion

- Material: compound is made of modifier – carbon deposit, gingerbread man, water vapour

- Object: modifier is acted on by compound – engine repair, horse doctor

- Possessor: modifier has compound – national debt, student loan, company car

- Product: modifier is a product of compound – automobile factory, light bulb, colour printer

- Property: compound is modifier – blue car, big house, fast computer

- Purpose: compound is meant for modifier – concert hall, soup pot, grinding abrasive

- Result: modifier is a result of compound – storm cloud, cold virus, death penalty

- Source: modifier is the source of compound – foreign capital, chest pain, north wind

- Time: modifier is the time of compound – winter semester, late supper; mowing class

- Topic: compound is concerned with modifier – computer expert, safety standard, horror novel

From the viewpoint of TE-Commerce these relations have to be modified to apply them fruitfully. For example, the difference between *engine repair* and *horse doctor* is quite remarkable — the first case is a typical extension in TE-Commerce to a term describing a device and one that can be applied systematically , while the latter case does not relate much to a conventional *doctor*. The *Topic*-relation exhibits a similar difference. While *computer expert* is the product of a systematic extension to the device *computer* and follows a similar pattern than *engine repair*, a *safety standard* or a *horror novel* show a totally different pattern. The situation is even worse for a relation such as *Product* (that sounds most interesting for TE-Commerce purposes). While an *automobile factory* shows a productive pattern in TE-Commerce — deriving the place of manufacturing for a product type — a *colour printer* is really a facet of a product type, whereas decomposing *light bulb* does not make much sense at all (apart for the purpose of explaining the reduction form *bulb*).

This is just a brief walk through these finely granulated relations. The point to take here is that these relations developed for describing word formation have to be modified before they can be successfully applied to TE-Commerce.

10.1.2. TE-Commerce lexical functions

The formalism of lexical functions is based on the combinatorial properties of lexemes. One of its applications is to build up dictionaries with very fine grained entries that describe meticulously the different expressions that can be constructed by using a lexeme. For example, the MAGN-function applied to a lexeme specifies an adjective or a phrase with adjectival function that conveys a greater intensity of the main quality of the lexeme, thus *Magn(tea)=strong.* As the values of lexical functions are more often than not hardly predictable and idiosyncratic to individual lexemes, they have to be specified in the lexicon[6].

The set of lexical functions described currently contains about 300 different functions, ranging from broad functions such as MAGN to very specific ones, for example SON (typical sound of an object). For the purposes of TE-Commerce, this set has to be adapted and modified.

The newly introduced TE-Commerce lexical functions center around the concept of a business domain and its possible facets (see also above, Part B). For example, *hotel* builds up a domain as can be observed by the expression of facets such as humans involved in the domain (*hotel manager*), activities (*booking a hotel room*), concrete nouns (*hotel furniture*), properties (*five-star hotel*) etc. As with general lexical functions, TE-Commerce lexical functions show both systematic as well as idiosyncratic values. The following examples are German, but given that lexical functions have be proven to apply to several languages, the findings for TE-Commerce lexical functions should not differ largely between different languages[7].

For example, a person specializing in a particular domain can systematically be build up by adding -spezialist or -fachmann ("Computerspezialist", *Reisefachmann*, *Chemiespezialist* etc), which is possible for a large range of domains. However, some domains have a additional different specification for specialists, as can be seen by examining words such as *KFZ-Meister*[8]

The standard way of applying a TE-Commerce function is to start by the branch head that spawns the different types of entities and relations that are part of a branch, just as described above. For example, the concrete noun *stein* is a branch head that can enter a TE-commerce lexical function such as WORKER which yields a set of occupational titles for this branch: *steinmetz*, *betriebsmeister steinindustrie*, *dipl.-ing. steine erden*, *steinschleifer*, *steinbearbeiter*, *steinsetzer* etc.

Examples of TE-Commerce lexical functions comprise:

- *Humans*

- WORKER, a human that produces or offers services in the branch (WORKER(stein) = steinmetz)

[6] See [Wanner 1996].

[7] Most work on lexical functions has been done in Russian and French through the ECD, but there are also papers covering for example Spanish, German or Korean.

[8] This touches upon legal regulatory practices. Nevertheless, it is also the value for a TE-Commerce lexical function which can be paraphrased using *a specialist in.*

- MANAGER, a human who is in charge of a business in the branche (WORKER(stein) = steinmetzmeister)
- HUM_METAPHORIC, the metaphoric name for professionals in the branch (HUM_METAPHORIC(IT) = IT-Guru)

- *Organizations*
- BRANCH, the organizational structure for businesses in the same field (BRANCH(chemie) = Chemiegewerbe, Chemieindustrie, Chemiebranche)
- PLACE where an activity takes place (PLACE(Jura) = Rechtsanwaltskanzlei)
- VENDOR that sells a product or offers a product (VENDOR(tabak) = Kiosk)
- MANUFACTURER that produces or assembles a product (MANUFACTURER(Elektroenergie) = Kraftwerk)

- *Concrete objects*
- PRODUCTS in a branch (PRODUCTS(Unterhaltungselektronik) = CD-Player)
- REPLENISH — a continuous supply that is needed to keep a device running (REPLENISH(Drucker) = Toner Cartridge)
- ACCESSORIES — additional components that enhance the functionality of a device (ACCESSORIES(computer) = Externe Festplatte)

10.1.3. Orthogonality in terms

For many branch heads there exist default values for the TE-Commerce lexical functions introduced above. These values are in general either the result of affixation or compounding of the branch head. The right hand segments of the resulting compound shall be called orthogonal terms as they combine with many different branch heads and carry only little or no topic-specific meaning.

For example, the TE-Commerce lexical function BRANCH is usually expressed by the following right segments in compounds

- *branche*
- bereich
- gewerbe
- industrie
- wesen

- wirtschaft
- business
- sektor
- system,systeme,systemen
- körper
- funktion
- infrastruktur
- *geschäft*
- fachgeschäft
- shop
- outlet
- store
- handlung
- großhandlung
- markt
- abholmarkt
- fachmarkt
- laden
- *geschäftsstelle*
- filiale
- zweigstelle

In some cases, an orthogonal term is ambiguous with a branch head. For example, *ausstattung* plays quite a different role in *Raumausstattung* than in *Babyausstattung* where it occurs as genuine orthogonal term and can be replaced by other orthogonal terms of the supply-type. The same holds for *Bau* which occurs as orthogonal facet in compounds such as *Fahrzeugbau*, but as a branch head in compounds such as *Hochbau.*

10.1.4. Instances-of

The instance-of relation connects named entities with categories or classes. As the process of gathering instances is typically a case of categorization — including overlaps and borderline cases — the following paragraphs take categories instead of classes as the basis of the instance-of relation.

If categories are organized hierarchically, the instance-of relation can be thought of as adding leaves to the branches. For example, the named entity *Helmut Kohl* is be an instance of the class politicians. *Samsung X-05* is an instance of notebooks.

Collecting instances might be done either by starting from the category name and systematically gathering instances — for example, looking for all politicians or all notebooks — or in a bottom-up way by looking at entities of the same kind and later looking for a descriptor that applies to all of them. The latter approach can be done by exploiting contexts of instances, such as query containers (for a detailed discussion of query heads and containers, see Part C).

A related issue was tackled in the KnowItAll project[9] It uses rule templates to detect instances from corpora. Starting form rule templates, a learning heuristic detects

> NP1 *plays for* NP2 & properNoun(head(NP1)) & head(NP2) = *Cosmos New York* →
> instanceOf(Athlete,head(NP1))

For example, through the sentence *The matured Franz Beckenbauer played for Cosmos New York* the relationship *Franz Beckenbauer is an instance-of an Athlete* can be established.

One way of obtaining instances-of is looking for heads appearing with the same or similar set of containers. For example, the following set of containers

X lyrics

X mp3

X wallpaper

X posters

X tabs / X guitar tabs / X bass tabs

X photos

X pictures / pictures of X

X memorabilia

X chords

X bootlegs

X ringtones

X music scores

[9] See [Etzioni et al 2004].

X discography

allow to extract a considerably clean list of bands from a query log. The results are sorted here by number of how many of the above listed containers they were found in.

There are several applications in TE-Commerce for the instance-of relation. One example is broadening a search by adding all instances of a class to a specific instance. For example, if searching for *Samsung X-05* does not produce results, other laptop models by the same maker can be tried as well. Taken the other way round, a query for *Samsung laptops* could be refined by listing the instances below this class.

True intensional query answering needs world knowledge of the sort *X belongs to Y*. It is only with such knowledge that a query *biography american presidents* could return a biography on Gerald Ford, even if the term *american president* is not or at least not prominently presented in the biography.

The main advantage of such an approach is that no conventional lexicon lists named entities. However, a demanding challenge of this approach is that most entities can be listed under many different headings. An interesting approach to handle multi-categorization of entities is researched at the CIS under the title *EFTG-Netz* that uses an intersection operator to finely describe how members named entity are related to topics and temporal-spatial embedding. For example, *20th century economic politics* can be decomposed as the intersection of the topics *economy*, *politics* with the temporal positioning *20th century.*

10.2. Spotting Similarity

What does it mean for two terms to be similar? Similarity can be understood in different ways, such as similar phonetic realization, sharing aspects of meaning, showing the same distributional properties, belonging to the same register and many other conceivable correspondences. Even if ruling out similarity that has nothing to do with the meaning of terms, a wide range of different relations could hold between similar terms.

Even antonyms have something in similar: both a valley and a mountain are topographic entities defined by elevation. From the standpoint of distributional patterns, the terms dog and cat are also similar as both are cohyponyms. Morphologically related terms such as *carpenter* - *carpentry* are doubtlessly similar. Even terms that are not connected on the level of language, but rather through a business logic can be called similar, such as forestry and paper or metal-sheets and car body. Another example of such loose associations are those pairs that are based on shared preferences, following the line *people who bought X also bought Y.*

Regardless on what grounds the similarity relation should be built upon, general properties of the relation $sim(a, b) \rightarrow [0, 1]$ (assuming normalized values between 0 and 1) between two terms a and b can be made specified with regards to symmetry, reflexivity and transitivity[10].

[10] See also [Weeds/Weit/McCarthy 2004].

Symmetry

A symmetric relation $sim(a, b)$ means that for every a, b $sim(a, b) = sim(b, a)$. Usually, this is what one would expect from a similarity relation. However, an asymmetric similarity value, as the alpha-skew value proposed by Lee[11], could make sense as it hints at the difference between hypernym and hyponym: An hyponym can be substituted by its hypernym in all contextes but not vice versa.

Reflexivity

Naturally, one would expect similarity to be reflexive, meaning that $sim(a, a) = 1$. Following what was laid out about the symmetry, it is conceivable, however, to let the number of meanings influence the value for $sim(a, a)$. For example, the value for similarity to itself could be used to indicate if a word is more polysemous (words such as *key*) than others. However, all conventional similarity metrics are reflexive.

Transitivity

If introducing a new binary relation between terms that are similar — through setting a threshold on the similarity function values — transitivity can be analyzed. If this new relation sim_B would be transitive, it would be true for all a, b, c that

$a_{sim_B}b$ and $b_{sim_B}c \rightarrow a_{sim_B}c$

Obviously, this cannot hold for any similarity relationship except strict synonymy. Otherwise it would be possible to move away a bit further from the meaning of the original term at each iteration (for example, *oar – ship – submarine – torpedo* etc).

The issue on what data such a metric could be tested is discussed below. Two different approaches are discriminated: syntagmatic patterns that look for similarity between terms that are both present in a text versus paradigmatic patterns for extracting similar terms when they do not occur together.

Syntagmatic patterns

Syntagmatic patterns apply to textual resources in which related terms appear together. In general, such related terms were put in there by one human, making them consciously deployed related terms. A prominent example of a syntagmatic pattern that can be used to extract related terms are co-occurrences in meta keyword lines. Here, very similar terms are intentionally used to capture traffic even if the users searched with different term variants. See below, "Enriching lexical resources"", for a discussion of how meta keyword lines can yield high qualitative lists of similar terms.

A further case of syntagmatic patterns are patterns that can be applied to prose text[12]. By starting with a seed list of related terms and using a template of what patterns should be extracted, it is possible to set-up a bootstrap approach[13]. For example, starting with the seed terms *laptop* and *notebook* and a pattern template N1 * * N2 (i.e. first noun, two arbitrary words, then the second noun) could yield the

[11] See [Lee 1999].

[12] See [Turney 2006].

[13] See [Hearst 1992] and [Hearst 1998].

pattern *N1 also called N2*, which again can be instantiated by many other synonymous terms.

Using free text (as can be retrieved from a Web search engine) bears the danger that all kind of terms are produced when applying such a pattern. Below are examples of what could be found using the Google query *auch * genannt*[14]:

- Gallen, auch Eichäpfel genannt
- Andromedanebel (M 31), auch Andromedagalaxie genannt
- M 33, auch Triangulumgalaxie genannt.
- Samenhülle, auch Tegmen genannt
- rattenzellengesteuerte hybride Roboter, auch Hybrot genannt
- Fantasiemasken auch Tüllmonster genann
- Keyword Dichte auch Schlüsselwortdichte genannt
- Direktversicherer wird auch Erstversicherer genannt
- Diabetes mellitus Typ 2 wird auch Zuckerkrankheit genannt
- Kreditzusammenfassung, auch Umschuldung genannt

While all of these pairs are valid (quasi)-synonyms, they are very inhomogeneous, not at least in their relevance for TE-Commerce. Moreover, a considerable number of instantiations of this pattern will yield spurious results caused by metaphorical or ironical usage. Restricting the texts to specialists' corpora could help to lower the amount of this noise.

Paradigmatic patterns

Paradigmatic patterns for the extraction of related terms can be compared to collaborative filtering. In the case of paradigmatic patterns the two related terms are not present at the same time, but appear in similar positions. Related terms found thereby could for example stem from differences in word usage, such as some people preferring *laptop* to *notebook* or vice versa. Through observing a large number of evidences paradigmatic patterns can harvest on collective wisdom. However, it has to be made sure that the terms in paradigmatic relation really are substitutable by each other.

One way of ensuring this is when they can be related to an external entity, as happens with anchor texts that point to the same page. Naturally, this is not a fully reliable case, as some web page creators might have set the link to the same page with different things in mind or with different stylistics — for example, both anchor texts *BMW* or *BMW Hompage* or even *der neue X5* could occur with links pointing to

[14] Tests conducted on Nov. 5, 2006.

www.bmw.de. A remedy that helps to narrow the space of possible intentions when setting anchor texts to a more lexical stance is to use Wikipedia.

Wikipedia authors may use anchor texts that are different from the title of of the article they are pointing to, The syntax for this is the double pipe symbol, e.g. [Nordrhein-Westfalen——NRW]. In this case, NRW appears in the text, but the link goes to Nordrhein-Westfalen.

By gathering of all these anchor text and cumulating them by the article title their links point to, a first list of possible related expressions can be gathered. As this contains too much irrelevant entries from the perspective of TE-Commerce, they have to be filtered which can be done by using MKW-DE as a filter list. The resulting list can serve as a good starting point to determine related terms. Here are examples for *hotel*:

hotelanlage hotel
hotelier hotel
hotelkette hotel
hotellerie hotel
hotels hotel
liebeshotel stundenhotel
luxushotel hotel
luxushotels hotel

In a third step, the most frequent co-anchor texts are enriched by other meta keywords that co-occur with these frequently in MKW-DE. The rationale behind this is to combine the paradigmatic patterns of co-anchor texts with the syntagmatic pattern of co-occurrence in meta keyword lines. Through this, clusters of highly related terms can be extracted. A sample of these clusters (each line starting with the most frequent term in MKW-DE):

hotel
hotels, urlaub, reisen, ferien, pension, reise, ferienwohnung, tourismus, ferienwohnungen, zimmer

design
webdesign, internet, werbung, homepage, grafik, web, gestaltung html, multimedia, flash

architekt
architektur, bauen, planung, architekten, bau, sanierung, haus, bauleitung, umbau, neubau

auto
kfz, gebrauchtwagen, pkw, werkstatt, service, unfall, vw, audi lkw, reifen

These clusters can then be combined in a second step by determining how similar they are to each other (see Part A, Term Spaces). It can be seen from the examples above that quite different grades of relatedness are extracted via this method.

A very simple way of extracting synonyms by their distributional properties is to look for patterns in a query log of the type A _ B and list those instantiations of this pattern that occurred with many pairs of A and B. The resulting list of these patterns on YO yields the following top counting synonym candidates:

107 civic accord
66 truck pickup
59 ranger explorer
57 prelude civic
53 ranger mustang
53 prelude accord
53 olympics games
51 truck mustang
49 thunderbird mustang
49 mustang explorer
47 full free
45 olympic games
42 rock music
41 truck ranger
41 truck explorer
39 thunderbird ranger
38 taurus ranger
38 probe mustang
37 ranger probe
37 nude naked
36 taurus mustang
36 mustang bronco
36 free crack

This pattern extracts mostly different models of the same brand (Ford Mustang – Ford Escort). This is indeed a fundamental problem of any unsupervised approach that builds upon either syntagmatic or paradigmatic patterns: How to discern terms that just behave very similarly from genuine similar terms? Different models are just one example of terms that behave very similar, but should not be mixed up with each other. Many co-hypnomys share the same distribution, even though they are distinctly different, for example *Herren – Damen* (consider all the garments that can occur with both sexes).

A final illustration of what is possible through association metrics is provided by the following pairs were extracted through looking at the top co-occurence measure based on Dice with Yates' correction. The co-occurrences were observed on MKW-DE. The first number is the number of co-occurrences, the number in parenthesis are the overall frequencies for the terms. Related terms are marked with a tilde, spurious entries with ?, while generality is marked with < and >:

```
211 hotel (14624)   reiseveranstalter (503)
1559 hotel (14624) ? restaurant (3027)
348 hotel (14624) ? golf (1677)
192 hotel (14624) ? gaststtte (556)
128 hotel (14624) > gstezimmer (252)
202 hotel (14624)   busreisen (652)
220 hotel (14624) ? restaurants (894)
682 hotel (14624)   sauna (9344)
218 hotel (14624) ? ski (1016)
1697 hotel (14624)   reisen (6723)
68 hotel (14624) < hotellerie (108)
393 hotel (14624) ? veranstaltungen (3632)
1045 hotel (14624)   reise (3360)
368 hotel (14624)   tagungen (668)
607 hotel (14624) ? gastronomie (1983)
687 hotel (14624)   wellness (2789)
403 hotel (14624) < touristik (1062)
677 hotel (14624)   essen (3543)
338 hotel (14624)   last minute (911)
757 hotel (14624)   freizeit (4695)
359 hotel (14624)   ferienhäuser (1073)
```

10.3. The two basic relations of synonymy and generality

In light of the the trade-off between coverage and precision in gathering instances of lexical semantic relations, the breadth of various relations can be cut down to two basic relations, synonymy and generality.

10.3.1. Synonyms and Quasi-Synonyms

Synonyms have to be examined from the perspective of active and passive language capabilities. Very seldom a total balance between the two synonyms, much more often one is the preferred synonym and the other term a lesser accepted or used term.

Strict synonyms have a way of making themselves rare. There are only few examples which are rather resistant to the nagging doubts of connotational differences. Almost all variants can - if looked closer upon - be traced down to stylistic, regional, temporal or domain-specific variants. If one does eliminate also translations and abbreviations, the bag becomes almost empty. In fact, native and Latin or English variants form, together with abbrevations build a considerable part of what can be empirically validated as synoynms.

Strict synonyms, defined by their universal substitutability, are a hard find. As different lexemes tend to evolve semantic nuances, there is seldom complete agreement on synonyms. Some variants gets connotation of register and are no longer perceived as neutral terms As the frequency of synonymic variants will often differ in magnitudes, the less common term is likely to be viewed with suspicion, even if its meaning perfectly matches its more common partner (consider Pädiater - Kinderarzt). Many synonyms are only valid in a limited set of distributions. For example, *einrichtung* and *möbel* works in the context of *Einrichtungshaus* and *Möbelhaus*, but not with *Klinische Einrichtungen* vs. *Klinische Möbel.* Reduced forms as synonyms were already discussed above (see Morphological-Syntactial matches). Additional examples for reduction forms can be discerned by applying the list of orthogonal terms as these often do only play the role of a semantic empty suffix. *Tapezierservice* is the same as *Tapezieren* and *Tabakwaren* does not differ decisively from *Tabak*[15].

A similar case to synonyms are variations on named entities. Apart from different spell-out conventions (for example, writing person names with a middle initial or not), a case of genuine different expressions referring to the same entity are exonyms. Place names from different languages for the same place could prove to be helpful for multilingual retrieval (for examples, *Monaco - Munich - München*).

10.3.2. Generality

Generality is a very intuitive concept. It can be observed through different paraphrases, including *X is a kind of Y.* A table is a kind of furniture; clearly, the opposite does not hold — by this token, furniture is more general than table. A *Samsung X-05* is a specific notebook, it could also be said to be a specific electronic device.

Generality relation is not commutative. It is not generally transitive, as it is often restricted to a local domain. For example, a *table* can be said to be a kind of *dining room furniture* and a working table can be said to be a table, yet a *working table* is certainly not a kind of dining room furniture.

Admittedly, such examples stretch what *kind-of* paraphrases usually hold. It can nevertheless not be hoped to establish a transitive relation of generality spanning several domains by applying one common metrics to all pairs.

The obvious morphological evidence that if a term A is a part of a compound term B, it is likely to be more general than B is a baseline approach. Note that through some orthogonal right segments of compounds counter this heuristic, for example *Hotelgewerbe* which is certainly not less general than *Hotel.* As a slightly refined approach, generality can also be observed through co-occurrence patterns.

While in many cases, the heuristic “if A appears in more lines without B than B does without A” hints for a generality relation between pairs of often co-occurring A and Bs, it does not always yield the desired effect. The pair *auto* – *cabrio* could be recognized by the heuristic, given that *auto* appears in 9.838 out of its 10.417 total lines without *cabrio*, but *cabrio* only in 401 out of its 980 lines without *auto.* More

[15] The *-waren* indicates a slightly more general term, however.

than half of the *cabrio*-lines appear with *auto*, but only 5.5% of the *auto*-lines with *cabrio*.

However, it does not work always that smoothly, as can be seen from the pair *touristik – hotel*. Although *touristik* is clearly the more general concept with regards to *hotel*, this relation cannot be derived from frequency counts. In the MKW-DE set, the raw number of occurrences for the contingency table are:

count for hotel total 16.898
count for hotel and touristik 1.082
count for touristik total 2.526
count of all lines 551.707

Regardless of what similarity measure is applied (see above), there is no way to discriminate this pair from these two other pairs that are clearly not exhibiting a generality-relation (hotel – meer) or in which the more frequently appearing item is really the more general one (webdesign – logo), given that ultimately all similarity-measures take their features from the contingency table:

count for webdesign total 18.371
count for webdesign AND logo 1.149
count for logo total 2.725

count for hotel total 16.898
count for hotel AND meer 1.018
count for meer total 2.757

The numbers are so close to each other (less than 10% deviance from the original *hotel – touristik* pair) While both pairs have items that are somehow related, the Without keeping track of lexical units, it cannot be hoped to achieve a list of meaningful similar and/or granular terms dynamically, even through very rich resources full of TE-Commerce-relevant vocabulary.

11. Enriching lexical resources

In addition to keeping track of vocabulary, a term handling framework for TE-Commerce also has to provide a process for adding vocabulary. In this chapter, a mechanism is presented that build upon existing lexical resources. It will be demonstrated how a large existing vocabulary repository facilitates additions and allows to capture substantial parts of external term collections such as query logs or Web meta keywords.

The proposed enrichment approach centers on the notion of Sense-Morphemes. Starting from a set of known Sense-Morphemes for a given topic, candidates for new topically related terms are gathered and through those, new Sense-Morpheme candidates. Through this bootstrapping process, both the original vocabulary repository as well as the candidate resources can contribute to the enrichment.

Given that this approach is performed on very large "real-life" resources (meta keywords, query logs, business directories, Web corpora), issues of noisiness will have to be considered when automatically enriching lexical resources.

11.1. MetaMatch and Sense-Morphemes

Based on the framework of Sense-Morphemes, an enrichment process (MetaMatch) for lexical resources is presented in this section. In this process, additional lexical material is acquired through applying already gathered Sense-Morphemes and lexical units.

After an initial process of determining a Term Space related through co-occurrence to the lexical units already known, a filtering process using the Sense-Morphemes and Term Space filtering routines introduced above is started. Finally, not only are new lexical units extracted, but also additional Sense-Morphemes. This allows a continuous semi-supervised enrichment process scalable to very large corpora and flexible enough to extract valid lexical units from corpora of different quality (see Part A, "Term Spaces" for issues of a Term Space quality).

11.1.1. Applying Sense-Morphemes

As was laid out earlier in the chapter on morphological-syntactical variation, the two main properties of a Sense-Morpheme is robustness in meaning and variability in expression. The latter property requires a new sort of filtering routine capable of catching morphologic variants. Just like the grep tool that allows to filter lines by regular expressions, a morpheme-grep needs to be established that only prints line containing a morpheme or one if its variant expressions. For example, applying this morpheme-grep with the Sense-Morpheme *[musik]* should list terms such as musikalisch

musikalien

musisch

musikexperte

Note that this differs from a substring (wildcard) search in two ways: Firstly, it only captures morphologically motivated segmentations — not mixing up ”barcode” with ”bar” —, secondly it goes beyond substrings in capturing variants (musik is no substring of musisch). Using the affixation tables described above, it is not needed to list ”wechseln” and ”wechsler” as separate entries. These kind of variations can be tackled by recognizing allomorphy.

Using Sense-Morphemes on a list of segmented compounds from a TE-Commerce (query-log, Web term list or meta keyword repository) is a way to quickly enrich a lexical resource. Following from the right-hand-rule, choosing those compounds where the known Sense-Morpheme appears on the right hand side is in general a safe approach. However, Sense-Morphemes such as *schneider* that are ambiguous require additional consideration.

One remedy is to apply the Left-Right-algorithm described below (Part C, “Head / Container principle for search queries”) to detect additional Sense-Morphemes for each meaning of the Sense-Morpheme. In the case of *schneider* such additional disambiguating morphemes could be *kleidung* and *messer*. Then, only these compounds are added to a list representing one meaning of the Sense-Morpheme that occur frequently with one of the additional disambiguating morphemes. For example, *Apfelschneider* is then removed from the list of *schneider [textil]*-extensions, because *Apfelmesser* appears frequently compared to *Apfelkleidung*.

11.1.2. Detecting new Sense-Morphemes

Conducting a first order co-occurrence on the representative seed Sense-Morphemes of a category helps to set up a first list of candidates. For the example here, all meta keyword lines containing both *sanitär* and *klempner* have been extracted.

The following Sense-Morphemes can be extracted from this co-occurrence list through applying the CoreMaker (see above, Part A, “TE-Commerce Term Spaces”) algorithm to it and manually pruning erroneous entries:

sanitär
klempner
heiz (383)
bad (264)
installateur (59) , including installation, installieren
lüftung (43)
haustechnik (29)
armatur (24)

klima (20)
dusche (10)

The filtered list only retains those lines that contain a known Sense-Morpheme of the category. It looks as follows:

787 sanitär
748 heizung
518 klempner
450 bad
288 klempnerei
216 installation
208 bäder
207 haustechnik
199 klima
193 lüftung
146 bauklempnerei
139 gasheizung
139 dusche
130 heizungsbau
127 installateur
115 armaturen
106 badezimmer
105 fußbodenheizung
104 sanitärtechnik
102 heizkörper
98 heizkessel
98 sanitärinstallation
85 badewanne
81 badsanierung
79 badplanung
77 duschen
69 badmöbel
68 ölheizung
58 klimatechnik
56 heizungstechnik
56 heizungsanlagen
56 badewannen
48 klempnerarbeiten
46 installationen
43 badausstellung
43 heizungen
43 schwimmbad
40 wasserinstallation

38 schwimmbadtechnik
37 gasinstallation
36 sanitäranlagen
35 badrenovierung
35 heizungsanlage

This list is a virtually clean list of the top representatives for the category *Sanitär*. Further refinements can be achieved by adding more Sense-Morphemes.

11.2. MetaMatch as a generative process

As a generative process, MetaMatch allows to fold out all Sense-Morphemes and orthogonal terms to their different spell-out. These combinations can then be filtered by a TE-commerce term collection such as a query log or a meta keyword repository. In practice, the generative process equals to apply MetaMatch in its usual normalizing function to all elements in the term collection.

Some example of the generative process on MKW-DE are presented below, starting with terms that are folded through their orthogonal components (note that singular/plural variations are not listed):

buchprüfungsbetrieb – buchprüfungsunternehmen – buchprüngsfirmen

chefkenntnis – cheftraining – chefseminar – chefunterricht chefkurse – cheftagung

tankstellenfirma – tankstellenbetrieb – tankstellenunternehmen – tankstellenabteilung

meetingmanager – meetingchef – meetingdirektor

bioethanolgewinnung – bioethanolproduktion

unfallabteilung – unfallwerk – unfallzentrale – unfallcenter unfallbetrieb – unfallamt – unfallzentren – unfallanlage

Combining a Sense-Morpheme with orthogonal facets, the term *Abfallberatung* would be represented as *498-BERATUNG* with 498 indicating a Sense-Morpheme that can be unfolded to *abfall/müll/reststoff/wertstoff* and BERATUNG indicating the orthogonal facet that can be unfolded to *berater/beratung/consultant/consulting/beratungsdienste / beratungsdienstleistung*. Again, all combinations that do not occur in MKW-DE have been filtered:

abfallberater
abfallberatung
abfall-consulting
müllberater
müllberatung

reststoffberater
reststoffberatung
reststoff-consulting
wertstoffberater
wertstoffberatung

The normalized form 498-761 contains two Sense-Morphemes, besides from 498 Abfall (see above) also 761 that expands to *eimer/kübel/behälter*. Again, only those combinations are reported that were found in MKW-DE:

abfalleimer
abfallbehälter
abfallkübel
mülleimer
müllbehälter
müllkübel
reststoffbehälter
wertstoffbehälter

Note that ambiguous Sense-Morphemes can only be filtered through negative lists or checking for co-occurrence of the resulting forms[1]. For example, the spurious pair *pflaumenbrand – pfaumenfeuer* is due to the ambiguous head *brand* with "alcoholic spirit" vs. "plant disease" as its two meanings.

11.3. Filtering processes

Cleaning Term Spaces implies the detection of artifacts and spurious terms. In general, all terms that are originated in how the Term Space was created (the "overhead" of the Term Space), but do not at the same time reflect content properties should be cleansed before the Term Space is inspected. For example, one often encounters artifacts in query logs that are due to internal testing routines or the query monitoring process, but do no tell anything about users.

Outliers can be detected by intersection, especially those that occur because of the origin of the term spaces, for example natural language queries that will hardly occur anywhere else. Pruning the peculiarities of a Term Space requires a filter that is larger than the original term collection. This is a corollary following from Zipf's distribution, given that usually more than half of all term types occur only one time in one term collection.

[1] However, not all valid synonyms will co-occur even in a large meta keyword repository. A negative list would contain all larger units that must not be conflated through folding out Sense-Morphemes. This list could be set up by editing and auditing the most frequently occurring results of MetaMatch.

11.3.1. Top query log frequency

Terms that appear with very high frequency in general Web search deserve special attention. The top frequent terms are often short, ambiguous or offensive terms. Inspecting the intersection of the (for example 10.000) most frequent terms with results of the enrichment processes usually yields many problematic terms.

In addition, it ensures that exactly these terms are inspected manually and become cleaned that are queried the highest number of times and would thus cause the highest detrimental effect.

Another application of the spreading notion is to remove overly generic terms by means of general language data. Removing the most frequent terms from general corpora — such as news-wire — allows to detect terms that probably contribute only little to a TE-Commerce meaning.

11.3.2. Spamming

If using resources that were created by many different people without a central editing instance, it has to be made sure that no malicious intentions in term usages interfere with the extractions. One prominent type of such malicious intents in the usage of terms are spam websites that contain a wide range of highly frequent vocabulary without any relevant content to it. Especially in meta keyword lines one may encounter illegitimate use of vocabulary, because of search engine ranking manipulating. In all cases, the vocabulary presented is only intended to attract traffic.

11.3.3. Removing noise in MKW lines

As discussed above, meta keywords are a very rich resource for lexical enrichment, albeit their usual infiltration with spam terms. Conducting a cleansing operation on a meta keyword table can be done by following the steps below:

- Remove all known junk terms and lines containing known junk terms (a line that contains a known junk term is discarded, as the other terms might be problematic too — i.e. contagiousness of junk)

- Remove all lines that do not obey standard syntax (comma or semi-colon separated values) — another syntax of MKW commonly deployed use spaces instead of commas to delimit keywords. In this case, terms consisting of more than one word are not longer distinguishable

- Remove all lines with too many MKWs per line (for example, more than 50)

- Remove all lines with all weird MKWs (no MKW appears elsewhere or meets a frequency threshold)

11.3.4. Spreading

If a term appears frequently in too many categories, it can be largely ruled out that it is a good representative of any category. This method of reducing noise through discovering high spreading terms resembles the inverse document frequency metrics in Information Retrieval[2]. It is based on associations between terms and categories, for example provided by a database of websites belonging to businesses for which the category is already known.

One way to measure spreading is to use a threshold for the number of occurrences in a category and then count the number of categories in which the term's occurrences surpassed this threshold.

A test conducted on Web-extracted keywords for which the SIC category of the website was known, yielded the following terms as having the highest spreading value (the number indicates the number of categories

542	email
542	click here
529	united kingdom
525	rights reserved
522	customer
521	products
520	email address
516	further information
507	privacy policy
506	high quality
504	head office
501	telephone number

All of these terms are clearly not characteristic of what a company does, but rather reflect what companies typically write on websites.

On the other hand, some bona fide terms appear in numerous categories, but are nevertheless applicable for a restricted range of categories. For example, *credit card*, *consultants* or *website hosting.* For some categories, such as banking, consultancy or Internet Services, these are valid keywords.

[2] See [Baeza-Yates/Ribeiro-Neto 1999].

Part C.

TE-Commerce - the case of queries

12. Query Types and how to recognize them

In a first attempt to unveil what is inside queries and how to treat them adequately, the complete query space is divided into several types of queries.

It has been pointed out that one of the most important shortcomings of traditional search engines is their equal treatment of every query. Looking at very frequent user queries the differences between queries become instantly apparent. In this section, the process of systematizing such differences is initiated in order to give rise to a general query classification. This classification will be based on a utmost linguistic notion, namely the simple sentence and its underlying predication. Subsequently, this division will be narrowed down to E-Commerce relevant queries.

12.1. Taxonomies of Web queries

The taxonomies of Web queries presented below are partitions of the Web query space into broad groups based on users' intents.

User queries come in different shapes and are issued in order to achieve different aims. This calls forth different strategies to deal with queries. This can easily be seen when taking into consideration the different types of results that can answer a query (see Part A, Interactions). Some queries are not optimally answered with a list of links and snippets from the destination pages. If a user searches for *homepage fc bayern muenchen*, there is in principle no need for an additional screen before sending her to the official homepage. This is not to say that the user interface needs to be behave totally different for different query types — having a homepage locator module in the crawler could nevertheless bring additional benefit, for example by highlighting the official homepage in the result set.

In this section, the standard taxonomies of Web queries as discussed by the seminal papers of [Rose/Levinson 2004] and [Broder 2002] are introduced and discussed. In combination with these classed of user queries, based on the users' intentions, the discrimination of queries with a local vs. a global scope are discussed, following largely the works of [Gravano/Hatzivassiloglou/Lichtenstein 2003].

Introducing the notion of predicates and arguments — in the context of elementary sentences (Harris and Gross) — leads to a new matrix of Web queries. This matrix allows not only to reconcile Broder's and Rose/Levinson's taxonomies, but also makes clear why uniform approaches to query answering are inherently limited. Based on this matrix, it can be argued why differentiations on the side of query processing, crawling/indexing and result presentation are necessary to resolve different kinds of queries.

12.1.1. User goals in queries

In contrast to classic IR that is based on the model of information need of users, experiences with Web search engines taught that information alone is not sufficient to describe user goals. In the context of Web search, users could also look for a specific URL (navigational queries) or perform transactions via the Web such as buying or downloading (transactional queries)[1].

Along with conducting an online survey of AltaVista users, Broder examined query log data (a random of 1000 query types from one day of AltaVista searches) to corroborate his thesis that only a minority of user queries are informational queries[2]. Some of the example queries provided by Broder deserve special discussion.

For navigational queries, Broder lists the following five examples and probable targets for each of them:

- Greyhound Bus – http://www.greyhound.com
- compaq – http://www.compaq.com
- national car rental – http://www.nationalcar.com
- american airlines home – http://www.aa.com
- Don Knuth – http://www-cs-faculty.stanford.edu/ knuth/

The classification of "Don Knuth" and "compaq" seems at least disputable. It is conceivable that a webpage containing a product overview, a price comparison or reviews are an equally good answer to *compaq* than Compaq's homepage is. For *Don Knuth*, a link to his works or encyclopedic information on him could also be in the set of relevant results, next to his homepage. The only clear case of a navigational query is *american airlines home* as the sought for a homepage is made explicit in the query. The most frequent examples of such an explicit marking of a navigational search are URL queries.

Based on what kind of interactions follow from the search, two other types of queries are discriminated. The user goal behind informational queries can be fulfilled through reading – no further interactions are required. Obviously, informational queries span a broad range of topics and granularity. An informational query can express a very specific need — such as *height of mount everest* — but also general ones, for example with a query *soccer.*

The more detailed taxonomy of user goals in Web searches by [Rose/Levinson 2004] divides the broad sectors into finer classes. Apart from a finer subdivision, it differs

[1] Broder 2002

[2] Both methods yielded comparable results. See Part A, "TE-Commerce Interactions", for a discussion on different empiric approaches to user studies.

from Broder's taxonomy in that it introduces a class of resource queries that replace transactional queries.

In a similar way than Broder's examples, the instances of navigational queries brought up by Rose/Levinson are only on the first glance straightforward. The example query *duke university hospital* is classified as navigational query. It may be safely assumed that retrieving the homepage of Duke's university hospital is a good way to answer this query. If taking a look at the content of the query, it can be seen that it contains a named entity. What makes this query different from a query such as *München*? The latter also contains a named entity and there is also an authoritative site for it. This illustrates how subtle the differences between these types of queries are if looking at queries in detail.

The distinction between "advice" and "undirected" types of informational queries is also marked by a very thin line. Apparently, "advice"-type queries are supposed to contain some call-for-action element that lacks in "undirected" ones, which can be paraphrased as "tell me all about X". However, as the examples (*walking with weights* for "advice" and *color blindness* for "undirected") reveal the "undirected" type might be justly interpreted as looking for an advice (looking for a remedy or relief to *color blindless*).

12.1.2. Arguments and predicates in queries

If a subset of queries is to be understood as having a propositional value, it s possible to apply the notion of arguments and predicates to them.

This opens up the possibility to understand queries as skeletons of elementary sentences in the tradition of Zellig Harris (kernel sentences that undergo transformations to build other simple and complex sentences) and Maurice Gross.

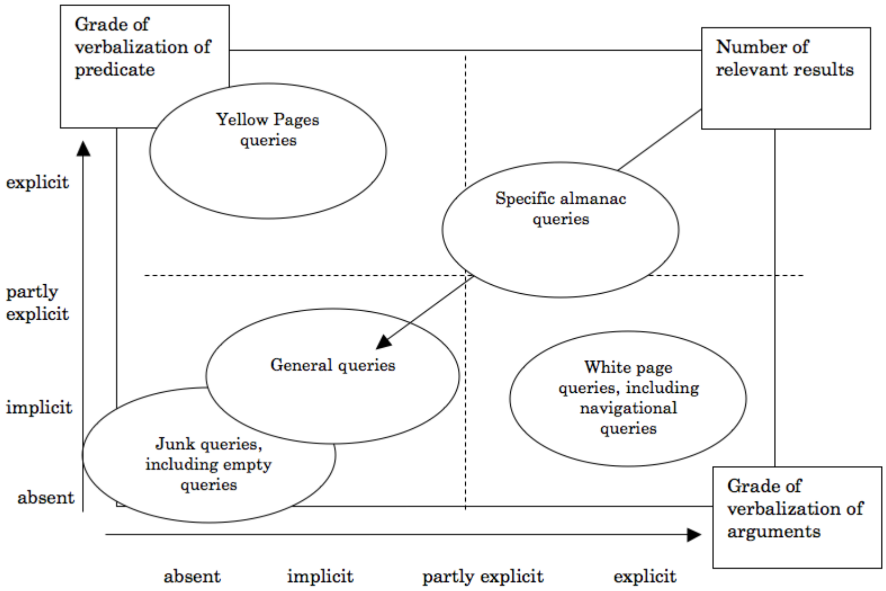
Combining the taxonomies of Web queries based on user intentions with the propositional content of queries that is expressed in a predicate-argument structure leads to a new matrix of Web queries. The main advantage of this matrix is that it not only shows why different query types have to be processed differently, but also why different result indexes have to be prepared.

The two axes designate the grade of verbalization of the predicate and arguments in the queries, spanning from absent to implicit and explicit verbalization. The more explicit the query predicates and arguments are formulated the smaller is the number of relevant results for these queries. Adding more intensional restrictions to the set of results leads to a smaller extension.

A special case are those queries for which no propositional content can be recognized. If these are disqualified as "junk queries" it means not more than that they are unresolvable by operational methods, including finding related queries (see below). For many examples of junk queries — such as numerous repetitions of the same letter — one cannot think of any way how to meaningfully resolve them.

Queries with explicit predicates and no arguments are in general Yellow Pages queries and can be best answered by an index of businesses' records. The predicate might appear in different morphologic forms, for example as a nomen agentis (*carpenter Leeds*) or an abstractum (*carpentry Leeds*). A prototypical Yellow Pages query also contain a geographical modifier. Without geographical modifiers it is often not easy to distinguish between general queries and Yellow pages queries. For example, *jewelry* could be interpreted as a general, encyclopedic search, but also as a product search. Such product or service queries can be subsumed under a broader concept of Yellow Pages queries, given that they are also ideally answered with a link to a vendor or manufacturer.

White Pages queries, on the contrary, only contain arguments. Their predicate is always implicit and one of the kind "what are the contact details for X". As with Yellow Pages queries, it is sometimes not easy to decide where to draw the line between general queries and White Pages queries.


Grade of verbalization of predicate
Number of relevant results
Yellow Pages queries
explicit
Specific almanac queries
partly explicit
General queries
White page queries, including navigational queries
implicit
Junk queries, including empty queries
absent
Grade of verbalization of arguments
absent
implicit
partly explicit
explicit

A matrix of Web queries

This matrix places on one axis the grade of verbalization of the predicate (from fully absent to implicit to explicit) and on the other axis the grade of verbalization of the arguments. Different query types are defined through their positioning in this grid. With regards to the ideal number of results, the more verbalized arguments and predicates are, the smaller becomes the set of relevant results. This is indicated by the diagonal arrow.

12.2. Web Navigational queries

In the query taxonomy evolved here, navigational queries are restricted to URL retrieval queries. They only apply when a user knows a website because it was viewed by her at one time, but has not memorized the URL or bookmarked the page — or simply chooses to type into the query box for sake of speed and convenience[3]. In general, two different approaches are feasible: either entering a part of the URL (often introduced by www), or a text chunk on the page.

Web navigational queries can serve as a substitute to more than bookmarks, as they allow to track content over different locations. In addition, when a piece of content is taken off the Web by one site, it might still be retrievable at other locations.

12.3. White Pages queries

In White Page queries, the user has a previous knowledge of the name of an entity (precise or imprecise) already obtained. The notion of White Pages does not need to be restricted to person names, but can be extended to other types of entities, such as web sites, companies or specific products. Though it is not always clear what informational needs stands behind a White Pages query (it could be something else than preparing a contact), yet it is clear that only information attached to the specific entity is helpful.

The term White Pages query should in the following be understood in analogy to the White-Page telephone directories. How can a term for a special sort of telephone directory be applicated to the model of internet search? Naturally, redirecting web queries from special sites to a machine readable telephone directory - as the web interfaces of network providers (`www.telefonbuch.de`) present - does only provide a more comfortable access to the data of telephone books. It can not to do justice to queries as entered on a general search engine.

White Pages query consist solely of arguments. They have an implicit predicate, which can often be paraphrases as "I want to know about X" or "How can I contact X".

12.4. Yellow pages queries

The logic of YP-queries is constituted by a need for which the user does not know beforehand what exact entity may provide it. Typical YP-queries do not only contain a formulation of the need, but also a geographical element. The reason for that is naturally that users submitting a YP query usually need to deal personally with the business.

Syntactically, YP-queries therefore consist of predicates and optional supplements which are mostly of a local type, but may also include specialties such as opening hours or price (for example, *cheap 24h restaurant london*). The YP predicates may appear as occupational titles, vendors, branch categories or problems (such as *acne*). From the perspective of TE-Commerce, *cosmetics*, *cosmetician*, *acne*, *beauty shop* etc. all share one common underlying predicate. The different realizations of this predicate can be captured by virtue of the TE-Commerce lexical functions described above (see Part B.).

[3] The "I'm feeling lucky" button on Google.com serves exactly this need.

Usually, YP sites offer three fields that can be searched — although these three fields can be realized in only two or even one search slot. These fields, namely “who” / “what” / “where” field are connected by an AND-operator. Only those searches that specify the “what”-field are proper YP-queries. From these perspective, the content of the “who”-field acts as an additional specification or a filter on the results.

With real-life users, some degree of mix-up between the different fields inevitably occur. One common behavior is to read the first field as surname and the second field (the “what”-field) as first name. Another typical issue is to insert the type of White Pages-information requested in the “what”-field, for example *mobile phone number.*

12.5. Almanac queries

A substantial part of web queries aim at encyclopedic information in a broad sense. Among such encyclopedic information are birth dates of persons, capital of states, city population or the name of inventors. Such searches are ideally not answered with a listing of URLs that require the user to find out whether the information is really available at these webpages. Instead, they should produce precise answers, ideally with references and other indications of their reliability.

Current Search Engines such as `Google` or `Yahoo` have introduced such mechanisms that trigger a special flash-in for a limited number of queries. For example, a query such as *president Haiti* or *define abrasives* produce a separate flash-in on `Google` that is based on a Web extraction module specially tailored for such classes of queries.

A demonstration of this principle can be provided by the class of queries asking for the etymology of a term. Here, the existence of an entity is known, but not the origin of its name. Queries such as *warum heisst X X* or *warum heisst C so?* might be well answered if retrieving the proper parts of Wikipedia entries. The following table shows prominent etymological queries in a YG query log and the corresponding Wikipedia entries that answer the questions put forth:

warum heißt dollar dollar???

“Das Wort leitet sich aus der alten deutschsprachigen Mnzbezeichnung Taler [...] ab”

warum heißt der teufelsmoor teufelsmoor

“Der Name Teufelsmoor leitet sich von doofes Moor (taubes Moor) ab.”

warum heißt das schwarze meer schwarzes meer

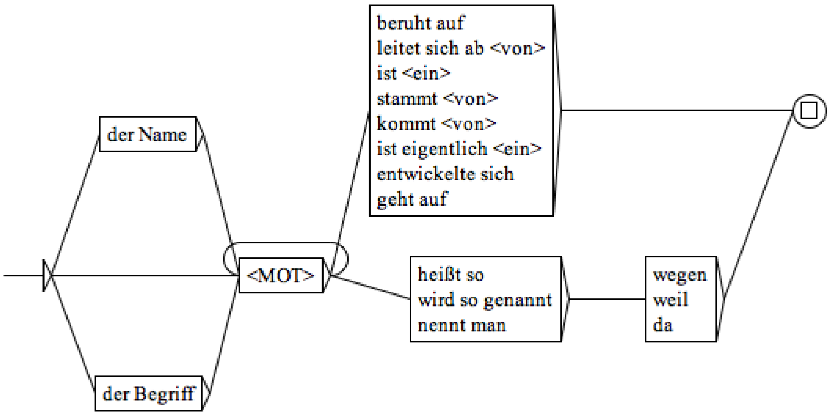
“Eine verbreitete Deutung zur Erklärung der Farbe des Meeres geht auf das Vorkommen besonders vieler sulfidogener Bakterien zurck, die durch ihre Sulfat reduzierende Aktivität Eisensulfid ausscheiden, welches das Wasser schwarz färbt”

warum heisst marienkäfer marienkäfer

“Wegen ihrer Nützlichkeit fr die Landwirtschaft glaubten die Bauern, dass die Käfer ein Geschenk der Maria (Mutter Jesu) seien warum heißt der kaiserschnitt kaiserschnitt

"Sectio caesarea, aus dem Lateinischen, zu deutsch "kaiserlicher (caesarea) Schnitt (sectio)"

All of the presented explanations can be extracted from the corresponding Wikipedia articles via the following local grammar



This automaton extracts etymological explanations that occur on the right hand side of the recognized units and the element for which the explanation is sought at the *¡MOT¿* slot.

A further example centered on the word *Dampfmaschine [steam engine]* illustrates the variability of encyclopedic facets. Below, such facets will be formalized under the notion of query containers. If looking at the list of queries containing *Dampfmaschine* and removing those queries that look for model steam engines (a clear case of TE-Commerce queries), one ends up with the following results from YG:

Dampfmaschine Bauplan
dampfmaschine funktion
Funktionsweise Dampfmaschine
Dampfmaschine Funktion
Dampfmaschinensteuerung
Dampfmaschine funktion
Dampfmaschine James Watt
erste Dampfmaschine der Welt
erfindung dampfmaschine
Dampfmaschine von James Watt
Dampfmaschine bild James watt
Dampfmaschine Schemazeichnung
erste dampfmaschine
dampfmaschinenbaupläne
dampfmaschine bauplan
James Watt Dampfmaschine Schemazeichnung
Erfindung der Dampfmaschine

Erfindung Dampfmaschine
dampfmaschine Zeichnung

Even this short selection reveals a high amount of variability in the patterns that contain the type of encyclopedic information sought for (*X Zeichnung — Schemazeichnung X — Bauplan X* or *Erfindung der X — Erfindung X — erste X*. Apparently, structuring these patterns will not only reduce the number of queries one has to deal with, but also allow to extract other elements for which the same type of encyclopedic information is sought for. For example, other inventions for which schematic diagrams are sought for in YG are the following:

Feuerlöscher
Heizung
Gaskraftwerk
Schalldämpfer
Subwoofer
Nistkasten
Armbrust

Through detecting and normalizing almanac facets, it is not only possible to reduce the amount of labour needed for setting up specially tailored extraction devices — ideally build as local grammars[4] —, but also to detect more entities that often appear in connection with almanac searches. One application of this is for example to present a special drill-down for searches containing this kind of entities, such as birth date, mini biography or occupation for a (prominent) person.

12.6. On queries with a local scope

Local queries have gained attention from different viewpoints in recent years. From the perspective of TE-Commerce, it is apparent that many local queries have an E-Commerce value to it. Moreover, finding the best answers to those queries requires local knowledge, an asset typically ascribed to YP providers.

The issue of local scope within a query is related to questions of the geographical scope of websites. This scope can be determined by extracting geographical names in webpage content[5]. Such an extraction may be extended to examine the geographical origin of other websites pointing to a website[6]. Through databases of IP numbers and their corresponding region, the location of a webserver can be approximated. However, the physical location of a webserver does not always tell much about the geographical scope of the pages it hosts.

[4] For the advantages in using local grammars for such purposes, see [Gross 1997] and [Gross 1999b].

[5] This requires an disambiguation module, given that many place names appear in several countries. See [Leidner 2004].

[6] See [Ding / Gravano / Shivakumar 2000].

If taking a look at user queries, even on first glance differences in the local scope become visible[7]. On a most basic level, the query space is divided into two subspaces, one with queries containing a geographical reference and one with queries that do not. However, this does not necessarily tell much about the local scope of a query. Consider for example a query such as *Easter Island history* contains a geographical reference, but it is obviously not localized in a way that *houses for sale Munich* is. Only in the latter case, the geographical reference translates into a range of postal addresses. A website containing information on the history of the Easter Island does not need to state anything about streets, towns or postal codes on Easter Island. In contrast, a user looking for houses in Munich expects to be provided with addresses (even if these addresses do not appear on the webpage itself, but are provided afterwards).

Following this stance, a query such as *flowers* is just as ambiguous as a query *new york pizza*. While the latter case is a lexical ambiguity (the city New York vs. New York as a pizza style), the former one can be interpreted as either looking for vendors or flowers — i.e. contact addresses of vendors — or looking for general information on flowers. Interestingly enough, if substituting the query *flower* with *wildflowers*, the localized reading is largely ruled out, because wildflowers are usually nothing that can be bought in stores. This again underlines the connection between TE-Commerce queries and local queries. Of course, non-localized (or "global") queries can also have a TE-Commerce value. This becomes immediately apparent when looking at queries which have a transactional intent that can be fulfilled online, for example *pdf reader download*.

12.6.1. Extracting German local queries

A first test to estimate the set of terms that tend to occur with geographical identifiers was conducted by extracting queries which contain German cities, pruning the city names and then counting how many different cities occured with the rest of the phrase. In this fashion, the query *hotel hamburg* was first reduced to *hotel*. Afterwards, *hotel* was ranked according to the number of different cities with which it appeared, regardless if the phrase had the city name on the left-hand or right-hand side of *hotel*. Note that the query frequency is thus ignored, only the number of different types are taken into account. This minimizes the effect of confusioning frequent terms that only appear with one city (*berlin alexanderplatz*). We are interested in terms that spread over a large number of city names, not in terms that occur frequently with single placenames.

This is the top section of the list that resulted:

7335	immobilien
5680	stadtplan
4640	hotel
3950	stadt
3248	kino

[7] A more detailed overview is provided in [Gravano/Hatzivassiloglou/Lichtenstein 2003].

2268	sparkasse
2200	arbeitsamt
2133	job
2096	volksbank
2059	gemeinde
2041	feuerwehr
2021	krankenhaus
2014	hotels
1868	gymnasium
1629	zeitung
1604	tierheim
1566	vhs
1548	finanzamt
1465	ferienwohnung
1407	volkshochschule
1363	amtsgericht
1292	wohnung
1292	stadtwerke
1189	autohaus

All these terms occur significantly often with different city names. The reasons for that, however, differ quite remarkable. Some terms represent instantiations of organizations that appear in most cities, for example *amtsgericht*, *stadtwerke*, *vhs / volkshochschule* (synonymous terms). Each of the individual local instances is a business or agency in its own rights, they do not form a unity. Rather, they stand for organizations that typically are present in a German city. Note that usually, they are not referred by through a proper name but by the combination of organizational name and city name, e.g. *volkshochschule dachau.* Other terms reflect common needs for a specific location, such as ‘wohnung” or “job”. The last class follows the logic of Yellow Pages in that it contains business types (for example “autohaus”) that are requested for a specific locality.

An additional division can be made based on the relationship of the user posing this query and the location entered. Some terms that appeared above, for example “ferienwohnung”, suggest that the user is not a resident of the location. In other cases, the user could be moving to the city. A query of the list above that could potentially show a reflex of arriving residents is “wohnung” or “stadtplan”[8]

12.6.2. Local knowledge

Local knowledge is regarded as the key asset of YP providers, meaning that they know the businesses, needs and peculiarities of a place. This notion can also be extended to the Web, even though local search on the Web usually means searching all places through one common interface.

[8] Of course, also residents might look for a new apartment or could be in need of a city map.

From an analysis of 48.029 query occurrences (1.000 query types) of YG that contained *Munich* it followed that more than 60% of these are looking for an entity which has one authoritative URL. While it can be argued that additional results could also be valuable to the user (leading to other sites not thought of when formulating the search), s search agents should at least mark this authoritative URL separately.

Such entities comprise for example *Flughafen München*, *Messe München, Arbeitsamt münchen* or *Körperwelten München.* All of these searches have one canonic URL that should be sufficient for the search, and at least needs to be included and highlighted in the result set.

The large part of the remaining searches (30% of the 48.029 occurrences) can be classified as being Yellow Pages searches. For these types of queries it is crucial to list bona fide business homepages in the result set.

These findings underline the importance of a special local-component of generic Web search engines. Following this line of thought, local knowledge can be interpreted accordingly as knowing about the websites from a locality and their local-relevant content such as addresses or opening hours.

13. Head/Container principle for search queries

Assuming as a prerequisite that a normalized term representation of queries is constructed in a first pass (including basic spell-out normalization operations, MWU detection and term variant recognition as described above), the repetitive properties of a query log are examined in a greater detail in this chapter. These properties will be described under the head/container framework which is derived from observing how shorter queries are part of larger queries, but convey the main intent of the query. The head/Container framework thus bridges the

13.1. Repetitive patterns in Query Logs

The most basic form of repetition is the literal duplication of a given query string. The ratio of tokens to types can serve as a signature of a query log. It usually lies in a corridor spanning from 2.5 - 7. Greater deviations from these numbers should be met with suspicion on how the log was gathered.

Obviously, if the query log is cut off at a frequency threshold, the numbers are quite different compared to a full query log down to frequency 1. In YO-3, the following figures can be observed:

# query types 142.429.367
# query tokens 3.631.854.318
average 25.6

Comparing similar logs — in this case, AV1 and AV2, which represent two subsequent months on the same Search Engine — should yield similar figures, if these figures are really able to serve as signatures of a log. Indeed, the numbers for AV1 and AV2 are very similar (despite their different size):

AV1

# query types 169.355.866
# query tokens 447.257.030
average 2.64

number of queries appearing only once: 125.727.533
in percentage of query types: 74.2%

AV2

# query types 111.229.844
# query tokens 290.599.105
average 2.61

number of queries appearing only once: 83.745.132
in percentage of query types: 75.3%

The comparison of these figures for Web Search Engines to GY is elucidating. In the GY query log, the frequency descends much steeper than in the generic Web search query logs.

# query types 3.702.853
# query tokens 18.736.196
average 5.06

number of queries appearing only once: 2.287.672
in percentage of query types: 61.8%

Broken down by logarithmic classes of percentage of all types, the following distribution is returned on AV1:

Number of query types	Percentage of all types	Percentages of all tokens
16.936	0.01%	12.6%
169.356	0.1%	24.0%
1.693.559	1%	39.2%
16.935.586	10%	59.1%
84.677.933	50%	86.3%
169.355.866	100%	100%

This table has to be read as stating for example that the top 50.000 queries represent 1% of query types, but cover 8.110.580 tokens, a 44.3% of query tokens. It demonstrates how tackling the top queries allows for a relatively large impact in terms of search event coverage.

At the other end of the frequency distribution, the query log tail contains rarely asked search terms that nevertheless contribute to the total distribution pie. Most individual query types appear only once.

The same table for GY looks of course quite different:

Number of query types	Percentage of all types	Percentages of all tokens
373	0.01%	28.8%
3.729	0.1%	48.5%
37.285	1%	61.1%
370.285	10%	74.2%
1.851.426	50%	81.1%
3.702.853	100%	100%

13.2. Head/Container detection

Looking at complex queries, i.e. queries that consist of more than one word, it is often observable that a kind of hierarchical relationship holds between the parts of the query. This relationship has to be recognized and preserved when delivering results to the query. For example, the query *pictures beatles* is certainly never answered well if the result is not about the Beatles. Even if the result presented does not contain a picture, it might still be a good starting point for finding those[1]. The part of the query that is essential for a relevant result will be called head, whereas the part of the query that reflects the facet of a query will be called container.

Heads appear isolated or combine with a characteristic set of containers. , i.e. there exist largely populated classes of heads that are discernible through their combinatorial properties with containers. Usually the isolated occurrence is the most frequent distribution for heads. If looking at the following queries containing the head *britney spears*:

[1] See also [Broder 2002] who reported that for 15% of all searches, the desired result is a collection of links rather then one site related to the topic of the search (results of a survey).

freq.	term
1982611	britney spears
129009	britney spears nude
56920	britney spears naked
46055	britney spears pictures
38984	britney spears pics
24200	britney spears lyrics
14803	nude britney spears
13744	britney spears wallpaper
10327	pictures of britney spears
10269	britney spears nude pics
9771	britney spears photos
8720	britney spears porn
6847	sexy britney spears
6834	anti britney spears
6751	naked britney spears
5999	britney spears fakes
5399	britney spears breasts
4968	hot britney spears pictures
4861	pics of britney spears

The containers that appear in this list *pictures*, *lyrics*, *nude* (which is in fact a reduced form of *nude pictures* or *nude videos*) occur with thousands of different heads other than *Britney Spears*.

The distribution of lines alone, however, is not sufficient to discriminate heads and containers. If looking at the top frequent queries containing the container *lyrics*, there is no immediate difference observable that distinguishes its distribution to that of *britney spears*.

freq.	term
1863989	lyrics
547931	song lyrics
211967	music lyrics
88558	rap lyrics
42945	linkin park lyrics
37329	country music lyrics
36449	leonies lyrics
31165	shakira lyrics
29654	ludacris lyrics
28570	lyrics search
28030	lyrics to songs
27572	mandy moore lyrics
26673	R&B lyrics
25731	incubus lyrics
25180	hip hop lyrics
24800	country lyrics
24200	britney spears lyrics

In both cases, the isolated occurrence (*britney spears* and *lyrics*) is the top frequent one and the term *lyrics* combines more or less frequently with other terms. Even the top frequency is very similar for these two examples. Supposing that nothing is known on the semantic of the terms (for example that *Britney Spears* is a named entity), what indications in the distribution are available to deduct the status of a term as head or container?

Separating heads and containers based solely on the frequency list is not a trifle task. Head and containers share a structural similarity, given that both combine mostly with instances of the other type of terms and not with instances of their own type. Making it even harder to distinguish between the two types, terms that in general act as containers in queries can also function as heads in other queries. In the example above, *lyrics search* has *lyrics* as its head.

The main difference on the level of frequency and repetitions is that there are far more heads than (normalized) containers. This can be observed by the decline in frequencies in the lists above, which is much slower in the second list. This effect is even amplified if the containers in the first list are normalized. In the top 100.000 query lines of YO, there are 34 queries containing *britney spears* types, but 278 containing *lyrics*. The first 1 million query lines have 324 lines containing *britney spears* and 3.517 queries containing *lyrics*. For the first 10 million lines, the numbers are 2.476 and 32.756, respectively. It is therefore possible to discriminate between heads and containers by counting the occurrences of terms in the log. Of course, these number have to be normalized by the total frequency of a term. Given that heads should have a higher *standalone* frequency (frequency of the query that just consists of the head) than containers do with all other things being equal, the ratio of standalone frequency to number of all occurrences (or at least occurrences in a large segment of the log) is typically higher for heads.

term	standalone freq	# occ.	ratio
child	53535	488	109.7
rainbow	53510	125	428.08
herturn	53414	2	26707
work	53368	442	120.74
leather	52713	403	130.8
escort	52665	153	344.21
bodysolutions	52586	3	17528.66
clonecd	52560	22	2389.09
firstunion	52538	17	3090.47

Although both heads and container may vary in their expressions, container do so to a much larger extent. Some examples of head variation have already been presented in the context of company name variations (see above, Part B, "E-Commerce Term Management"). Further below, several container types and their variant expressions will be illustrated.

13.3. Contexts, Instances and bipartite graphs

Based on the context and instance operations already introduced (see above, Part A, Term Spaces) an iterative algorithm to segment queries into heads and containers is presented in this section. As a prerequisite, a data structure that allows a quick context and instance operation such as the index structure described above is assumed.

For this purpose, two functions are introduced, the head and the container extension. The head extension starts with a set of containers and calculates all heads that appear within these containers. Conversely, the container extension starts with a set of heads and calculates all containers that appear with these heads.

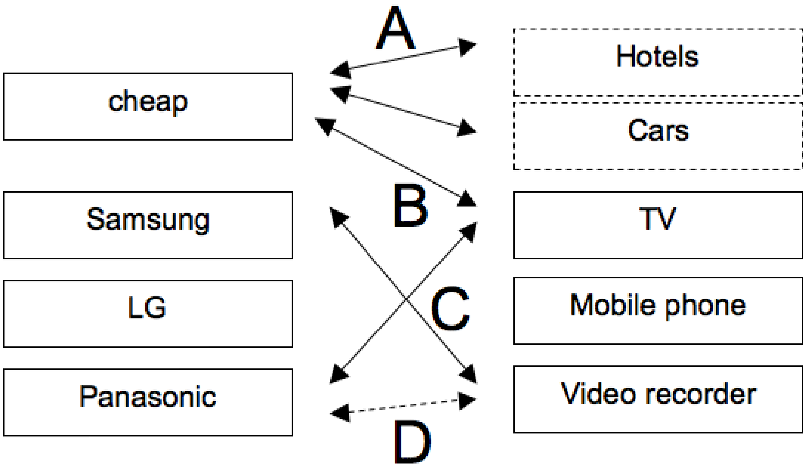
Following the extension step, a pruning step retains the top third in frequency of these extensions that are not in the top third with regards to spreading. For example, a container extension such as *cheap* combines with millions of different heads. Its spreading value is therefore very high and it will in almost all cases lie in the top third spreading units. The results from the pruning step can be used to start another extension step.

The task of setting up head and container lists can be viewed as bipartite partitioning of a graph. Elements of the container and head lists do not in general combine with other elements of the same list. This property leads to a disjunction between the container partition and the head partition of a bid-log or query-log.

In the context of finding semantic relationships based on large corpora, [Biemann / Osswald 2005] presented a so-called "pendulum algorithm" that uses a bootstrap approach. The two steps of finding items and verifying allow the generation of a relatively clean new set based on few seeds. The bootstrap approach works with rules using tags. Related approaches can be found in named entity recognition approaches, such as [Riloff/Jones 1999].

The notion of bipartite graphs applied to heads and containers is illustrated below. On the left hand side are typical brands in the electronic entertainment domain, on the right hand side are typical products in the same domain. Arrows B and C represent edges in the bipartite graph that corroborate the classification. Arrow D represents a combination of brand and product type that should be there if the graph contained all combinations, but the particular combination is not observed in the data. Obviously, there will in general be many missing combinations between the two partitions in a head/container-based bipartition of a query-log. It is therefore necessary to allow the detection of these bipartition without requesting the existence of all possible edges between the nodes of the partitions.

Moreover, the detection process even has to take care of outliers, i.e. edges between nodes within one partition. While this breaches the theoretic model of bipartitioning, such a behavior is likely to occur in real-life query data by virtue of second-level containers. Arrow A represents such an outlier (a second level container) that has an edge into the set of product types, but many more edges outside this set.



Bipartite Graphs and Head/Container calculus.

13.4. Common Query Containers

After having introduced the concept of containers from the viewpoint of repetitive patterns and having associated it with semantic components of a query, a selection of common query container classes and their instances will be presented in this section.

13.4.1. Second Level Container

Very general containers that can combine almost with any kind of heads will be called "second-level" containers. Instead of modifying the head, they rather modify the whole type transaction that is triggered by the query. A typical example of second level containers are all variants of *online*, such as in queres *search hotels online* or *search hotels on the web*. In the context of Web search, these containers can be largely omitted. They can be detected through their extremely high spreading – they combine with a very large set of different heads.

Second level containers are extremely volatile with regards to their position in a query, as can be seen from these two lines that appear with similar frequency with many permuted variants:

> download free mp3 [11] — free download mp3 [60] — mp3 free download [184] — download mp3 free [10] — free mp3 download [221] mp3 down load free [74]
>
> games download free [14] – download games free [20] – free download games [59] — download free games [31] – games free download [35] – free games download [71]

Second level containers such as *download* or *free* act as free inserts and can appear at almost all positions of a query, for example *free download games – download free games.*

13.4.2. Containers, specific to types of heads

Several exemplary container collections are presented below. The containers can be extracted by starting from a small list of seed containers and enriching this list with the left-right-algorithm described above. The resulting containers build up subspaces of the query space that are rather distinctly separated.

As an example of the head type ORGANIZATIONS, these containers help to extract soccer clubs from a query log:

X ergebnisse
X fussball
X spiele
X fanartikel
X ticket
X eintrittskarten
X fanshop

As an example of the head type LOCATIONS, the following list of containers extracts travel destinations from a query log. These travel destinations heads co-occur partly also with real-estate-related queries.

X hotel
X hotels
X unterkunft
lastminute X

unterkuenfte X
flug nach X
last minute X

A further example is the container list that occurs with books as heads, an example of ARTIFACTS. Note that these containers reflect canonic literature heads rather than contemporary books, as can be seen by their bias to literature instruction:

X interpretation
fragen zu X
inhaltsangabe X
personenbeschreibung X
X charakterisierung

The next large lists of containers all represent requests for a specific media type of content delivery. These containers are extremely frequent in general Web search and are of a second-level nature with regards to the content of the query. Today's Web Search Engines have specialized services for providing media other than text, for example images or videos.

The containers listed below are normalized to the term appearing on the right hand side of the dot. This list illustrates how many container variants exist that can be reduced to a small set of normalized container types:

bilder.pics
fotos.pics
bilder von.pics
neue bilder von.pics
plakat.pics
wallpaper.pics
pics.pics
pictures.pics
gallery.pics
poster.pics
foto.pics
fotos von.pics
photos.pics
photo.pics
wallpapers.pics
jpg.pics
galerie.pics
bild.pics
gallerie.pics
galerie.pics

bildschirmschoner.pics
kalendar.pics
e-mail.contact
adresse von.contact
fotos.pics
hintergrundbilder von.pics
screensaver.pics
poster von.pics
hintergrundbilder.pics
galleries.pics
hintergrundbilder von.pics
pictures of.pics
neue bilder von.pics
biografie.bio
biographie.bio
biographie von.bio
biografie von.bio
lebenslauf.bio
biography.bio
lebenslauf von.bio
leben.bio
kurzbiographie.bio
lyrics.text
songtexte.text
texte.text
songtexte von.text
liedtexte.text
liedertexte.text
texte von.text
text.text
liedtexte von.text
mp3.music
midi.music
midis.music
topless.nude
nude.nude
naked.nude
nackt.nude
ganz nackt.nude
nacktbilder.nude
sexy.nude
nackt bilder.nude
nacktfotos.nude
sexy bilder von.nude

shirtless.nude
ganz nackt.nude
nude gallery.nude
nacked.nude
sexy pics.nude
nude pictures.nude
nackt free.nude
nackt fotos.nude
porno.porn
pornstar.porn
pornos.porn
pornobilder.porn
sex.porn
pornostar.porn
hardcore pics.porn
interview.interview
interview mit.interview
interviews.interview
video.video
dvd.video
videos.video
filme.video
avi.video
fakes.fake
clip.video
video clip.video
fake.fake
fanclub.fan
fanpage.fan
fanartikel.fan
fan.fan
fan club.fan
fanpage.fan
fanshop.fan
fan page.fan
news.news
nachrichten.news
neuigkeiten.new
werke.info
discographie.info

14. Query Spaces

Gathering the methodic instruments laid out in Part A and Part B, the Term Spaces that are created by user queries are examined in this chapter. Among the questions that are going to be addressed here are the following:

- What does it mean for a query to be equivalent to another query?
- What possible normalization steps can a query undergo and to what extent do these steps apply to real-life large query logs?
- What are the stylistics of queries, including query length and query syntax?
- What topics can be discerned on a large scale in query logs?
- How can the tail of a query log be best treated?

The query log resources that are going to be examined are YG (to frequency count 3) and YO-3. In addition, the Yellow Pages query log YG is compared to these general Web search logs.

14.1. Query Normalization and Equivalence

Before any sound statements on the performance of search systems can be made, the question of what user queries are equivalent to each other has to be answered. Obviously, the answer to this question is to a large degree dependent on the nature of the search system and the type of query issued to it. In some cases, the only equivalency holds between literal identical queries, for example if a user tries to retrieve a webpage by entering an exact quotation from it.

Assuming queries, however, that are a formulation of a user's informational or transactional intention, it is apparent that there is often more than one way how to word such searches. In TE-Commerce the usage of singular and plural, for example, does in general not constitute a different query intention. Therefore, a query *flat panel tv* and *flat panel tvs* is equivalent from the perspective of transactional intention.

As a process that connects equivalent queries, normalization steps transform the spell-out of individual queries in a more abstract form. Equivalent queries share the same normalized form.

The main benefit of query normalization is an improved consistency in query handling. All the normalization steps proposed here preserve the user's intention modulo to TE-Commerce. Therefore, their results should be the similar or identical on

TE-Commerce search sites. Applying the normalization steps described below would increase the consistency and reliability of search sites. As a side-effect, they can be used as test battery, as divergent or even disparate results would be an indication of a search system's drawbacks (see also Part D, Improving a YP system with vocabulary enhancements for a Yellow Pages test battery).

The normalization procedure takes an individual query and transform it via a pipeline of normalization levels. From a macro point of view on the search system, the question of how these normalization procedures affect a complete query log arises. This means nothing more than applying the normalization procedures on all the different query types in the log and observing the folding ratio of the log. In the following paragraphs, a lowercased version of the query log in the form of a pure frequency list is assumed to serve as starting point for the different normalization stages. The first steps introduced below are removal steps. They detect queries that cannot be answered via a TE-Commerce result, either because they are meaningless or belong to navigational URL queries.

14.1.1. Step 1: Junk removal

In a first filtering step, all queries that are apparently junk strings are removed. By a junk string is understood a string that in all probability does not carry any meaning and is produced either through an artifact or a user playing around with the search engine.

Indications that are characteristic of junk terms are laid out in Part B, for example repetitions of a single letter more than three times in a row. Instances of such junk terms that appear at least 5 times in the YO Log are presented below:

aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa

aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa

aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa

tttttttttt

tttttttttttttttttttttttttttttrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrr

rrrrrrrrrrrrrrooooooooooooooooooooooooooooooooooooooooo

aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaanj

aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatorquere

alabamaalaskaarizonaarkansascaliforniacoloradoconnecticu delawaredcfloridageorgi ahawaiiidahoillinoisindianaiowakansaskentuckylouisianamaine marylandmassachusetts michiganminnesotamississippimissourimontananebraskane vadanew

Search Search Search Search Search Search Search Search Search Search

‘! 1‘1‘1‘1‘

SDFPOJWSSSDSDFPOJWELGWSSDFPOJ GWSSDFPOJWELGWSSDFPOJWELGWSS DFPOJWELGWSSDFPOJWELGWSSDFPOJ WELGWSSDFPOJWELGWSSDFPOJWELGW SSDFPOJWELGWSSDFPOJWELGWSVFPOJ WELGWSSDFPOJWELGWSSDFPOJWELGW SS-

DFPOJWELGWSSDFPOJWELGWSSDFPO JWELGWSSDFPOWELGWSS-DFPOJWELGW

===]=]=]========]=]=]===== ===]=]=]====== ==]=]=]=== =====]=]=]=== =====]=]=]== ==== ==]=]=]========]=]=]=== =====]=]=] ========]=]=]==== ====]=]=]========]=]=

Apart from overlong words, other cases of junk queries are queries containing unrecognizable encodings or no alphabetic chars at all. Chars such as quotation marks, greater/lesser-than, hashes, asterisks, pluses or all kind of parentheses are deleted from the query. Queries that contain anything else than word chars, spaces, dashes, punctuation marks and few additional chars such as dollar signs after this removal step are discarded.

All queries that do not at least contain two alphabet chars are removed, just as all queries that contain one char appearing more than three times in a row. In addition, queries with a length over 200 are eliminated.

This had the following effect on YO-3:

TOTAL TYPES 142.429.367 – TOKEN 3.631.854.318
WITHOUT JUNK 127.077.330 – TOKEN 3.436.367.854

I.e. 89.4% of types and 94.6% of tokens remained in YO-3 after the junk query removal step.

In comparison, GY has a much larger number of junk terms which are here mostly telephone numbers and are discarded because less than two alphabet chars occur within those:

TOTAL TYPES 3.702.853 – TOKEN 18.736.196
WITHOUT JUNK 2.147.191 – TOKEN 14.601.783

I.e. 58.0% of types and 78.0% of tokens remained in GY after the junk query removal step.

14.1.2. Step 2: URL removal

The detection and removal of URLs works as a twofold process. In a first pass, all occurrences that fit the pattern *www*, a space or a dot, a string of alphanumeric chars, again a space or a dot and finally two or three letters, are gathered:

www[\ .] \w+[\.] \w \w \w?

The matches are stored and are used in a second pass to detect URLs without leading *www*, such as *hotmail.com.* Variants with and without dots — including deviations such as using columns instead of dots — are also covered, for example *www:hotmail.com.*

Naturally, queries such as *hotmail* will not be removed by this approach. In these cases, however, it seems fair to keep such queries in the query space, given that their character as navigational URL is not unambiguously discernible and that they cannot be separated thus from a query such as *bmw*.

Although URLs are not bad queries per se, they do not contribute to the term content of a query space. Therefore, they can be removed before entering the further stages of the normalization process.

Applied on YO-3, the following figures resulted:

TOTAL TYPES 127.077.330 – TOKEN 3.436.367.854
WITHOUT URLS TYPES 83.980.550 – TOKEN 1.017.800.016

I.e. 66.1% of types and 70.4% of tokens remained after the URL removal step.

In comparison, the GY log contained much less URLs and the number of types and tokens that became removed is smaller:

TOTAL TYPES 2.147.191 – TOKEN 14.601.783
WITHOUT URLS TYPES 2.107.321 – TOKEN 13.684.275

I.e. 98.1% of types and 93.7% of tokens remained after the URL removal step. Apparently, though, the GY log contained relatively more URLs in its top section than YO-3. This is due to a small number of very common URLs that appeared in this YP search (such as `www.ebay.de` or `www.google.de`) and the absence of a long tail of URLs.

14.1.3. Step 3: Ordering

The next step re-orders the terms in each query according to their alphabetic position. Bases on the representation of the Term Space described in Part B, MWUs are shielded from being rearranged in this step. For example, the query *central park new york* is not transformed to *central new park york*.

Applied on YO-3, this led to a reduction in query types from 142.420.000 to 115.120.110, i.e. only 80.8% of the original size in types.

This shrinkage is considerably larger than what can be achieved by re-ordering GY. Here, out of 2.147.191 types, 1.968.146 remained. The size of the folded query space is still 91.7% of the original size.

14.1.4. Step 4: Container and Head normalization

Following the different term normalization steps laid out in Part B, the query space can be compressed without affecting the intention of the queries. Basic operations are the folding of spacing and hyphenation variants, singular/plurals and container variants.

For example, the query pair *pics bmw-x5* and *pictures bmw x5* would be folded into one through this step. Although this would probably squeeze the query space considerably, this step was not experimentally tested yet, given that the synonym and head/container lists are still in the process of being gathered.

14.2. Query Pies

14.2.1. Distribution of Query style

Even without knowing anything about the terms used in a query, some characteristics are immediately observable. Most queries consist of a set of literals, while few make use of advanced syntax, i.e. metacharacters. The usage of such interface specific elements is limited ([Silverstein/Henzinger/Marais/Moricz 1998] reports a 20.4% of queries making use of them, newer studies report an even lower value). A considerable part of those queries using advanced syntax use their own home-brew syntax that will not affect the search-engine. In the model of interactions laid out above, the similarity between the use of advanced syntax and navigational drill-downs was touched upon. For example, the use of the NOT-operator to exclude homonyms to a query could be functionally substituted by a query refinement box displaying the several readings and allowing the user to choose from it.

A further immediate observation regarding query strings is their length and the number of white-spaced separated words they consist of. The query pie approximatively can be divided into queries containing one word, those containing two words and those containing more than 2 words. [Silverstein/Henzinger/Marais/Moricz 1998] report for the Altavista Querylog from 2nd of August 1998 to 13th of September 1999 the following statistics[1]:

1 word in query	32,5%
2 words in query	32,8%
3 words in query	34,7%

In comparison, from analyzing YO-3 the following pie results (note that this log is truncated at frequency 3, so queries appearing only once or twice are not taken into account):

1 word in query	33,0%
2 words in query	37,5%
3 words in query	29,5%

The boost in two words queries is striking. Most two word queries are either build up by one multi-lexemic unit such as a named entity or reveal a head container structure.

Here are the top two word queries from YO:

[1] Percentage was recalculated on the basis of queries containing at least one word.

ask jeeves
cheap flights
car insurance
travel insurance
internet explorer
bbc news
british airways
estate agents
free ringtones
friends reunited
mobile phones
currency converter
national lottery
free sms
yellow pages
pc world
inland revenue
chat rooms
the sun
auto trader
harry potter
cheap holidays
cd covers
sky news

All of these top ranking two word queries fall in either the category of multi-lexemic units or are build up by a generic container (*free* and *cheap*) and a head term. In addition to this, all of these two word queries are relevant for E-commerce in the model we adopted. Apart from these basic syntactic properties, the semantic status of these queries gives rise to an interesting subdivision.

By analyzing a sample set of about 1.700 query tokens that contain only one word, it can be seen that most (70%) one-word-queries are named entities such as *skinceuticals, euroseal, trackpower, str8guys* or *2a59* — a model ID. In few cases, a bona fide TE-Commerce term or its misspellings appears as a one word query (such as *cafe* or *greetings*). Much more often one encounters one word queries that are misspellings of multiple word queries written without spaces (*sonydvdrom*).

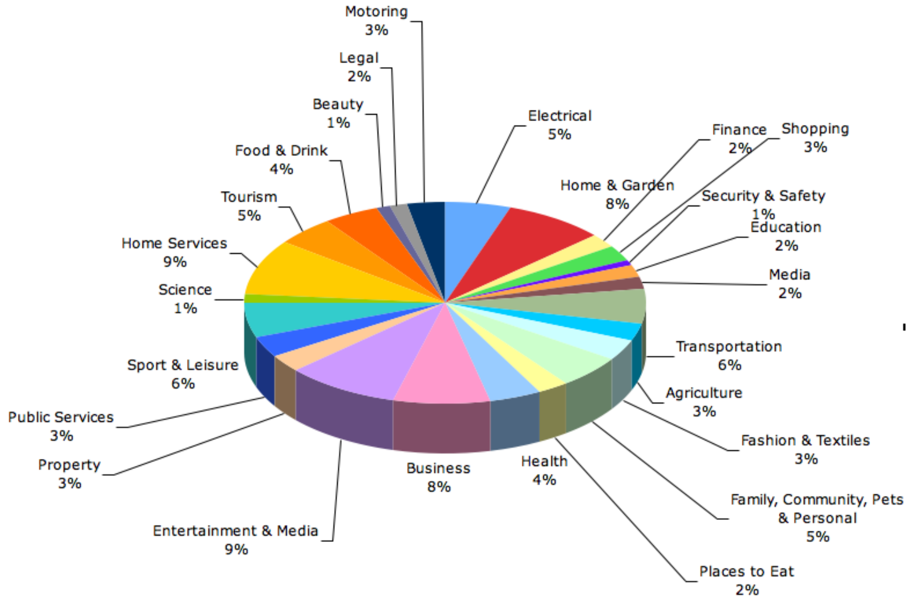
Queries that contain more than four or five words are usually those that are build analog to a natural language question. Examples from AV1 are *where can i learn about the movie star wars, which movie won the academy award for best film in 1996, who was the first african american to win an academy award, where can i read reviews of the current movies* or *where can i find information about howard stern.* In addition, one encounters queries that make excessive (and sometimes ineffective) use of syntax operators among the longer queries: *epson and stylus and color and 600 and print and problem* or *+pizza +”olive oil” vegetarian -restaurant +recipe -kids -lori’s -father.* A final group of comprehensible longer queries are those that search for a named entity with a long name, for example *i don't want to miss a thing aerosmith*. In AV1, the threshold seemed to lie at ca. 18 words, beyond which all queries are either teeming with syntax operators or are plainly incomprehensive.

14.2.2. Distribution of Queries per sector

Taking a look at the topical distribution of queries, an Overture/IMRG study from May 2004 reported the following shares in shopping-related searches per sector[2].

Entertainment & Leisure 61%
Travel 20%
Health & Beauty 2%
Home & Garden 3%
Food & Drink 2%
Property 6%
Apparel & Footwear 3%

A much larger test, based on 18.342.634 (42%) recognized query types from a set of 43.896.967 types (YO-3 down to frequency 5) and a list of 22 million terms that significantly often appear on websites of businesses registered to a known YP category, extracted the following pie:



Distribution of lines with non-lexical words (blue).

[2] A summary can be found online at `http://www.netimperative.com/2004/05/06/RESEARCH_Etail_search` [Nov. 1, 2006].

The terms were weighted differently depending on the logarithmized frequency within the websites that belong to a businesses registered to a category. Through this, the best term representatives of a category had the biggest say. For example, the category "Finance" was triggered by terms such as *balance sheet, financial services, mortgage info, treasury services* and more than 370.000 terms more.

Motoring	3%
Electrical and Electronics	5%
Home & Garden	8%
Finance	2%
Other Shopping	3%
Security & Safety	1%
Education	2%
Media	2%
Transportation	6%
Agriculture & Groceries	3%
Fashion	3%
Family, Community	5%
Places to Eat	2%
Health	4%
Business	8%
Entertainment & Media	9%
Property	3%
Public Services	3%
Sports & Leisure	6%
Science	1%
Home Services	9%
Tourism	5%
Food & Drink	4%
Beauty	1%
Legal	2%

14.3. Saving the tail

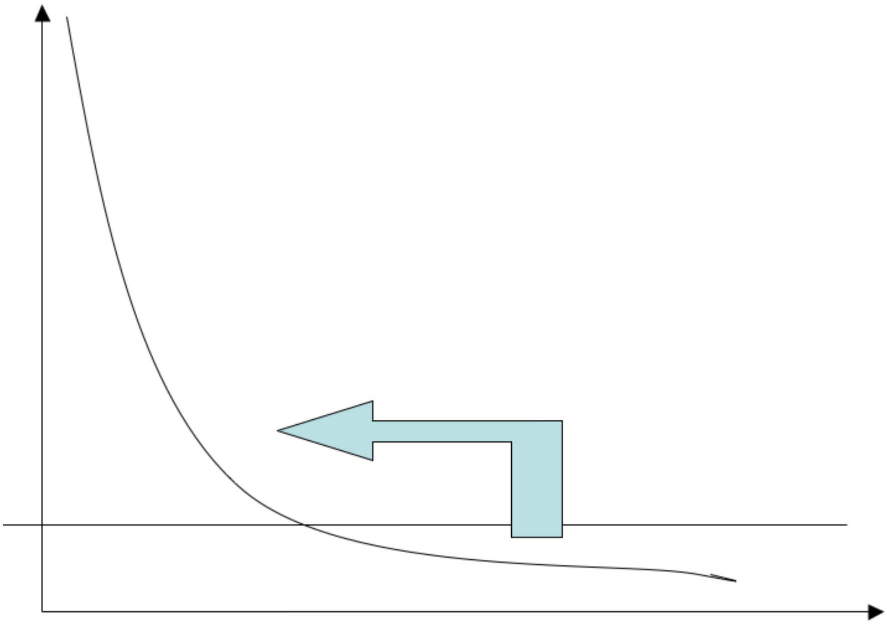
The tail of a query log, i.e. the large number of rare events in its frequency distribution, has enticed many because of the apparent wealth of information that is hidden there.

Chris Anderson pointed out to the importance of filters to direct users from hits to niches, i.e. from the short tail to the long tail[3]. This process is described as business opportunity and being boosted by new Web phenomena such as collaborative filtering, tagging, social recommendations etc.

The task that the tail poses can be seen quite differently from the perspective of TE-Commerce — exactly the other way round, leading from the unknown to the known

[3] See [Anderson 2006].

as a means of recuperating terms. This relates to the hypothesis that most of the low frequency items can be related to queries from the top frequency spectrum by recognizing different types of query variation.



Recuperating the long tail by associating it with elements from the short tail.

A sample test was conducted on a 500 sample out of 15.412.932 hapax legomena queries of YG. An extract of this is presented below. It was tried to find the closest matching query with frequency eight or higher. The mapping steps were entitled to lose some granularity, but should not make a substantial loss with regards to the user's intentions.

The different components that would ensure these mappings are put in capital letters. They refer to the term normalization routines as laid out in Part B. Instances in this sample that could not have been related to higher frequency queries are marked with exclamation marks, instances for which the association seems dubious with a question mark.

A first example is the query *einstellung e mule* from the tail. Through a cascade of transformations, it can be associated with the query *emule einstullungen* that appears 276 times in YG:

$$\xrightarrow{SING/PL}$$

einstellungen e mule

$$\xrightarrow{SPACING}$$

einstellungen emule

$\xrightarrow{ORDER}$

emule einstellungen

The following examples are presented in less detail. The necessary transformation modules are written in capitals and are explained below:

win98 nach xp installieren → 8 windows 98 unter windows xp installieren HEAD_LEX STOPWORDS

olympia office center →1790 Olympia HEAD_LEX::Brand

perfomance ipaq 3970 → 136 ipaq 3970 HEAD_HEURISTIC::Product_Designator

Hotel Frankfurter Buchmesse → 215 hotel frankfurt HEAD_LEX::Geo

univisiet ?? → APPROX_WORD 12 univision OR 213 universitaet

domenikbarber !!

VOL+Mustervetrag+ansehen → 280 mustervertrag APPROX_WORD (mustervetrag   mustervetrag) CONTAINER _LEX::media

schwarzenberg neustädter hof → 1101 Schwarzenberg HEAD_LEX::Geo

automotive and kongress !!

Lederrundecke rot → s 41 rundecke COMPOUND CONTAINER_LEX::material CONTAINER_LEX::color

schiesserei in Frankfurt 10.02.2003 !!

bibel arche noah → 116 arche noah HEAD _HEURISTIC::Frequency

3510i download → 19 3510i software CONTAINER_SYN download software

hpmhftvc.dll !!

besäumlade !!

billen videos → 11 billen CONTAINER_LEX::media

Transporte Spedition 02371-4 → 11 Spedition Transport Gefahrengut Logistik JUNK_REMOVER ORDER

haniel medical center → 79 haniel CONTAINED

Technics und Bedienungsanleitung → 1135 Technics HEAD_LEX::Brand

wrapsrezepte → 14 rezepte wraps SPACING ORDER

weisangen knzelsau → 1940 künzelsau HEAD_LEX:Geographical name

genieser → 16 harald genieser ?? CONTAIN

Gericom Garantie → 1364 gericom HEAD_LEX:Brand name

"Zucker ziehen" → 8 zuckerziehen buch OPERATOR_REMOVE SPACING CONTAIN

C&AMode → 8 c&a mode SPACING

hausjraun akt 40 nackte hausfrauen APPROX_WORD(hausjraun¿hausfrauen) CONTAINER_LEX(akt bilder) ORDER

Alte Mädchenschule !!

Final Fnatasy clips 14 final fantasy bilder APPROX_WORD(Fnatasy¿fantasy) CONTAINER_LEX(clips bilder)

A brief explanation of the components used in the transformations above:

STOPWORDS
Variants due to occurrence of filler words.

ORDER
Permutations on the term order.

SPACING
Variants due to spacing.

CONTAIN / CONTAINED
The query from the tail is contained in a query that is more frequent or vice versa.

CONTAINER_LEX
Container variants, listed in a dictionary.

HEAD_LEX
Heads listed in a dictionary that allow detection of head variants in queries through a look-up.

APPROX_WORD
Matching through edit distance to a word with a logarithmically higher frequency.

OPERATOR_REMOVE
As a preprocessing step, every search engine operator (such as + or “) becomes removed

JUNK_REMOVER
Incomprehensible parts of a query, for example sequences of digits become removed.

Part D.

Case studies of Applications

15. Applying taxonomies in TE-Commerce

Although taxonomies and ontologies[1] are often praised as the solution for accessing and structuring textual information and excessive research has been devoted to applying ontologies in the web[2], examples of real-life Web E-Commerce applications that utilize ontologies do not spring to mind easily.

Two lines of arguments why this is the case can be distinguished. One general objection points to inherent drawbacks of the concept of taxonomies, while a softer variant points to drawbacks of real-life instances of taxonomies.

The first line of argumentation might be called the "voodoo-classification" argument. It challenges the principle value of a priori established relations for open domains such as the Web[3]. This argument denies the very principle of taxonomies — the ordering and structuring of the world through concepts in one central repository. Especially in an open environment such as the Web without central authorities, coordination or clear edges, the expenses of forcing a view of the world onto users would render such endeavors infeasible.

A second line of argumentation does not deny the principle benefits of using a controlled and structured vocabulary, yet points out that current ontologies are not capable of capturing the breadth of actual user queries. As discussed previously, in part B, "E-Commerce Term Management", squeezing millions of search queries into the usual ten thousands of taxonomy types is not a trifle task. Moreover, automatic extending and adapting ontologies in the same meticulous fashion expected from a manual curated ontology is not yet available.

15.1. Shortcomings of Taxonomies

In this section it is laid out why using current taxonomies, even those decidedly intended for E-Commerce purposes, is not a feasible way for capturing the breadth of E-Commerce-relevant terms, if no enhanced matching algorithms or substantial modifactions to the taxonomies take place. This holds especially for analyzing queries, but also for treating other E-Commerce term repositories such as advertisers' keywords or vocabulary on business websites.

[1] The differences between these two concepts can be neglected from the perspective of terms.

[2] The European Centre for Ontological Research, situated at the University of Saarbrücken, bundles many effors in this area (/www.ecor.uni-saarland.de). Members in Germany outside the University of Saarbrücken are the Ontology Research Group of the University of Bremen (www.fb10.uni-bremen.de/ontology/) and the CCSW at the DFKI (semanticweb.dfki.de/).

[3] Clay Shirky, "Why ontologies are overrated", online at www.shirky.com/writings/ontology _overrated.html [Nov. 1, 2006].

The shortcomings presented are both inherent to the concept of taxonomies, but also have their origin in historically evolved pecularities of widely used taxonomies and their descendants. A first step will take a look at the structural overhead of taxonomies. Structural overhead is here understood as a property often observed in taxonomies that puts a high cost on term insertions, emphasizes relations in contrast to content and sets up a construct of high level nodes with little value external to the taxonomy.

In a second subsequent step, the negligence of spell-out for taxonomic concepts is demonstrated. This issue falls between reasons inherent to taxonomies and properties of currently prevailing examples of taxonomies. Clearly related to historically evolved properties is the last shortcoming discussed which lies in content gaps and content residues. Here, terms that are missing in prevailing taxonomies and terms that appear almost exclusively within taxonomies, but not in any other real-life data, are examined.

15.1.1. Structural Overhead

The standard work process in taxonomies is a very cautious process that focuses on structure and normalization of concepts[4]. Although various methods for machine learning of relations and new concepts from corpora have been proposed, the bottleneck of entering *terms* into a carefully curated taxonomy remains. Before any term is added, its relation to existing concepts in the taxonomy has to be examined. As many terms are no clear synonyms to a term already existing in the taxonomy, a whole cascade of decisions have to be made, all of which require an editorial choice and high attentiveness (see the diagram at the end of this chapter).

This cascade of issues that have to be dealt with leads to a considerable cost of insertion for new terms. A special-topic taxonomy such as a medicine thesaurus might thus need hundreds of man-years to create and cost up a good six-digit figure in EUR for licensing. Fur most TE-Commerce applications this is far from being feasible.

A related aspect of the high term insertion costs is the ratio of terms to relations. This ratio is the number of unique terms in the taxonomy divided by the number of all relations. Typically, these relations operate on IDs. If presented in the relation table format (see above, Part B), counting relations is a mere line count of all tables of the taxonomy (assuming de-duplicated tables) and counting terms a line count of the table storing relations from IDs to term descriptors. Disambiguated terms that have more than one ID should count as several terms, following a standard in reporting the size of taxonomies. A ratio of 1 indicates a flat list with no relations between terms, as then the number of terms equals the number of relations, including the relation between ID and term descriptor.

The average depth of a term is a related metric. It is only applicable for hierarchic taxonomies that provide a path towards the top nodes for each term. The average depth is defined here as the algorithmic mean of the length of all navigational paths. A navigational path to a node t is the concatenation of nodes from the (in general

[4] See [Gomez-Perez / Fernandez-Lopez / Corcho 2004], chapter 3, “Methodologies and Methods for Building Ontologies”.

virtual) single top node to t. For example, a navigational path to *soccer shoes* could be *sports — sporting equipment — soccer equipment — soccer shoes.*

For illustrating these metrics, the WAND business taxonomy[5] is used as an exemplary taxonomy for E-Commerce purposes. It is a multi-lingual taxonomy with over 40.000 preferred terms that is mainly used to power B2B-portals, but has also been applied to enhance search functions on Local search and Yellow Pages search.

With a maximum depth of 10 levels, here is that table that summarizes how many nodes are at each level of the hierarchy (with the virtual root node at level 0), sorted by the number of nodes:

count	*level*
20695	3
18221	4
8663	2
8573	5
7476	1
2876	6
1116	7
59	8
17	9
15	10

This yields the average depth of 3.38 with a standard deviation of 1.39. This indicates a fairly balanced hierarchical tree, yet with some branches that go much deeper than the average three to four levels. The large number of types on level 1 is largely due to specific chemicals that all appear below the top node.

A further aspect of structural overhead is the requirement in taxonomies to group related items under one node, even if there is not a commonly used name for this node. For example, it is clear that the concepts *beds, sofas, futons* and *couches* should be grouped together, but there is no widely used name for this group. Usually, the labeling of such nodes used coordinated terms. Some of these terms are so widely used in category systems that they have become lexicalized and show traces of a meaning distinct from the components, for example *Home, House & Garden.*

This issue becomes especially virulent in the top classes of taxonomies. Here it is not uncommon to face constructs that only exist within the sphere of taxonomies. Even on first glance, one can discern terms such as *Mining And Quarrying Of Nonmetallic Minerals, Except Fuels, Apparel And Other Finished Products Made From Fabrics And Similar Materials* or *Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks* to be part of the SIC classification. Obviously, some processing of these complex term aggregates have to take place before they can be successfully used in a search scenario.

[5] See `www.wand.com`.

One issue that occurs in many examples of real-life hierarchical taxonomies is that the nature of the hierarchy relation is not clearly defined and differs throughout the taxonomy. Moreover, tree-like hierarchies are not well-suited for incorporating independent features as subdivision principles.

In E-Commerce taxonomies, a typical source for inconsistent and redundant subdivision of types are the mixing of product types and transaction type. Transaction types refer to the different types of business activities such as buying, renting, repairing, maintaining, manufacturing, distributing etc. As almost all kind of products can be transacted following one of these transaction types, there is either the option of starting the taxonomy with these transaction types or with product types. In the former case, all different product types have to be reduplicated, in the latter case the same holds for the transaction types.

The WAND taxonomy as an example shows a mixed behavior. While the standard subdivision principle sets up subtypes of products, there is a completely separate branch *services* that stores all the service transaction types. Not surprisingly, this branch of the taxonomy tree is at the same part redundant and shows inconsistencies compared with the rest of the taxonomy. It is redundant when a product type simply re-appears as a service and inconsistent if the corresponding service to a product type or the corresponding product to a service. For example, there is no type *Bottle Cleaning Services* in the WAND taxonomy, although there are *Bottle Cleaning Machines* — in most other cases of cleaning machines or equipment, the corresponding service type is present in the taxonomy.

A further problematic issues of setting up a hierarchy springs from the existence of rest classes that are needed because the division principle below one type does not yield a covering partition of it. These rest classes usually feature negation or constructs with *other* or *except*, for example *Other mining and quarrying not elsewhere classified* or *Growing of other fruit, nuts and spice crops* (see also below, "Pointwise mappings"). Usually, these rest classes percolate through the complete taxonomic tree, leading at the top node to a "super" rest class, for example SIC's "9999" class.

It is in general difficult to draw a line between genuine product types which should be added into the taxonomy and instances of product types which would only contribute to redundancy if being added. Instances-of follow from the combination of one or more product features and are thus more systematic than product types. For example, while *window frames* is certainly a distinct product type, *aluminum window frames* or *rectangular window frames* are rather not, considering that in them other feature values could be substituted and features could be combined for them *aluminum rectangular window*. Obviously, this line is not easy to draw. Indications of an instance are a broad set of feature values and a systematic spell-out of features. This makes *women's sweater* more prone to be a genuine product type than *red sweater* or *wool sweater*. However, the latter example hints at another dimension of this separation that cannot be grasped by observing terms. Those product designators that can be associated with distinct lines of business are genuine product types, making *wool sweater* a much more probable candidate for a product type than *red sweater*. It is hardly conceivable that a business specializes in red sweaters, but much more so that a business specializes in wool sweaters. To incorporate such differences into a taxonomy means enormous efforts and collides with the hierarchical structure of taxonomies, as independent features require a multiplication of branchings in the hierarchy.

15.1.2. Neglecting of spell-out

Most taxonomies are intended for human users that are able to recognize a concept in different spell-outs and can adapt their look-up once they are acquainted with the taxonomy.

Even if taxonomies incorporate common synonyms, they do not regularly integrate morphologic and orthographic variations. For humans searching the taxonomy (especially in a language such as English with a very systematic morphology), this is not a severe shortcoming. It becomes an issue, however, if the taxonomy should be used for automatic processing of texts or queries (see below).

The need in taxonomies for disambiguated concepts clashes with many common searches, as they are often very broad and/or ambiguous. In some cases, a taxonomy might provide larger terms that contain the search as a substring and help thereby to resolve it. For example, the query *frames* could be resolved to *browser frames*, *picture frames* or *bed frames*.

Intersecting the WAND taxonomy with its 83.351 types (37.494 preferred and 45.857 non-preferred) with YO-3 shows just how distant the spell-out principles in a query log and in a taxonomy are. Although all WAND taxonomy types are valid representatives of business-related concepts, 52.138 out of these 83.351 types were not found in the query log[6]. These are more than 62.5% of all types. 34.245 of the not-found were synonyms, i.e. the ratio of "found" to "not found" does not differ largely between synonyms and preferred terms.

A similar picture follows from comparing the German WAND taxonomy with meta keywords gathered from the Web. Comparing 5.787.485 lines of German meta keywords yielded that only the minority of the 51.953 German WAND taxonomy terms[7] (24.295) appear anywhere in these meta keywords, whereas 38.780 do not.

One example of the different spell-out principles is that in the WAND taxonomy all services have the suffix *service*. Therefore, *painting* is present as *painting services* in the taxonomy. Of course, such a behavior will not be found in a query log. Obviously, keeping consistent editing rules in the spell-out of taxonomy types will clash at some points with the incorporation of frequently used terms. The best solution

15.2. Making taxonomies fit for TE-Commerce

After having discussed typical shortcomings of taxonomies, what methods can be used to make them fit for TE-Commerce applications? In the following paragraphs, several

[6] About 5.000 of these types represent specific chemicals.

[7] Version from beginning of 2005.

approaches to adapt taxonomies to TE-Commerce requirement through enlarging the vocabulary and adding structures are discussed.

15.2.1. Merging and mappings of taxonomies

When discussing the challenges of merging and mapping taxonomies, the widely discussed area of structural-driven mappings will be disregarded here, as they can only build upon an already existing term matching framework[8].

Merging and mapping taxonomies has a wide range of applications. One first purpose is to import data assigned to one category. A second purpose is to close the gaps in one taxonomy and broaden the vocabulary available. Finally, merging taxonomies may help to create a super-taxonomy that aggregates all of its component taxonomies.

Data-driven mappings

A time-saving, but not overly precise approach to set up a mapping table is possible if records labeled in both category systems are available. For example, many business address data sets in English-speaking countries use a variant of SIC code in addition to their own proprietary category system. Consider a real-life case where business data is provided categorized to a proprietary category system of approximately 1900 headings and to NACE codes. The NACE codes are available for more than 90% of businesses, while only about half of the businesses half a proprietary YP code. On the surface, it seems conceivable that a mapping between both heading systems can be achieved by counting co-occurrences of categories for individual businesses.

While this approach does in many cases yield good results, it is prone to several types of errors. Errors might occur because companies are registered too more than one code in a category, reflecting different lines of businesses. Through this, mappings can enter the system that tell more about what lines of businesses go together well in a company than what the equivalent terms in the different category systems are.

For example, the following three mappings followed from the process of aligning high frequent correlations from the YP codes (to the left) and NACE codes (to the right):

- Image Consultants – Physical well-being activities
- Furniture Fittings – Manufacture of chairs and seats
- Central Heating – Installation & Servicing Plumbing

Apparently, some co-occurrences of categories are just co-incidental or reflect different lines of businesses than representing valid mappings between the category systems.

In general, failures of data driving mappings occur because it is tried to describe two things at one time. Firstly, an assignment process (for example companies to categories) that is not necessarily consistent and secondly, the correlation of assigned

[8] See [Valiente 2001].

classes with each other. At both stages, errors and inaccuracies can occur which then multiply in effect.

Pointwise mappings

Pointwise mappings from one category system to another can build upon the term variation principles presented above. In this section, the challenges that occur on a detail level when conducting such mappings are discussed.

Typically, most problems that arise when attempting to merge different category systems are not merging issues per se, but rather signals for weaknesses, redundancies and gaps in these category systems. Obviously, mapping to a very rich and detailed category system is easier than to a restricted and not well formulated system.

This list presents common issues when conducting pointwise mappings:

- Mapping between a type with a specified feature and a type without a specified feature (for example *hotels 10-19 beds – hotels*)
- Mapping between types with different logic of subdivision (for example *hotels 10-19 beds – sport hotels*)
- Missing granularity in the target system – in this case, some information inevitably gets lost in the mapping process (for example, *holzschnitzer – handwerk*)
- Missing domains in the target system – in this case, only a pragmatically related fallback to a broad level term is possible (for example *haushaltsauflösungen* mapped to *transport*)

The long-term solution to these mapping issues is to create a balanced and normalized category system that incorporates all the granularity needed for applications and also contains a feature normalization component. Such a system

A sample from a real-life mapping between two Yellow Pages / Business Directory header systems illustrates how different term matching types are used and also that in some cases the mapping has to resort to a broader level. This sample is taken from a group of alphabetically ordered categories (in the second column) that have been mapped to the categories in the first column.

Generators-Electric-Repairing → DC Electric Generators: BROADER / ORTHOGONAL

Decoys → Decoy Birds: BROADER

Decoys → Decoy Calls: BROADER

Dehumidifying Equipment → Dehumidifiers: ORTHOGONAL

Demagnetizers → Demagnetizers: IDENTICAL

Dental Equipment - Repairing & Refinishing → Dental Equipment and Supplies : BROADER / ORTHOGONAL

Diabetic Products → Diabetic Candies: BROADER / ORTHOGONAL

Diatomaceous Earth → Diatomaceous Earth: IDENTICAL

Die Casting Machinery → Die Casting Machines: ORTHOGONAL FACET

Engines - Diesel → Diesel Engines: PERMUTATION

Diesel Fuel → Diesel Fuels: MORPHOLOGICAL —SINGULAR/PLURAL

Donut Making Machines → Doughnut Makers: ORTHOGRAPHIC / MORPHOLOGIC

Dust & Fume Collecting Systems→ Dust Collectors: MORPHOLOGIC

Tallow → Edible Tallow: BROADER

Clocks - Dealers → Clocks: ORTHOGONAL FACET

Clocks - Wholesale & Manufacturers → Clocks: ORTHOGONAL FACET

Metals - Expanded → Expanded Metal: PERMUTATION

Fax Equipment & Systems → Fax Machines: ORTHOGONAL FACET

Fairgrounds → Fairground Amusements: REDUCTION FORM

Eyelashes-Artificial → False Eyelashes: SYNONYMS artificial–false

It can be seen that simple identity occurs rather infrequent and that in many cases the mappings follow a systematic and operational process, such as replacing orthogonal facets or looking for subtypes that are represented by substrings.

In general, between two separate term structures there are always numerous fields that are more finely elaborated in one structure than in the other. This leads to the notorious case of rest classes.

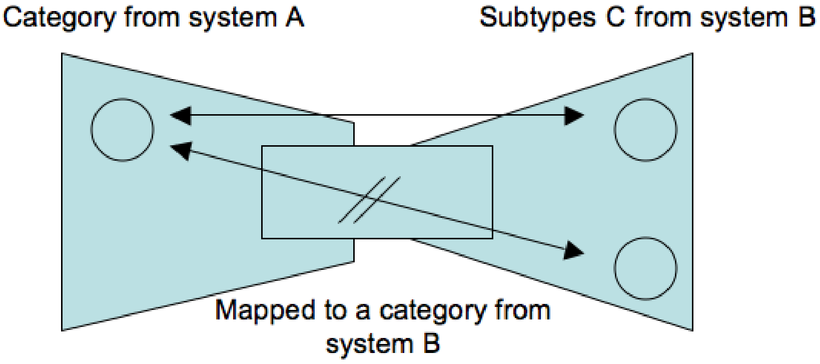
Rest classes commonly occur within category systems and hold all instances that do not fit a subdivision principle or are too few to create an own subclass. The need for rest classes arises purely from usability concerns, namely a limitation on the number of subclasses under one class and the length of the hierarchy path. However, rest classes lead to problems both with regards to the interface and to the categorization of records. On the interface, they hide more than they reveal. For example, consider a subdivision of hotels into wellness hotels, conference hotels, luxury hotels and a rest class holding all other hotels. While someone looking for a luxury hotels immediately perceives that this kind of hotel is in the database, someone looking for pet friendly hotels needs an additional click that usually leads to an unsorted list of everything that did not fit into the systematic. On the data indexing side, the presence of rest classes poses severe difficulties in the mapping process because rest classes from different systems might hold different objects, depending on the division principle into which they could not been integrated.

Mapping cascades

Cascading several mappings of taxonomy types in a row bears considerable dangers with regards to preserve the intention of the types. As mappings between existing taxonomies always bear the danger of losing some aspects of the mapped terms (especially those aspects in meaning that are explicitly in the type's description, but are rather conveyed through the structure), cascading mappings can mean to multiply errors.

One remedy is to introduce a large normalized "intralingual" taxonomy that allows to perform the mappings in a star-shaped manner and not in the form of a pipe. Such a normalized taxonomy has to provide the maximum of granularity needed for applications, because otherwise cross-mappings can occur that do not preserve the intention of types.

The scheme below depicts such a situation — a type from category system A is mapped to a slightly broader type in another category system B. By mapping to a taxonomy, the hierarchical nature of it should be exploited. The subtypes C to the mapped type B should theoretically match to A:



Granularity changes leading to invalid mappings.

While some transitions from A to subtypes C through the mapped type B will be acceptable (the upper arrow), because the subtype C happens to be equivalent with A, other transitions will end up twisted (the diagonal arrow), because the granular information contained in the type in A gets lost via the mappings.

A health check for mappings can be performed by counting the number of types in one category system that can be pulled out by a mapping and vice versa, how many types in the other category can be pulled out by it. If these numbers are high, it could be an indication that a sort of funnel described above exist which may erroneously cross-relate subtypes. For example, mapping *ice skates* to *sporting equipment* will lead to erroneous associations of subtypes to sporting equipment, such as *hula hoop rings*.

15.2.2. Densening ontologies

The task of structuring terminology in order to build up or enrich a taxonomy is a demanding and time-consuming job. Starting from a cleaned list of terms that are known to be relevant to a given domain, it is necessary to set some terms as canonical forms (preferred terms), attach other related or synonymous terms to them (the so called "Used-for" terms) and order the resulting sets into an acyclic structure. Such structures entail hierarchical relations (formed by subsumption, hyperonymy, meronymy or other partitive relations) and the diverse transversal relations (such as "related to").

Adding relations

A common deficit of hierarchical ontologies is the lack of relationships between different branches. This happens mainly because hierarchies inherently adapts to a vertical completion and extending mode. By vertical completions we refer to the process of dividing a broader term into narrower terms following a differentiating principle. As the hierarchy pushes along into more depth, re-occurrences of related concepts, inserted under divergent aspects, will occur. If we take the example *ice skates*, it might occur as subterm to a broader concept of *sporting equipment* (possibly with some concepts such as winter sports equipment in between). *Ice rinks*, however, the place where skates are commonly used, will regularly appear as a subterm to a broader concept of *sporting facilities*. The only relationship between the terms that can be inferred from this hierarchical structure is at best the very broad concept *sport* (if both *sporting equipment* and *sporting facilities* are subterms of such a concept). This relationship can hardly be very robust as it relates all kinds of sporting equipment to all kinds of sporting facilities (from baseball gloves to shooting ranges etc.). Here, a more precise linking between ice-skating terms is needed.

A discussion of relations that can be observed if relating nodes in the WAND taxonomy that are lexically related, but not structurally: They are not in the same subtree and there is no already defined relation holding between them.

Donkeys – Donkey Meat

Here, a meryonomy relation between the two relata can be observed. The pattern "Animal – its meat" is to a limited degree productive.

Abaca Fibers – Abaca Cloth

One element is a higher processed form of the other element. This is a very productive pattern of associations in TE-Commerce, i.e. the notion of vertically integrated chains.

Abatement Services – Noise Abatement Materials

One element designates a service that makes use of the other element that designates a material. This again is a very productive pattern of associations in TE-Commerce.

Aboriginal Organizations and Services – Aboriginal Law Attorneys
Here, both elements are connected through the re-occurrence of a group of humans. The extracted relation gains additional quality because occupational titles such as *attorneys* are not far away from *services.*

Abortion Services – Abortion Alternatives Counseling
This pair is a special case, because one of the elements designates the countermeasure to the other element. Other examples of such a pattern would be *insolvency avoiding* vs. *insolvency services.*

Absorbents – Absorbent Cotton
One element is the generic product type, the other a subtype in a IS-A-relationship.

Actor Trailers – Actor Agents
Opera Glasses – Opera Companies
These two examples follow the pattern "equipment for human group" – "services for human group". This association is in general too weak to produce meaningful relations.

The following cases do not work because the terms (*costume, accordion, make-up*) that triggered the extraction of a candidate pair of related terms are ambiguous and used in different senses for the different elements:

- *Costume Jewelry Stores – Ballet Costumes*
- *Accordion Skirts – Accordion Files*
- *Accordion Skirts – Accordion Doors*
- *Make-Up Air Heaters – Eye Make-Up*

Too general words should not be allowed to enter the process of extracting candidate pairs. This caveat is not only restricted to high frequency words. The following three examples feature with *absorption* and *elevating* working principles that can be applied to too many different domains and with and *aerated* overly general feature of products.

- *Sound Absorption Wall Fabric – Atomic Absorption*
- *Elevating Office Chairs – Elevating Conveyors*
- *Aerated Concrete – Aerated Waters*

Setting up polyhierarchies

As laid out above, perhaps the most important drawback of tree hierarchies in the world of E-Commerce is that they have to settle with dividing criteria of different kind (e.g. domain, type and feature). Another related aspect lies in the polyhierarchic nature of many concepts in E-Commerce.*Office lighting* for example might apply both to the lighting domain, as a subdivision of electronic equipment, and to office equipment, as a subdivision of business equipment. With the aim in mind to set up a classification system that reflected the way people group concepts, it seems inevitable to set up a structure allowing for more than one parent to a concept. In its broadest form, this would be an acyclic directed graph that is suitable to represent polyhierarchies.

A process to enrich an existing hierarchical category system with polyhierarchical relations can be set-up in close analogy to the process described above for finding new relations based on the detection of morphological, syntactical and semantical related terms. For example, this detection routine allows to associate the category descriptor *office lighting* to the two category descriptors *lights & sounds* and to *equipment for offices*.

Using the MetaMatch approach described in Part B, the first pass of the algorithm extracts category descriptors that are termwise related. The first pass delivers pairs (both as A, B as well as B, A) together with the hierarchical context of the categories which is represented as the hierarchy path.

By using the approach described above to detect generality, it is possible to obtain additional indications of what ordering is to be preferred. Additional heuristics that can be applied in this context to determine the order of categories, are the length of hierarchy paths (a short path of a category suggests that this category should be treated as new parent category) and the presence of additional matches to sibling categories. For example, if matches can be found via term matching from *tft displays* to *flat panel manufacturers* and *display manufacturers* and the latter two have are subcategories, then their common supercategory (e.g. *computer equipment manufacturers*) could be considered as the new parent category to *tft displays*.

15.2.3. Extending taxonomies

As a means of enriching an existing taxonomy, an extended terminology can be inserted by looking for slots in the taxonomy for which a rich vocabulary can be found. Extended terminology in a taxonomy implies the acquisition of a very large number of terms with the trade-off of lesser granularity, as these terms are inserted as a flat list.

As the selection of such a slot is crucial, it helps to focus on those nodes in the taxonomy that represent frequently searched concepts and spawn a large vocabulary. For example, the latter demand is met when choosing *spiders* as a slot, because there is a wealth of vocabulary provided by the different subgenera of spiders. However, such an extended vocabulary would have only very limited value to search-type applications.

A better selection could be for example *wines*, as this is both a concept with a very rich vocabulary and belongs to a frequently searched topic[9]. In general, the extensions will be instances of the taxonomy type they point to, either combinations of features with types (*20 inch lcd monitor*) or combinations with brands and/or models (*sony lcd monitor*). The more specific these combinations are with regards to the taxonomy type, the more valuable is the extended term resulting from it — *cheap lcd monitor* is much less interesting to keep in the list of extended terms than *hdmi lcd monitor* is.

This results in a subclass of good slots, built up by product types for which a wide range of specific proprietary product designators can be extracted. Such slots comprise of product types such as *plasma tvs, dvd player* or *washing machines*, for which thousand of models and product type – feature combinations can be found.

Regular inspection routines can be applied to the extended terminology, promoting some terms to structured taxonomy types, grouping others into synonyms and discarding all inapt terms.

15.3. Using ontologies in TE-Commerce search

The main fields of application for ontologies in TE-Commerce search is basically a query recuperation or repair tool. Here, a search that produces few or even zero results is substituted by a search for which a number of results are available. Detecting the query terms in a taxonomy can recuperate the query by trying similar terms or more general terms. Secondly, a large result set that is due to a broad search can be broken down if the narrower terms to the query terms are used as refinements. Lastly, using a taxonomy may help to hint at related topics to the user that are of potential interest to her, but were not present in the original formulation of the query.

15.3.1. Recuperation of searches

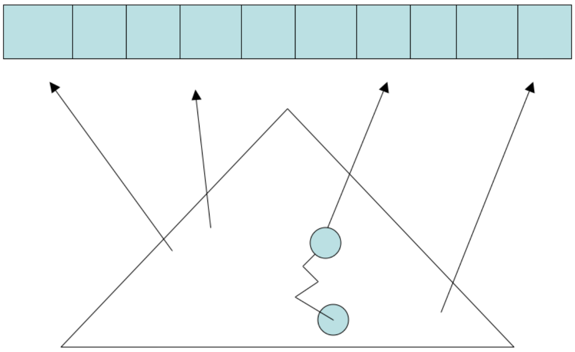
In the case that a search produces a low number of results or even zero hits, one way to repair the search is trying to look-up it in a taxonomy and substitute it by related terms. In general, one would expect that terms reflecting broader concepts should increase the number of results (but see above for anomalies with regards to hierarchy level and popularity of terms).

The substitution process could operate on simple searches, by replacing one string through another which is presumed to have more chances of finding results, for instance a broader term or the preferred term to a synonym. Alternatively, it could replace a simple search by an advanced search that makes use of the OR-operator and inserts all related content of the taxonomy in the OR-clause, such as synonyms, preferred term or related term.

In search contexts that operate with categories, the most natural way of recuperating failed searches with a taxonomy is to find the closest cross-reference to a category in

[9] Wines are often chosen as an exemplary domain. See Natalya F. Noy and Deborah L. McGuinness, “Ontology Development 101: A Guide to Creating Your First Ontology”, online at `http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html [Nov. 1, 2006]`.

the taxonomy. If the category system covers large parts of the taxonomy, locating a concept in the query within the taxonomy should be sufficient to recuperate it by delivering back the most appropriate category or group of categories. The scheme below depicts a hierarchical structure with an entry node and the path towards a category.



Finding the closest cross-reference in a taxonomy.

The circle at the bottom depicts the concept in the taxonomy (the triangle) that was detected in the query. The arrows from the taxonomy represent mappings from the taxonomy to a set of categories (the boxes above). The hierarchical structure of the taxonomy helps to find the closest type for which a mapping to the categories is available (depicted as the zig-zag line).

15.3.2. Refinement of searches

If searches produce a large result set (which happens for many Web searches), a break-down of the result set according to semantic criteria is desirable. Using a taxonomy, this break-down can be achieved in a controlled, covering and non-redundant way.

As was touched upon when treating clustering (see Part A, Term Spaces), many current examples of topical drill-downs do not provide a controlled division of results. The presented subsets are often overlapping and not exhaustive for the original search. Taxonomies could serve better in this respect, assuming that it is possible to find a corresponding node for the search in the taxonomy. If such a corresponding node is pulled out — using the head/container calculus described above (see Part C) — the drill-down can be based on its immediate children. In the back-end, this could be achieved by substituting the query with the immediate children selected for a drill-down.

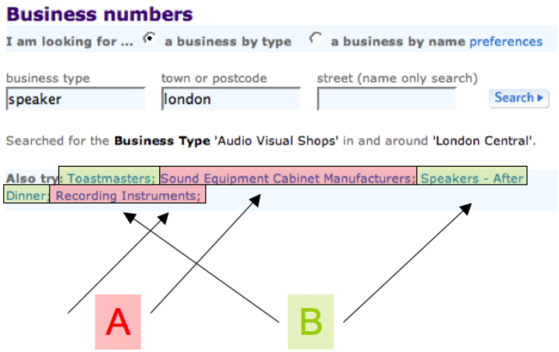
Obviously, such refinements are only feasible if the results have enough information to it. In the context of Yellow Pages search with only categories as the only source of information for a result, one cannot go deeper in the drill-down than the level of categories[10].

[10] In fact, most specialties that are listed in today's Yellow Pages business listings are general properties such as opening hours or accepted methods of payment.

15.3.3. Disambiguation

As described above, numerous searches, especially very short and frequent ones, have more than one meaning. These ambiguous searches are ideally treated by displaying a drill-down based on the different meanings. Assuming again that a matching device helps to find correspondent nodes to a search in the taxonomy, ambiguous searches would produce more than one corresponding node from different branches of the taxonomy.

For example, searching for *notebooks* could produce *Notebooks* as a a taxonomy node both in the branch *School Supplies* — *School Paper Products* and *Notebooks* in the branch *Computers* — *Mobile Computers*. Either high-level terms from these branches or representatives of these (for example frequent queries that fall in the same branch) can be displayed as disambiguation links. In Yellow Pages types of search, the disambiguation would lead to a grouping of categories based on the different meanings:



Business numbers

I am looking for ... ● a business by type ○ a business by name preferences

business type: speaker

town or postcode: london

street (name only search):

Search ▸

Searched for the **Business Type** 'Audio Visual Shops' in and around 'London Central'.

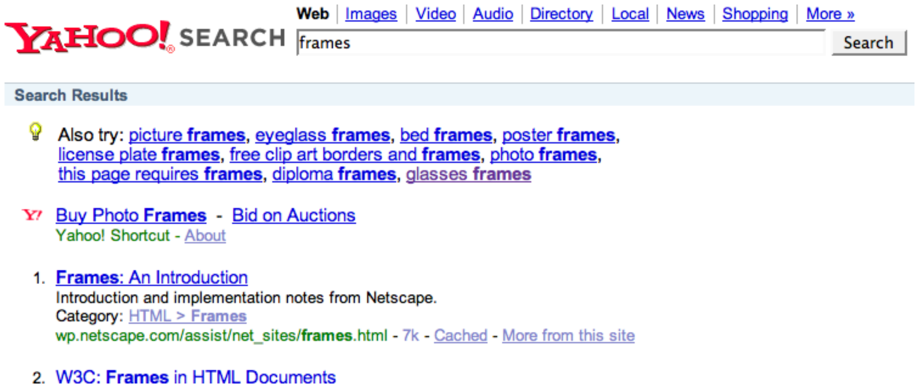
Also try: Toastmasters; Sound Equipment Cabinet Manufacturers; Speakers - After Dinner; Recording Instruments;

Grouping suggested categories by different meanings (A: speaker as audio equipment, B: speaker as professional).

Current clustering Search Engines (such as `www.clusty.com`) are yet not capable of distinguishing between different levels of branching that occur for search results. For example, the search for *foundations* pulls out a division of clusters that are based on instances of foundations such as the *Gates Foundation* as well as a clusters that are based on the different readings. The meaning foundation (of a building) is represented by the cluster named *conrete*[11].

A similar notion to search refinement and disambiguation is the display of related searches. These may help to clarify the user's needs or help to detect related information not thought of when formulating the query. On `www.yahoo.com`, a related search feature shows popular queries that contain the original query. This has often an effect of a semantic drill-down or disambiguation, but it is not conducted in a controlled way. For example, these are the related searches for the query *frames*:

[11] Searching for *foundation* yields rather different results on `www.clusty.com`.


Web | Images | Video | Audio | Directory | Local | News | Shopping | More »
YAHOO! SEARCH
frames
Search

Search Results

Also try: picture **frames**, eyeglass **frames**, bed **frames**, poster **frames**, license plate **frames**, free clip art borders and **frames**, photo **frames**, this page requires **frames**, diploma **frames**, glasses **frames**

Buy Photo **Frames** - Bid on Auctions
Yahoo! Shortcut - About

1. **Frames**: An Introduction
Introduction and implementation notes from Netscape.
Category: HTML > **Frames**
wp.netscape.com/assist/net_sites/**frames**.html - 7k - Cached - More from this site

2. W3C: **Frames** in HTML Documents

Related searches on yahoo.com for *frames*.

Using a taxonomy, it would be possible to group the different readings, whereas they appear here in no visible order. Ideally, items such as *picture frames, poster frames, photo frames* should be in one group, *eyeglass frames, glasses frames* in another and in a third group the rest of the related terms.

```
if term $t$ is already in the taxonomy in a variant form $t'$ then
    if $t$ is a better representative of the concept than $t'$ then
        promote $t$ to preferred term.
        demote $t'$ term to synonym.
    else
        if $t$ fulfills the editorial guidelines to be added as a synonym then
            insert $t$ as synonym.
        else
            reject.
else
    if $t$ represents a genuine missing concept then
        determine where the missing concept should be inserted:
        $T_{hyponym} \rightarrow$ hypernym(s) of $t$.
        if $t$ fulfills the editorial guidelines to represent a concept as preferred term
        then
            insert $t$ as hypnonym to $T_hyponym$.
        else
            find a normalized representation $t''$ for $t$
            insert $t''$ as hypnonym to $T_{hyponym}$.
            if $t$ fulfills the editoral guidelines to enter the taxonomy then
                determine the relation $R$ between $t$ and $t''$ (strict synonym, related
                term, other relations):
                insert $t$ as $R$ to $t''$.
    else
        reject.
        (Note that this case happens quite often when going through actual term
        collections. Rejecting does not imply that the term is meaningless. It
        could just fall into a class that is not part of the taxonomy's design, for
        example a brand name.)
```

Procedure `Insertion of` t `into a taxonomy`

16. Paid keyword spaces

Paid keyword spaces can be regarded as the core of TE-Commerce, as this term space is restricted to terms that have a potential E-Commerce relevant reading. A paid keyword space comprises of all terms that advertisers have booked at one time. It is accumulated by SEM providers as one part of their ad database[1].

Apparently one of the most interesting term spaces for E-Commerce is build by all the terms on which companies bid in order to place advertisements. It is interesting out of two reasons: firstly, finding out what terms are deemed to be exploitable economically by companies and secondly, what are the terms that could potentially be used to generate revenues but are not used currently.

In addition, as the selection of keywords is done under the rule of economic constraints and incentives, the quality of paid term spaces is in general much higher than other that of other term spaces built up without paying. This mechanism reminds of prediction markets that tend to produce more precise predictions if the money of predictioners is at stake[2]. In the process of booking keywords and placing bids on them. advertisers predict not only what users are looking for on the Net, but also how they word their needs. Although there are still campaigns that try to accumulate traffic for a very broad range of keywords (even including non E-Commerce-relevant terms), the possibility to measure exactly the success of a campaign in terms of clicks together with the economic pressure to optimize campaigns leads to a deliberate and careful choice of booked keywords. This holds for both long tail and top frequent keyword approaches, as well as for any mixed approaches. Especially long-tail approaches try to model potential variants of keywords (including all the variation principles discussed above).

The different keyword bidding mechanism deployed in paid keyword schemes bring up the issue of term variations. On the one hand, the advertisers save money if they can sidestep keywords under fierce competition (what is commonly called bid wars) and varying terms is one way to do this. On the other hand, if the paid keyword providers are able to detect and normalize term variations, they can deliver more ads (leading to immediately higher results), have more control over their paid keywords and are able to incense bid competition. The interaction between ad triggering, ad ranking and keyword bidding mechanisms make paid keyword spaces an ideal field for studies in term handling.

Going beyond the keywords, the actual texts of the advertisements reveal additional interesting properties. Clustering operations might be performed on the target site for

[1] There are only few studies so far that make use of such data. See as one example [Carrasco/Fain/Lang/Zhukov 2003].

[2] Confer [Surowiecki 2004].

ads, on their title or on parts of the teaser.

Useful information on paid keywords contain in addition to the keyword string the name of the advertiser, the URL triggered, the bid price (maximum and average) and performance indications, such as searches, clicks or conversions. This is what the advertiser can inspect via the administration login on the SEM provider's site.

If only the keyword strings are taken into account, a flat list of terms can be aggregated over all advertisers. This list will be termed "bid log" in analogy to a query log. Its most striking properties in comparison to a search log — which will be empirically validated in the following sections — can be subsumed under the following headings:

- Edit attentiveness
- High amount of repetitiveness
- Re-construction of term variance

A dump of the complete keyword database of over 2.1 million paid keywords for the German market has been provided for research purposed from the former Espotting Media, now Miva in end of 2003. This list of keywords booked by advertisers comprise all keywords entered, regardless if they receive clicks and generate revenues. They are grouped by individual advertisers, but unfortunately there is no indication where one group starts or the other group ends. Through manual inspection, though, it is often easy to discern a group of bids belonging to one advertiser.

16.1. Bidlog properties

Edit attentiveness

Generally speaking, an entry in a bid log is much more carefully edited than a query posted to a search-engine. The time taken for deliberation is of course much longer than the average search engine user spends for formulating her informational need. It is not only the time factor that plays a role here, but also the potential monetary impact of bids, compared to freely available querying. Although bidding on weird terms that do not produce clicks would do no harm, the danger of losing money for clicks that bring no qualified traffic prevents largely that nonsense or irrelevant bids enter the system. As bids that perform too low are removed after some time, the bid repository becomes cleaned up over time. Yet even the bid repository examined here which also included non-active bids shows a much higher amount of edit attentiveness than the average query does. This edit attentiveness does not mean that all terms in the bid log are well formed, given that many advertiser tried to incorporate misspellings of their keywords.

In the bidlist containing 274.848 different word types (only counting words consisting of letters, not digits or digit-letter combinations), 130.220 words (= 47 %) are outside the CISLEX. The frequent terms of the remainder reflect named entities, as the list of the top frequent non-lexicon words reveal:

sony
gameboy
nokia
samsung
siemens
epson
brother
villas
playstation
hanedler
sega

One example of this list, *hanedler* is a misspelling, the rest are prominent brands or model names. Even after applying the spellcheck-list described above (see Part B, Orthographic Matches), 103.953 (= 38 %) of words remained as possible misspellings. As will be shown below, most of these misspellings are obviously systematically inserted to reconstruct the way users might misspell the term. Among the most frequent genuinely misspelled terms were: *doppelhaushalfte, immobilienen* and *appartment.* Many examples of misspellings in the bid log are very systematic, however — for example, concatenating words to one large word (*hotelberlin*). Although the editing of bid log data is usually down with much care and deliberation, the effects might not be visible on the level of correctly spelled terms.

High amount of repetitiveness

A bid log shows a very high amount of repetitiveness. A first-glance inspection reveals many entries that are very similar to each other and only vary in one part:

1 zimmer wohnung berlin
1 zimmer wohnung dachau
1 zimmer wohnung erding
1 zimmer wohnung fuerstenfeldbruck
1 zimmer wohnung ismaning
1 zimmer wohnung krailing
1 zimmer wohnung muenchen
1 zimmer wohnung unterhaching
1 zimmerwohnung muenchen

These entries obviously belong to one customer (the entries of one customer appear as a group, see above). This kind of repetition reflects both the re-construction of term variance (*zimmer wohnung – zimmerwohnung*) and the transformation of the advertiser's inventory into ads. Assuming that an advertiser has a list of products and services that are offered, these will often be transformed into ads by instantiating patterns. In the example above, the pattern is "1 zimmer wohnung X" with place names around Munich for instances of X.

Without data on the associations between keyword bid and bidder, it is hard to tell what amount of repetitiveness is due to one bidder that systematically instantiates the same pattern or is created successively through several bidders.

Re-construction of term variance

Given the literal matching — following a minimal preprocessing (comparable to what is discussed above in Part A, Term Spaces) — it does not surprise that advertisers incorporated a re-construction of term variance. Bidding on alternative variants may have the advantage of moving to a niche in terms of bid competition. At that time, bidding was possible for each literal string, i.e. also singulars and plurals could be bid on independently. Obviously, the term variants appearing in the bid database reflect the matching options that are available for delivering the ads. This issue will be explored in more depth below, see “Matching bids and queries”.

1 tages contactlinse
1 tages contactlinsen
1 tages contaktlinse
1 tages contaktlinsen
1 tages kontactlinse
1 tages kontactlinsen
1 tages kontaktlinse
1 tages kontaktlinsen
1 tages linse
1 tages linsen
1 tagescontactlinse
1 tagescontactlinsen
1 tagescontaktlinse
1 tagescontaktlinsen
1 tageskontactlinse
1 tageskontactlinsen
1 tageskontaktlinse
1 tageskontaktlinsen
1 tageslinse
1 tageslinsen

Even though most SEMs today offer some type of matching option that goes beyond literal strings — at least singular / plurals — there are always nichés for advertisers that try to capture traffic at a low price. It is obvious from these examples that the variations appearing hear ar e constructed — they represent what someone guesses to be search variants. Many of these keywords entered once the system, but never occurred in a query.

16.2. Matching bids and queries

One of the main benefits for SEMs that term handling can provide is an enhanced matching between bids and queries, leading to more ads being displayed for relevant queries. Other than in generic search applications, these enhancements have a direct impact on revenues. The trade-off between precision and recall can be evaluated dynamically through observing the click-rates and conversion rates for the additionally displayed ads.

Besides broadening the traffic brought to an advertiser, an enhanced matching can also higher the quality of the traffic. In many cases, advertising customers will have a limited budget and are not able to monetize additional traffic adequately. Through recognizing the intentions of a query at a more granular level, traffic can be directed to more specific offers, allowing more advertisers to occupy one field. Even if the total exposure rates for these advertisers decrease, the increased click and conversion rates that are made possible through the better targeting can make such technical improvements attractive to advertisers.

A simple way of allowing more matches to bids without losing any relevance in the bid display is to normalize singular/plurals, spacing and hyphenation (see above, Part A, "E-Commerce Term Spaces"). Through this baseline matching routine — which was implemented on the Espotting network for English in 2003 and for German in 2005 — the number of clicks on bids increased over night about 10%.

One way to achieve a enhanced semantic matching between bids and queries is to apply a taxonomy (see above, "Applying taxonomies in TE-Commerce"). In the examples below, it is demonstrated how one bid can be enriched with variants and subtypes of the WAND taxonomy in order to allow its exposure for a broader range of keywords. The existence of a matching process that determines which node of the taxonomy is best suitable for a given bid is assumed.

Car Insurance

enriched-synonym: Car Insuring
enriched-synonym: Automobile Insurance
enriched-synonym: Automobile Insuring
enriched-synonym: Automotive Insurance

Office Supplies

enriched-synonym: Office Equipment
enriched-synonym: Office Products
enriched-subtype: Office Paper Products
enriched-subtype: Overhead Transparencies
enriched-subtype: Letter Openers
enriched-subtype: Name Badges
enriched-subtype: Paper Perforators

enriched-subtype: Office Clips and Fasteners
enriched-subtype: Pencil Cases
enriched-subtype: Pencil Sharpeners
enriched-subtype: Presentation Boards
enriched-subtype: Toner
enriched-subtype: Cartridges

Plasma Screens

enriched-synonym: Plasma Display Panels
enriched-synonym: Plasma Display
enriched-synonym: Plasma TV

Dating Agencies

enriched-synonym: Date agency
enriched-synonym: Dating Services
enriched-synonym: Marriage and Dating Services
enriched-synonym: Marriage Bureau Services

International Phone Calls

enriched-synonym: Oversea Calls
enriched-synonym: International Calls
enriched-synonym: Oversea Phone Calls
enriched-synonym: International Telephone Calls

Finally, the notion of heads and containers is very suitable for dealing with bids and especially the matching between queries and bids. Through normalizing containers, for example, a bid such as *pics britney spears* can be related in a controlled way to a query such as *photos britney spears*. Besides an enhanced matching functionality between queries and bids, it is also possible to detect missing bids — bids that would match frequent queries and that are related to already existing bids.

If factoring out those queries that do not pull out bids because the query–bid matching process is suboptimal, the genuine missing bids can be subdivided into bids missing because of missing containers or because of missing heads. One sample of finding missing heads illustrates how these gaps can be detected.

Looking at the bid data in ESP, a large number of music/popstar-related bids can be observed. Most of these are of the form head – container, with the artist's name being the head and the kind of service or product that is promoted the container. If collecting a large number of such bids, the following list of containers can be extracted (see above, Part C, "Head/Container principle for search queries"):

X wallpaper
X lyrics
X songtexte
X chords
X musik
X mp3
X live
X noten
X reinhören
X texte
X tabs
X lyric
X songs
X fanpage
X cover

These containers yield the following list of heads as instantiations. The heads below are sorted by the number of container in which they appear. Note that this ranking is not necessarily concomitant with the query log frequencies, yet is even more reliable for the purpose of extracting heads.

grönemeyer
robbie williams
nena
avril lavigne
bob marley
radiohead
korn
eminem
oasis
dsds
red hot chili peppers
pink
tatu
tracy chapman
beatles
ärzte
rammstein
deutschland sucht den superstar
depeche mode
bon jovi
queen
nirvana
michael jackson

linkin park
no angels
rolling stones
christina aguilera
wolfsheim
die ärzte
rosenstolz
shakira

In this list, some amount of head variation can be observed (*die ärzte – ärzte* or *dsds –deutschland sucht den superstar*). The list of heads extracted via this container list is remarkably clean. With this process it is not only possible to detect what heads are missing in the bid list, but also how large the search volume for these bids are.

Applying the notion of heads and containers to the bid log allows to separate the concepts in a bid log from the term variants. For the SEM companies, this would enable a bidding process that is not based on individual term and their variants but rather on individual offers. The competition would then concentrate on commercially relevant concepts on different levels of specificity, from domains down to individual products, services, brands or models etc, instead of trailing off to the numerous term variations.

Determining the corona of term variations around a bid would then lie within the competence of SEMs or smaller businesses that manage online marketing campaigns. The advertisers themselves would only submit their products or service catalogs or even just their website from which their offers can be extracted.

In this line, it is noteworthy that not all terms in the bid-log are TE-Commerce terms on first glance, but also some terms that do not aim at transactions, but rather at attracting traffic to a website. For example, a bid such as *fdp darmstadt* intends to attract traffic to a particular political party's site instead of enticing transactions. However, even in such cases the performance of the keyword can be measured on a per-per-lead basis, for example in this case how many users downloaded pamphlets or signed up to a newsletter. The logic in bidding and monitoring the success of bids remains the same.

16.3. Predicting the worth of a keyword

Two issues will be addressed in this section that deals with the worth of paid keywords. Firstly, what ranking mechanisms are used for paid keywords and what bidding strategies have subsequently evolved. Secondly, what conclusions can be drawn from analyzing both the click volume and the bid heights of keywords, especially with regards to the value of keywords to SEM providers and to advertisers.

If more than one ad can be displayed for a keyword — either because several advertisers have placed their bid on this keyword or because broad matching options brought ads from related keywords into play — a decision on the ranking has to be made. Naturally, the most important feature that it is decisive for the ranking of a bid is the bid height. However, today's large SEMs use several modifications of a pure bid height approach.

The first issue that has to be taken into consideration is what happens if two bids are of the same height. The simplest decision is to rotate the available advertisements on a random base. Another approach is to prefer the longer standing bid — which of course should entice the newcomer to bid higher.

A significant change to the ranking process was introduced by Google's Adwords that took a kind of relevance feedback into account[3]. Ads that received a higher click-through rate could outrank ads with a higher bid. In its original configuration, the rank of an ad was calculated by multiplying click-through and bid height. Ads with a high perceived relevance are thus boosted, while advertiser that are willing to spend a more for a click have a chance to hold their rank.

As of today, Google's Adwords program does not disclose fully what ranking factors come into play when determining the rank of an ad. Factors that are mentioned are the global performance of a keyword (average clickthrough rate for all advertisers), relevancy of the ad teaser text and quality of the target page[4].

MSN's Adcenter is very similar in this respect to Google's Adwords, with the exception of additional target bids. Incorporating more targets (socio-demographic defined groups, based on the preferences of these for individual MSN channels) requires to place an extra bid. These extra bids also boost the rank in the MSN Live search results page[5]

For both SEMs, it is not clear how the different matching options come into play. It seems conceivable that first a pool of potential ads for a query are retrieved and the ranking then works as if these ads were all pulled out by the default matching option. An ad that came up by broad match but performs well in terms of click-through for this keyword may thereby outrank an exact match ad. Through monitoring the ad's performance and occasionally rotating ads, so that additional ads can be tested for performance, it is possible to let an evolutionary process decide what ads to keep.

Yahoo currently uses a simpler, but more transparent scheme. It groups ads by the different matching options. First the "Direct Match" ads, ranked by their bid height, then the "Advanced Match" ads, also ranked by their bid heights. [6].

A feature that all SEMs currently deploy is auto-bidding. Auto-bidding allows to set a maximum cost-per-click, but the real cost-per-click will be adapted so that it is just high enough to maintain the rank of the ad.

Consider for example the following list of advertisers and their maximum bids:

- Advertiser's 1 Maximum Bid: 4.79

[3] This is of course a ranking based by expected revenues for the SEM. See also [Aggarwal/Goel/Motwani 2006] for a recent discussion of bidding models.

[4] `https://adwords.google.com/support/bin/answer.py?answer=6111&ctx=sibling [Nov. 1, 2006].`

[5] `http://advertising.msn.com/Home/Article.aspx?pageid=50?pageid=707&articleid=3200 [Nov. 1].`

[6] `http://searchmarketing.yahoo.com/de_DE/rc/srch/dtcfaq_mt.php [Nov. 1, 2006].`

- Advertiser's 2 Maximum Bid: 4.37
- Advertiser's 3 Maximum Bid: 4.36
- Advertiser's 4 Maximum Bid: 2.94

Advertiser number 3 benefits from auto bid, as the actual bid price for this advertiser is 2.95. Without auto-bidding, the full price of 4.36 would have to be paid. However, a side effect of autobidding has a detrimental effect for advertiser number 2, as the bid is driven up by number 3[7].

There is even software available that looks for such bid gaps and sets the auto bid for the second competitor just one cent below the maximum bid[8]. This makes the highest bidder still have to pay the full bid if using auto-bidding (because the next auto-bid is just one cent below). Effectively, however, the bidder in second position only has to pay one cent more than the bid in third position, but poses a virtual price to be beaten by the bidder above. This technique is apparently widely used

To determine the worth of a keyword to the SEM provider, one can multiply the number of clicks on a keyword with the average bid height for it. A more precise calculation takes the rank of the keyword into account and multiplies the number of clicks per rank with the bid height for this rank. Based on the first approach, the top valued keywords for Espotting in 2004 have been the following

reisen
lastminute
kredit
private krankenversicherung
fluege
handy
daten wiederherstellung
weiterbildung
sportwette
hotels
shopping
mietwagen
billigfluege
erotik
internet provider
kredite
abnehmen

[7] Confer also Stan Hauser, "The Pros and Cons of Bid Gaps", online at `http://www.gotlinks.com/earticles/articles/107171-the-pros-and-cons-of-_bid-gaps_.html` [Nov. 1, 2006].

[8] KeywordMax (`www.keywordmax.com`) and AtlasOnePoint (`www.atlasonepoint.com`) are the most well-known.

lastminute urlaub
lebensversicherung
urlaub
hotel
baufinanzierung
finanzierung

Looking at these examples, there seems two be two different ways how a term enters the top list. One part of these examples are very general searches from a limited number of domains, for example travel (*reisen, lastminute, fluege, hotels*). The other part are very specific keywords that have in all probability a high conversion rate and/or a high revenue per conversion for advertisers.

From the perspective of advertisers, the value of a keyword is the product of (average) conversion rate and (average) revenue per conversion, assuming a ceteris paribus situation, i.e. the capability to monetize all conversions equally. This product is the maximum price that an advertiser can pay if the campaign should immediately pay-off[9]. If looking at the actual list of highest bids from Espotting in 2003, the following terms appear (bid height in EUR-cent in paranthesis):

bueroservice 608
bueroservice koeln 526
bueroservice muenchen 526
telefonservice 478
bueroservice duesseldorf 430
bueroservice dortmund 385
tagesbuero 366
daten wiederherstellung 353
datenwiederherstellung 262
festplatten datenrettung 227
pkv vergleich 210
ratenkredite 201
business center 187
pkv private krankenversicherung 185
businesscenter 176
disk recovery 174
kranwaage 174
private krankenversicherung vergleich 173
berufsunfaehigkeitsversicherungen 171
sterntaufe 170

Out of the 10.000 terms that generate the highest revenues, only 3.473 are found in the list of the top 10.000 terms sorted by bid height. The rest of term that generated high revenues made it into the list through high click volume instead of high bid price,

[9] In some cases, advertisers will be willing to pay a higher price, for example to promote a new brand.

whereas the rest of the terms highest by bid prices were not clicked enough times to generate enough revenues for the inclusion in the top 10.000[10].

The highest bids that were not found in the list of strongest in revenue together are the following:

bueroservice duesseldorf
businesscenter
corrupt
pkv info
pkw kredit
altersversorgung
beschleunigungssensor
privat krankenversicherung tarif
kapitalbildende lebensversicherung
buero muenchen
buero koeln
buero duesseldorf
buero dortmund
sensoren
risikolebensversicherung vergleich

None of these appear higher than the 300.000th rank in YG. It is presumable that these terms have very high conversion rates and/or revenues per conversion. Looking at the content of these extremely high bids, as they relate to services that are obviously above the scales of pennies and dimes. This list suggests that cases of extreme bid wars should be restricted to terms that are not very often searched, making them not as valuable to SEMs than the impressive bid height numbers might suggest[11].

From the perspective of the SEM, it is especially the terms that have both a high click rate and a high average bid height that are most interesting, although the first factor is more important in this equation. This can be seen from comparing the list of top clicked with the list of top revenue-generating terms: More than 6.500 of the 10.000 top revenue-generating terms are also in the list of top 10.000 terms by clicks. This is almost double the number than the intersection between the top 10.000 revenue-generating terms with top 10.000 terms by bid height. Given that 2.315 terms appear in both intersections, i.e. are in the top 10.000 clicked terms, in the top 10.000 terms by bid height and in the top 10.000 terms by revenue, it can be seen that little more than 1.000 terms are very strong in generating revenues although they are not among the top clicked terms. These terms are in the majority either very specific services

[10] The cut-off threshold at 10.000 is of course arbitrary. Fur the purpose of qualitatively estimating what the most valuable keywords are from the perspective of SEMs, this metrics suffices, though.

[11] These numbers are very little compared to cases of bid war on Google AdWords, with bid heights going up to 40$ per click for specific topics, such as combinations of lawyer services plus geographical name. See `http://forums.searchenginewatch.com/showthread.php?p=56206 [Nov. 1, 2006]`.

(such as *sterntaufe* or *herzluftballon*) or belong to domains that have high average revenues per conversion, such as tourism (more than 30% of terms) or insurances.

Given that more specific keywords should generate higher conversion rates, but are less searched for, a decisive factor in boosting the value of a keyword is the average revenue per conversion. This rate is mostly characteristic for the domain of the keyword.

16.4. A database of campaigns

16.4.1. Scraping ads

In a first step, a corpus of campaign data, consisting of the keyword that triggered ads along with the title, the copytext and the URL of the advertiser was collected through querying Google.com with a high latency time and scraping the advertisments. For the set of 193.195 high frequent queries generating the highest revenue this resulted in a list of 792,009 ads. The number of web document hits that the query produced was also stored. An entry in this campaign database consists of a tupel < *web document hits, query log frequency, query, ad rank, ad title, ad teaser, displayed target URL*>.

In this excerpt of the database, the ad rank is omitted. Searches that did not produce any ads are marked with NOBID

- 220000000 – 1073826 – handy – Handy – hier bis zu 82% billiger! Es geht immer - billiger.de – www.billiger.de/Handys_ohne_Vertrag
- 220000000 – 1073826 – handy – Prepaid günstig bei Blau – Nur 16 Cent/min. 11 Cent/SMS Mit Blau.de bis zu 80% sparen – www.blau.de
- 181000000 – 1008430 – poker – NOBID
- 417000000 – 930611 – last minute – Last Minute-Angebote Superspar-Angebote von Expedia.de: Hier sicher buchen und wegfliegen. – www.expedia.de/lastminute
- 417000000 – 930611 – last minute – 100% Last Minute – Lastminute direkt vom Flughafen Verkauf von Restpläzen und Stornos – www.ferienknaller.de
- 417000000 – 930611 – last minute – billig-buchen.de – Urlaub buchen - rund um die Uhr. Aktuelle, supergünstige Angebote! – www.billig-buchen.de

Additional potential helpful information would comprise the bid price and detailed clickthrough information on paid keywords, i.e. which ad was clicked how often starting from which query.

The clustering operations presented in the subsequent section are based on displayed target URLs instead of advertisers, so URLs belonging to the same company build up separate groups. Bid prices on a large scale would help to spot keywords under high competition; an approximation of this can be done via the number of different URLs appearing for a keyword. The gravest shortcomings of the data available are the missing clickthrough indications; here, one has to rely on the natural selection of campaigns, a "survival of the most fitting (or most paying)"-principle that should in general apply to paid keyword campaigns. While this should work for the live campaigns appearing on the result screen of Search Engines, as Google for example calculates the ranking of ads also based on clickthrough data, it remains a desiderate for the flat list of keywords. Here, the only possibility is intersecting the bid list with query log data in order to find out what bids are frequently searched for and what bids never appeared in users' searches over a given period of time.

16.4.2. Clustering the campaign database

Re-organizing the database of online advertisements texts, URLs and triggering keywords leads to interesting views. One sorting option is to cluster identical URLs and texts together, which results in a list of synonymous or quasi-synonymous keywords chosen for triggering this ad. Another sorting that allows new insights into the campaign database is achieved by listing the different URLs for identical keywords. This results in the list of competitors on a field of business, for example here competitors on *computerzubehör*:

- www.bueroversandhandel-jena.de
- www.mediaonline.de
- www.kmelektronik.de
- www.amazon.de
- www.jacob-elektronik.de
- www.atado.de
- www.electronicscout24.de
- www.otto.de

Listing the URLs of advertisers by the number of keywords for which they display ads results in the list of the biggest long-tail advertisers. Some of these are domain-specific (for example real estate-related), but harvest many terms of this domain, making them appear in this list that has the number of keywords for which the URL appeared on the left:

- 37333 www.ebay.de
- 12216 www.shopping.com
- 11090 www.Preisvergleich.de
- 9882 eBay.de

- 8281 www.de.eBay.com
- 8016 www.onlineangebote24.de
- 7434 www.amazon.de/buecher
- 4700 www.shopping-de247.de
- 3629 www.immonet.de
- 3621 www.Preisvergleich.de/
- 3349 www.quelle.de
- 3134 www.wer-liefert-was.de
- 2746 www.otto.de
- 2691 www.bookings.de
- 2518 www.jekoo.com
- 2071 www.neckermann.de
- 2040 www.mercateo.com
- 1959 www.TripAdvisor.de
- 1701 bounce.deal-market.com

Looking at the associations between search term and the title of the ad is a method to extract the potential E-Commerce reading that a term can trigger — for example *hotel in amsterdam* for the query *amsterdam*. Comparing keyword and title also allows to discriminate cases where the offer is more granular than the search term and vice versa. Finally, clustering by the ad title yields a list of related searches, as can be seen below:

Hotels in Amsterdam appears for the following queries:

amsterdam billig hotel (more specific)
amsterdam (more general)
amsterdam pensionen
amsterdam pension
amsterdam hotels
zimmer amsterdam
unterkuenfte amsterdam
hotels amsterdam
amsterdam hotels
amsterdam hotel

One result of the scraping of paid listings is the discovery of ad gaps, i.e. frequently searched keywords that produce no ad at all. The overwhelming majority (in a sample test of 500 out of the 47.254 lines, more than 85%) of these gaps are due to policy reasons and to keywords representing named entities. As discussed above (Part A, TE-Commerce Interactions), several types of ads are not allowed by Google, including hardcore pornography and gambling. Bidding on a proper name is either prohibited by Google's Adwords policy (if the advertiser is someone else) or does not make much sense if the potential advertiser is listed anyway on top position. A more interesting type of ad gap is presented by geographical names. While large cities usually trigger ads — especially for hotels, flights or real estate — many smaller towns do not. There are some real-estate advertisers that try a long tail strategy on smaller locations, but they do not appear for the majority of smaller locations. Presumably, the reason lies in too low conversion rates for these ad, leading to a removal by Google. A further group of terms that do not produce any ads are terms that have no (or just a too detached) E-Commerce relevant reading, for example *Postleitzahlen*. Given that the ads undergo a high fluctuation, it does not make sense to list the few ad gaps that do not fall under one of the above mentioned categories. It is remarkable, though, how complete the field of legitimate searches is covered by Google's Adwords. Moreover, it is striking that the overwhelming majority of inoffensive searches are used to trigger ads and have thus a potential E-Commerce reading. The E-Commerce-relevant part of Web search — which is exemplified through the bid space — is still largely underestimated.

17. Improving a YP system with vocabulary enhancements

Considering that the majority of businesses are not accessible on the Web (see above, Part A, TE-Commerce Agents), it becomes obvious that YP (YP) and business directories are crucial devices in TE-Commerces.

In contrast to their importance, however, these sites suffer in general from a very high number of failed searches, sometimes up to 40-45% of all search occurrences[1].

The reasons for this enormously high rate seem threefold. Firstly, the indexed data fields are much more restricted on business directories than for example on general Web Search Engines. Usually, only categories, business names and a limited number of additional pointer to categories are available (see below for alternative search logics on business directories). Secondly, combining more than one form field ("Who?", "What?" and "Where?" field), might drastically decrease the number of results if a consequent AND-operator functions between the fields and no vicinity-based (for the "Where?" field) or similarity-based (for the "Whot?" and "What?" field) matching is in place. Thirdly, users are accustomed to the breadth of searches that are answered on the Web, do not necessarily think of YP categories when formulating their needs, but rather type in products, services, vendor types, problems or very general searches (see above, Part B, E-Commerce Term Management).

17.1. Evaluation of current YP search systems

Before a standardized test battery is going to be tested on several German, UK and US business directories, different search logics have to be explored. As these directly affect the users' possibilities to interact with the site, there is no single evaluation approach suitable for all sites, although the underlying test principle remains the same. Testing is conducted by sending pairs of related high frequent queries to the YP site and monitor if these searches produce any hits at all, if they product relevant results and finally, if both searches produce consistent results.

17.1.1. Search logic on directory sites

The standard business directory logic contains an index of business addresses with category information, optional also specialties for the businesses. In general, busi-

[1] Query Log analysis performed by the author on queries of several major German YP and directory assistant sites.

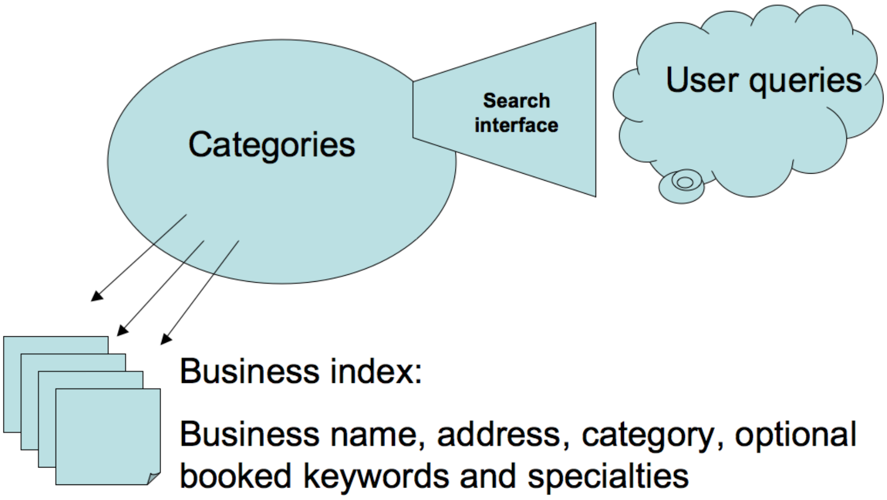
nesses have to pay for adding specialties, for example by upgrading their listing with highlighting and additional keywords.

From these index data follow the three typical form fields: a "What?" search on the category information, a "Who?" search on the business names and a "Where?" search on their whereabouts. On the interface, the "What?" and "Who?" form fields are often merged into one.

As a navigational aid, a hierarchy of categories is often available that allows to enter a drill-down starting from broad categories. At the end nodes of this hierarchy are the same categories that also appear in the result listings.

The categories that get displayed are not necessarily the same than with which the business listings are indexed. It is a common practice that the headings which appear on the Web are broader and further normalized than the categories used for indexing businesses which might stem from different sources are not necessarily edited to produce a consistent spell-out.

A typical result page on a YP site contains top listings that are graphically highlighted from the rest of the listings. These top listings contain are not only highlighted, but also contain more content, for example additional specialties of a business. A standard entry contains only contact information such as postal address and telephone number. This restricts the search space of the "What"-field to categories, pre-defined aliases pointing to these categories and advertisers' keywords:


Categories
Search interface
User queries
Business index:
Business name, address, category, optional booked keywords and specialties

Conventional search logic on YP site

Some business directories have index data that is rather based on keywords per company instead of categories, usually through self-registration of businesses or by harvesting printed content[2]. For these, the search mechanism have to provide an access to keywords and ideally also to keyword variants. The advantage of this model is that it allows more granular searches. However, if no category information and no

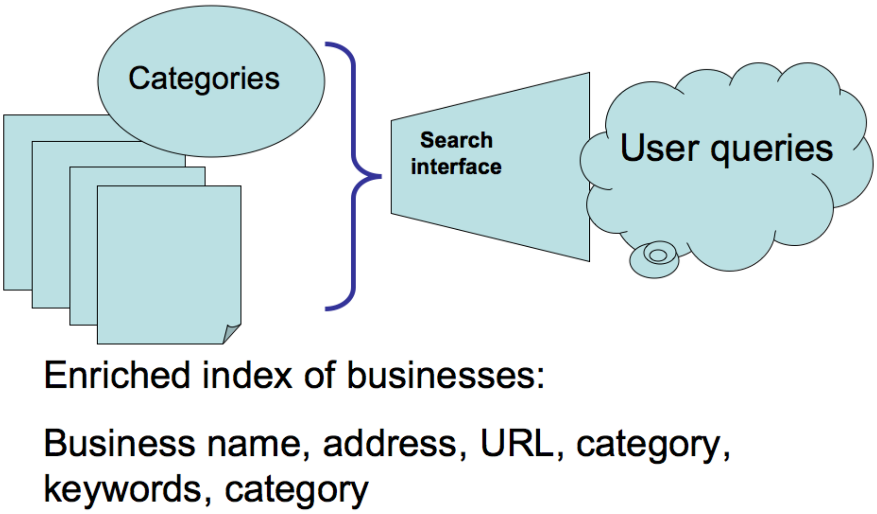
[2] For example, `www.scoot.co.uk`.

weighting of keywords are available it is not possible to discriminate between additional specialties of a business types and core line of businesses. For example, a hotel might list *restaurant* as one of its specialties, but should certainly be discriminated from restaurants proper.

In many cases, a mixed logic combining categories with additional keywords is used[3]. A discriminative feature of these integrative approaches is whether the keywords are controlled by the categories. Conventionally, if choosing a category or entering a query that can be matched to a category, a frequency counting of keywords is delivered back, as on `www.superpages.com`. Alternatively, if choosing a keyword, a frequency counting of categories is delivered back — yet without any control of what keywords can be associated to what categories. In both scenarios, one faces usually some amount of noise that is either due to imprecise categorization of businesses as well as due to irrelevant assignments of keywords to a company.

A possible way to overcome this noise — at least to a certain degree — is to extract keywords from the websites of companies. With this approach, it is possible to provide finer granulated information that YP categories to the user that at the same time can be immediately experienced as being valid descriptors, as they appear on the site of the business once the user clicked on it. These keywords with a very high perceived relevance will not be available for all businesses, though. As was laid out above (see Part A, "TE-Commerce Agents") only a minority of businesses have a homepage that can be found with reasonable resources on the Web and out of these, only about 80-90% have enough content to provide good keywords, especially if the keywords are restricted by the category of the company[4].

In a schematic summary, an integrative search logic that combines categories and additional extracted keywords for a part of the businesses would look like this:


Categories
Search interface
User queries
Enriched index of businesses:
Business name, address, URL, category, keywords, category

Search logic with categories and additional keywords on YP sites

[3] For example, `www.superpages.com`.

[4] Results from assigning a database of 22 million unique keywords to 500.000 UK business homepages.

The interaction cycle on YP search agents is different to that on Web search engines. Usually, YP search agents offer a link to broaden up the search or even automatically broaden the user's query in the case that few or zero results have been returned. This broadening up of searches uses in general a relaxed geographical constraint, less commonly also a moving-up to a hierarchically superordinate caetegory. For example, if *women's hairdresser / Cardiff* fails, possible modified queries could either have a broader geographical scope (*women's hairdresser / Wales*) or entail a broader category (*hairdresser Cardiff*). Different strategies for broadening the geographical scope are feasible, for example a radius search or moving to a larger administrative unit.

If a search that leads to few or zero results is not recognized so that it cannot undergo the process of broadening up its scope, an approximate matching can take place that tries to relate the literal strings of the input to known units. In many cases (see above, Part A, Term-driven E-Commerce), the suggestions from the approximate matching module are calculated based on a predefined list of keyword or category descriptors. It is not guaranteed that these suggestions will produce any results for the given location.

A typical YP categorization system numbers from few thousand terms up to maximal 20-30k categories. Usually, only few thousand terms are used for the display or navigation on the site, given that the screen size and the maximum acceptable hierarchy depth determine an upper limit to the number of categories displayed. Without deploying a considerable number of aliases and using an approximate matching that can work in real-time on the query, the huge discrepancy between the limited and closed term spaces of categories and the unlimited and open Term Space building up through the free text search will prevent the pleasant effect users know from search engines — no matter how weird the search is, there are almost in all cases results for it on the Web. Considering that also on YP sites more than half of query strings occur only once, regardless of the time span examined, it becomes clear that there is no upper bound of individual queries in sight when users type in their searches day by day.

In many cases, the user is not presented with the internally used categories of the business index, but with a simplification of it[5]. The reason for this is twofold: The full set of categories might contain category descriptors that are redundant or nor normalized enough to be displayed on the Web — on the other hand, displaying broader categories can be considered as a means to sell paid keywords for businesses that want to distinguish themselves from others. This consideration clearly clashes with the intentions of delivering more relevant results through vocabulary enhancements. Ultimately, however, directing highly qualified traffic will open up new possibilities for monetizing this service — the fields of high competition between advertiser will just be of a more granular kind, for example centering on keywords such as *hotels with golf course* instead of the general *hotels*. The final answer to this question should lie in more precise remuneration mechanisms for the YP operator that are not based on

[5] `www.11880.com` for example does currently not display the 8.000 datagate categories, but a summarization of these.

exposure, but rather on successful contacts enabled through the YP site. Such pay-per-lead models will create the highest revenues if as much information of the users' queries and the businesses' offers are recognized and matched together. Through this, the full TE-Commerce potential of YP search is unleashed and term improvements will be a central asset of YP operators and create immediate benefits to them.

17.1.2. A cascaded test battery

In YP search, the classic IR evaluation metrics of precision and recall are hard to apply and do not necessarily tell much about the performance of a search agent. The search results can in general only be inspected at a shallow level at the site of the YP search, often just in the form of categories. For example, if searching for *samsung X05* and being presented with hits from a *Computer & Notebooks* category, there is no way to tell from the presented information (category, company name, address) whether these vendors really have *samsung X05* have in stock.

It is therefore only possible to test whether the presented results fit the user's query, not if they represent businesses that can fulfill the underlying needs. If a business is also listed with its homepage, it is conceivable to check what specific offers it makes, yet this hard to transform into an operational testing suite. Focussing on the displayed results instead entails counting the number of hits, keeping track at the categories of these hits and of navigational aids, for example display of related categories.

A cascaded test battery that builds up on these three components — number of hits, categories of hits and additional navigational aids — evaluates strengths and weaknesses of YP sites without having to perform redundant tests. Just as in an oral examination, the test battery starts with rather simple tasks - such as singular/plural variations - and then moves forward to more demanding recognition and normalization problems such as morphologic variations and synonyms. While the performance on one level does not necessarily predict the performance on another, it becomes clear after a few tests how well a system performs on one level. It is not necessary to test hundreds of singular/plural variations when it becomes clear ager having tested a few that the system in not capable of recognizing inflectional morphology.

In addition to this, a higher level in the cascade represents a functionality that needs the functionality from lower levels to unleash its full potential. For example, capturing synonyms has a much greater impact if handling inflection also takes place. The layers are therefore ordered along a line starting with basic challenges and moving on to sophisticated challenges.

Layer 1: The basics (singular/plural, spacing, dashes etc)

In the context of YP search, singular and plural variation are almost in all cases equivalent in meaning. Whether users type in *attorneys* (presumably, because they want to see a list of attorneys) or *attorney* (because the one individual attorney chosen from the list is sufficient), the criteria for relevant results do not change[6].

Spacing and hyphenation variation — apart from the cases already presented above where these variants bear different meanings (Part A, Term Spaces) — need also to be treated equally.

The query pairs listed below thus be should all produce the same results on YP sites:

- video conferencing – videoconferencing
- dvd recorder – dvd-recorder
- back packs – backpacks
- textbooks – text books
- night clubs – nightclubs
- note book – notebook
- pharmacy – pharmacies

Layer 2: Orthographic variations and typos

At least the most frequent misspellings should be recognized by YPsearches. Examples of such misspellings are *accomodation* vs. *accommodation*, *labtop* vs *laptop* and *picnic* vs. *picknic*. These misspellings can either be solved by a dynamic approximate matching or by a look-up in a pre-calculated list of the most common misspellings.

If YP search agents have a index full with many keywords in it, they are usually able to produce at least some matches even through simple literal matching, because some businesses' keywords will also be misspelled. However, in this case a large discrepancy between the number of results for the correct spelling and for the typographical error can be observed.

The query pairs listed below thus be should all produce the same results — or at least lead to a "did you mean" link for the misspelled variant — on YP sites:

- accommodation – accomodation
- picnic — picknick
- laptop – labtop
- acupuncture – accupuncture
- physicians – physicans

[6] Arguably, a part of users typing in attorneys might look for law firms with more than one attorney. However, even though the singular form for vendor types is usually much higher in YP query logs than the plural form, there are still so many searches for the plural form that it can be ruled out that all these users search explicitly for law firms with more than one attorney — especially by comparing the frequency of singular and plural of *attorney* with the much lower frequencies of synonyms to *attorney*.

Layer 3: Morphology and Syntax

The query pairs listed below reflect morphological and syntactical variations sharing the same user intention (see above Part B, “Morphological-syntactical variations”). They are built up from variants that appear a similar number of times in YP:

- clinics medical – medical clinics
- car rentals – rental cars
- spa hotel – hotel with spa
- sports cars – sport cars
- aromatherapists – aromatherapy
- wedding planning – wedding planner
- file cabinet – filing cabinet
- translation services – translators
- nurses – nursing
- moving services – movers
- acupuncturists – acupuncture

Layer 4: Synonyms

The following list of synonyms tries to restrict itself on almost universally accepted and widely used synonyms or quasi-synonym. In the context of YP search the difference between synonyms and quasi-synonyms can be largely neglected, given that a business that is described only by a category should in most cases be equally describable by synonyms or quasi-synonym to this category.

- automotive insurance – car insurance
- automotive rentals – rental cars
- bicycles – bikes
- billiard tables – pool tables
- cellular phones – cellulars
- chairs – seats
- computer components – computer parts
- eyeglasses – spectacles

- laptop – notebook computers
- recreational vehicles – rvs
- recreational vehicles – campers
- refrigerators – fridges
- running footwear – running shoes
- sofas – couches
- swimwear – bathing suits
- swimwear – swim suits
- web site design – web site creation
- web site design – web site development
- soda – pop
- automotive detailing – automobile detailing
- indian cuisine – indian restaurant
- doctors – physicians
- attorney – lawyer
- wifi network – wireless network

Ambiguous searches

Numerous frequently searched single word terms are ambiguous in meaning, even if the possible meanings are restricted to the commercial domain. What is conventionally listed as ambiguous words (either a polyseme or a homonym) is in many cases resolvable under the premise that the meaning has to be commercially relevant. The selection of meanings might even be further restricted through the topical context of the search. For example, a user typing in *apple*, *jaguar* or *tiger* in the search box of an Apple online shop does not enter an ambiguous search. Likewise, searching for *matrix* on a mathematic bibliographic site does not leave room for interpreting the query as the movie title[7].

In the domain of YP search, genuine ambiguity arises when the different meanings reflect a commercially relevant concept. Examples are provided in the following list:

- Speaker (professional orator vs. audio equipment)

[7] All these examples have been taken from [Zeng/He/Chen/Ma 2004]. For general Web searches, these are all valid ambiguous words.

- Printer (professional in the printing business vs. computer peripheral)
- Nails (building material vs. manicurists)
- Foundation (philantrophic organization vs. foundation of a building)
- Plane (aircraft vs. woodworking tool)
- Frames (picture frame vs. eyeglasses' frames)
- Notebook (portable computer vs. stationery)

A systematic source of ambiguous YP searches is the suffixation with *-er* that produces agent nouns for which the agens can be either a human (speaker = professional orator) or a device (speaker = audio equipment).

Lastly, very broad searches (for example *shopping* or *computer*) have to produce a broad group of categories matching to this broad search. If the selection of categories is only based on counting frequency of keywords in the result set, it often happens that irrelevant categories sneak into this selection. For example, searching for *hotel* on `www.herold.at` also produces hits from the following categories: *Espresso- u Kaffeemaschinen*, emphGastronomiemaschinen or *Wurstschneidemaschinen* [8]. Businesses listed in one of these categories also listed *hotel* as a keyword. Apparently, going from a keyword appearing in actual listings to the categories of these listings does not always yield valid results, as the keyword might not reflect the main line of business.

A further case where multiple categories needs to be displayed is presented by different vendor types for a product. For example, *Italian food* may relate to Italian groceries and Italian restaurants alike.

Zero hit test

Starting with the most frequent YP search as derived from combining several YP query logs, it is tested what searches produce zero hits. These are obvious cases for vocabulary enhancements, as virtually all these searches can be mapped to one or more of the YP site's categories.

Through the intersection of several YP query logs it is possible to remove peculiarities and artifacts of individual logs. The top thousand queries from this intersection were tested using either the largest city in the country as entry for the "Where"-field or no entry at all, if the YP search allows to leave the "Where"-field blank.

While an empty result set might be a valid answer if the user searched for a very specific business in a small town (even then a geographical broadening of the search should take place), it is definitely a shortcoming if the user searched for a popular category through a frequently occurring query. Moreover, the search interface should indicate to the user whether the requested type of business was recognized, but the requested businesses not present in the index, or whether it was not recognized at all.

[8] As seen on Nov. 1, 2006.

17.1.3. Criteria

From the perspective of the user, judging the performance of a query on a YP portal is an integrative evaluation that combines how well a query is recognized as well as how the results are ranked and presented. As the quantity and quality of results does not only depend on factors that can be improved via enhancements to the search process, but is also affected by the available business listings, the testing criteria presented here can only be used as relative metrics. Their main application is a micro-evaluation on how consistently related queries are handled on one site. Even without knowing anything about the size of the business index of a site or its performance on other search terms, it is obvious that disparate results for *notebook* and *notebooks* represent an issue that calls for improvement.

A first metric is based on the number of results for a query. An obvious case that calls for improvement are common queries that produce no hits. As the tests are conducted without any geographical filter or — if that is not possible — by choosing a large city. Therefore, it can be safely ruled out that there is really no appropriate business matching the search. In detail, it can be shown that all of the common queries producing no hits could be related to a category with results in it.

In addition to the number of hits, the intersection of hits for the pair of similar searches is counted. Intersection can be measured both for a selection of categories or for listings of businesses. The percentage of intersection is calculated by dividing the number of those results produced by both searches through the number of all results. If search query A produces five results and search query B produces three results out of which two have been in the result set for A, than the percentage of intersection is $2/(5 + 1) = 33\%$. Given that the results on YP sites are typically a combination of paid listings and index listings, a small amount of discrepancies can be due to different numbers of paying advertisers and not originate in failures of query variance detection.

Finally, through listing the categories produced by the searches in the test results it is possible to spot irrelevantly answered searches. If the results are business listings without category indication, such an evaluation would require to know what these businesses offer which could only be provided by URLs or in a limited number of cases the business name (for example, a business named *Sandy's Beauty Salon* is certainly not a good result for a search for *notebooks*, see also Part B, "E-Commerce Term Management").

17.1.4. A sample analysis based on the test battery

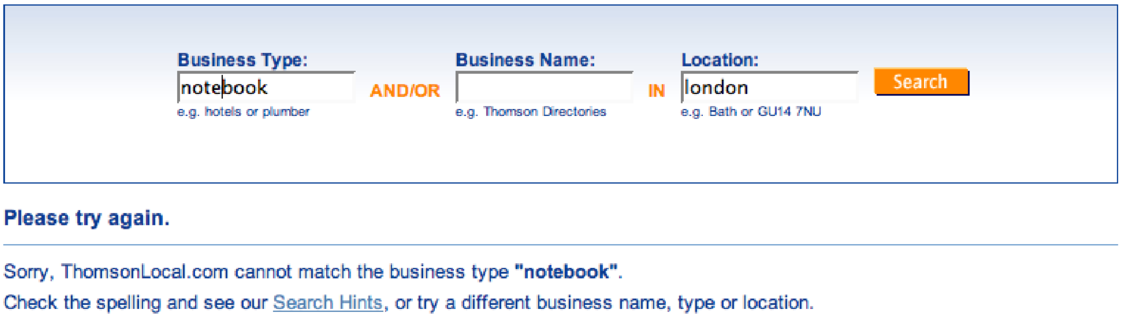
As one example of the several tests that have been conducted on German, US and UK YP sites and demonstrated that all of these are in need to some degree of vocabulary enhancements, a sample analysis of the site `www.thomsonlocal.co.uk` is provided below.

All queries have been issued with the location "London" (selecting "All areas"). The test search terms comprised of 100 pairs of orthographically, morphologically or semantically related searches. Ideally, these searches should

1. produce results at all
2. produce relevant results
3. produce similar results for similar searches, including singular / plural search variations (*restaurant-restaurants*), morphological variations (*travel agency-travel agent*) and synonyms (*notebook-laptop*)
4. be resolved to their different intentions if they use ambiguous words (*printer profession* and *computer peripheral*)

Zero hits

Many search words, even very popular ones (*notebook, notebooks, attorney, automobiles, eyeglasses, swimwear, nurses*) do not get any results. Both *notebook* and *notebooks* written as one word do not get any results. *Note book* written in two words does not get any computer-related results. *Note book computers* only gets one computer-related category (*Recycling & Disposal Computers*). On `www.thomsonlocal.co.uk`, as on many other YP sites, the result screen for a failed search does not provide more than general suggestions for the user:



Business Type: notebook

e.g. hotels or plumber

AND/OR

Business Name:

e.g. Thomson Directories

IN

Location: london

e.g. Bath or GU14 7NU

Please try again.

Sorry, ThomsonLocal.com cannot match the business type **"notebook"**.

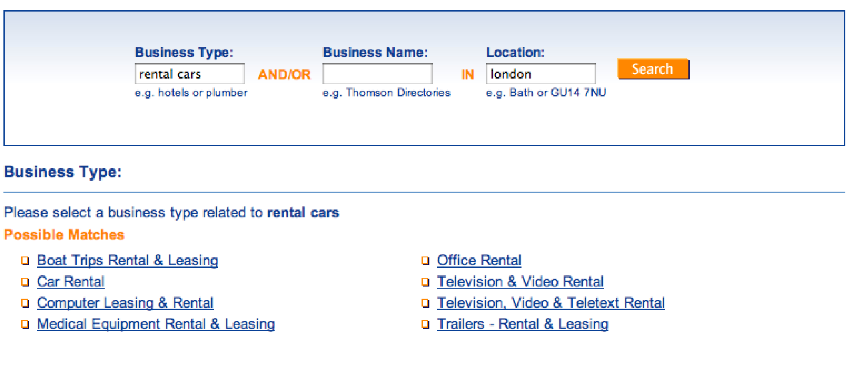
Check the spelling and see our Search Hints, or try a different business name, type or location.

A failed search on `www.thomsonlocal.co.uk`.

From the perspective of user interaction, a zero hit result is basically a dead-end. The user cannot recognize whether her request was understood, but could not be answered by a listing, or not understood at all. The only way to proceed is to start anew. From examinations of YP query logs, it can be seen that about 20% of users that experienced a failed search do not try to reformulate the search (see below, "Analysis of Failed Searches on YP sites").

Irrelevant category selection

Many search words do not lead directly to company listings, but offer a selection of categories to the user. While this mechanism is clearly helpful if the categories match to the search query, irrelevant categories will distract the user. Below is one example of irrelevant category selection for the search phrase *rental cars*:



Business Type: rental cars AND/OR Business Name: IN Location: london Search

e.g. hotels or plumber e.g. Thomson Directories e.g. Bath or GU14 7NU

Business Type:

Please select a business type related to **rental cars**

Possible Matches

- Boat Trips Rental & Leasing
- Car Rental
- Computer Leasing & Rental
- Medical Equipment Rental & Leasing
- Office Rental
- Television & Video Rental
- Television, Video & Teletext Rental
- Trailers - Rental & Leasing

Irrelevant selection of categories on `www.thomsonlocal.co.uk`.

While *Car Rental* is the correct category, the other "Possible Matches" are not related to the search term, apart from the occurrence of *Rental* in both of them.

Inconsistent results for similar searches

The singular/plural variation *book* *books* has a great impact on the search results: While book leads to a selection of seven categories, books let the user select from only four categories of which only two had been in the previous set of seven categories.

Similar inconsistencies in the selection of categories occur for many singular/plural variations (for example *printer – printers, note book – note books, lawyer – lawyers, pharmacy – pharmacies*).

The same holds for many synonymous searches (*attorney – lawyer, notebook – laptop, swimwear – swim suits, doctors – physicians, aircraft – planes*), producing disparate or even totally disjunctive results.

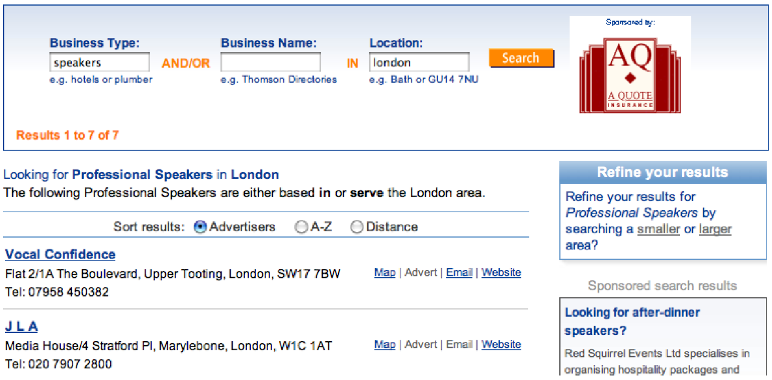
Disparate results comprise the following two scenarios:

- one query produces results (either a category selection or company listings), its synonymous variant gets zero hits

- both queries produce results (either a category selection or company listings), yet there is little or no overlap of results

Resolution of ambiguous keywords

The first example, *speakers*, is only resolved in one direction. The meaning "speakers (loudspeakers)" is not recognized by the displayed suggestions:



Business Type: speakers AND/OR Business Name: IN Location: london Search

e.g. hotels or plumber — e.g. Thomson Directories — e.g. Bath or GU14 7NU

Sponsored by: AQ A QUOTE INSURANCE

Results 1 to 7 of 7

Looking for **Professional Speakers** in **London**

The following Professional Speakers are either based **in** or **serve** the London area.

Sort results: ● Advertisers ○ A-Z ○ Distance

Vocal Confidence
Flat 2/1A The Boulevard, Upper Tooting, London, SW17 7BW
Tel: 07958 450382
Map | Advert | Email | Website

J L A
Media House/4 Stratford Pl, Marylebone, London, W1C 1AT
Tel: 020 7907 2800
Map | Advert | Email | Website

Refine your results

Refine your results for *Professional Speakers* by searching a smaller or larger area?

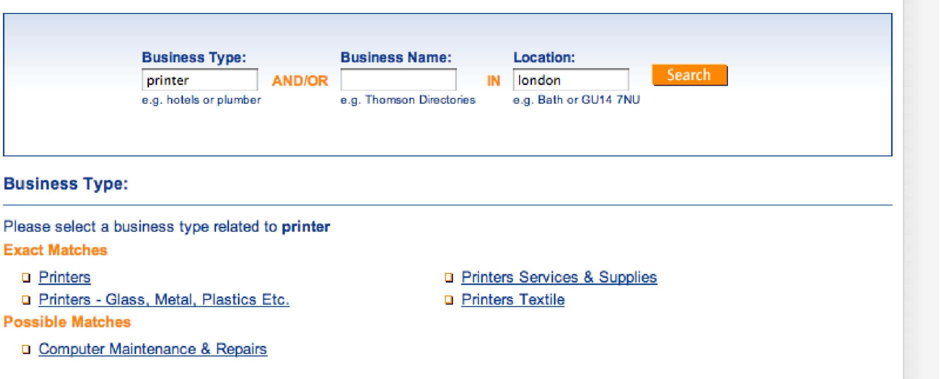
Sponsored search results

Looking for after-dinner speakers?

Red Squirrel Events Ltd specialises in organising hospitality packages and

Failed resolution of an ambiguous keyword

A second example, *printers*, is resolved to both meanings. While the meaning "printer (professionals)" is represented adequately, the meaning "printer (computer equipment)" is underrepresented in the results set:



Business Type: printer AND/OR Business Name: IN Location: london Search

e.g. hotels or plumber — e.g. Thomson Directories — e.g. Bath or GU14 7NU

Business Type:

Please select a business type related to **printer**

Exact Matches

- Printers
- Printers - Glass, Metal, Plastics Etc.
- Printers Services & Supplies
- Printers Textile

Possible Matches

- Computer Maintenance & Repairs

These findings are not exceptional for YP sites. While in this case, the singular / plural recognition is rather weak compared to other providers, it is rather common that YP sites fail to recognize morphological and especially semantical variants[9]. Many providers use a wildcard search on their categories, enriched with a limited number of fixed aliases. If they make use of an approximate matching solution, it is often only based on a pre-calculated lexicon and not capable of dynamically resolving searches to results in the index. A typical shortcoming of pre-calculated lexicon for spellchecking is the occurrence of suggested spelling correction that do not produce any results at all.

The results for several UK, US and German sites (including thephonebook.co.uk, ufindus.co.uk, scoot.co.uk, touchlocal.co.uk, yell.co.uk for the UK, dexonline.com, yellowpages.com, truelocal.com, AOL's / Google's / Yahoo's local sites for the US, gelbeseiten.de, herold.at, directories.ch, goyellow.de, 11880.com for the German-speaking

[9] The best performing site in the analysis of UK YP sites was `thephonebook.com` that were able to achieve an average rate of 75% intersecting results for morphological variations. Even this site had only 21% ratio of intersecting results for semantical variations.

market) revealed that all of these sites are not yet not capable of capturing morphological and especially semantical variations on a systematic level. While approximate search solutions help to detect orthographic variations, singular/plurals and some morphological variation, there is so far no site available that captures synonyms and relevant vocabulary for categories on more than an anecdotal level. The number of added terminology to the categories rank in general about several thousands to few ten thousands terms, while it would be rather ten times terms that are needed for a systematic recognition. While YP sites that incorporate Web crawl results in their index have in general far fewer failed searches than those sites that just use offline YP data, the consistency of search results, especially with regards to synonymous searches, is low on both types of YP sites with a slight advantage for those sites that only use offline YP data. This is because their result space is in general limited to the categories available and thus can be handled easier, at least for some exemplary or very frequent search terms, as these searches can be manually resolved to one or more categories.

17.2. Analysis of failed searches on YP sites

How can the logged user entries help to increase the performance of a YP site? In this section, the set-up of YP query logs, the nature of failed searches on a YP site, how they appear in the query log and what can be done to monitor and evaluate the improvements with regards to failed searches are examined.

A query log from a YP site has to minimally list all the entries in the different form fields, plus an indication whether the searches produced results. If the GUI consists of three fields (“Who”, “What” and “Where”), one entry in the query log should look like this:

```
”TIMESTAMP”;”WHO”;”WHAT”;”WHERE”,”SUCCESS”
”2006-06-13 00:04:23”;”claus gollwitzer”;”(null)”;”(null)”,1
”2006-06-13 00:04:25”;”(null)”;”Spezialtransporte”;”Braschwitz”,0
”2006-06-13 00:04:26”;”lübke”;”(null)”;”rheine”,1
”2006-06-13 00:04:29”;”urban”;”(null)”;”eggolsheim”,0
```

In addition, clickthrough and session information are helpful information that is worth gathering. A session could be marked by a hash value consisting of IP address, user agent, browser settings etc., thus allowing to identify the searches of individual users without breaching the privacy of individual users. In connection to session information, the clickthrough data reports what actions the user performed after submitting a query. The action could be either clicking on a result (ideally, the teaser text of the result is documented as well), reformulating the search or aborting the session.

Preliminary test results, based on 1.000 query occurrences from the telegate query log showed that on average, if a session does not produce a hit, 2.35 queries were entered by the user. If the initial query has no success, about 20% of the users do not proceed and reformulate their query. In one case, a user tried 18 times to reformulate the query without having success. In general, most reformulations try to broaden the search by leaving out one field or using only one word where previously two or more words have been entered (for example, omitting a first name in the “Who”-field).

Even without these additional data, there is a lot to be learned from the query strings and the indication of success/failure of a search alone. However, it is necessary first to single out what is the reason behind a failed search in order to determine which terms should be included into an enhanced vocabulary recognition module.

One way to do this would be re-engineering the complete search system on the YP site, including the index of businesses, advertisers' keywords, shortcuts, approximate matching components and other elements that are in operation on the live system. This is in general not a feasible solution. Apart from practical issues such as the size of the business index, this approach breaches the idea of having an objective test method that works outside of the live system.

Restricting the analysis to those entries that have only an entry in the “What?”-field is not a suitable approach either. Through this restriction, about 99% of all query occurrences[10] are removed. The optimal way of determining why a YP search fails and how it can be resolved is to shown on the level of individual queries how a failed query can be transformed into an equivalent query that produced results. As a result, it becomes clear which aspects of the query recognition process need to be enhanced. The transformation

$$Q_0 \rightarrow^T Q_1 \text{ with } T \text{ reflecting one of the variance principles as described in Part B}$$

can be established by applying the MetaMatch and an approximate matching functionality to the set of unresolved and resolved queries on the log. As this only explains the reasons why a query failed in the case it can be transformed into a successful query, this methodology has to be completed by a sample analysis of the remaining failed searches.

Based upon this, the notion of a top query log test battery can be developed. The top query log test battery consists of a considerable number of the most frequently occurring queries (in practice, a number of about 10.000 to 20.000 seems feasible). For these queries, their current status is noted and evaluated. The status of a query can be either failed or successful and what and how many categories and listings it produces. Also included into the status could be a flag indicating whether a query produced advertisements. The evaluation flag either indicates that the status of the query is the way it should be or, alternatively, how the ideal result of this query would look like. This top query test battery thus allows to monitor the process of improving query answering while at the same time focussing on the most important queries. If changes are made to the query answering process (for example changing the edit-distance threshold in an approximate matching module), their effects on the top queries can be studied and unwanted side-effects removed.

[10] Query-Log analysis conducted on the telegate log.

17.3. Reasons for inaccurately answered searches on YP sites

There are several reasons why searches could be answered inaccurately on a YP site. A broad segmentation of error sources separates three groups of problems. The first group consists of problems that arise from the underlying data index. The second group is built up by problems following from the access to these data. Finally, the presentation and interface may also be responsible for less than optimal results.

Data problems on the basis level are nothing more than accidental misinformation in one or more than one fields, for example an outdated telephone number. These type of errors usually have to be dealt with human editors, although their labor can be facilitated if information is extracted from the Web, for example by extracting contact information from company homepages. Other data errors can arise through a deduplication or address folding process that groups very similar or identical address records together. This process can produce two types of errors, either grouping of records that represent different businesses or overseeing records that represent identical businesses. Typically, office room sharing (for example through lawyers or physicians), leading to identical postal addresses for several business names is a source for such errors. Finally, the categorization of businesses might lead to problems in answering searches. More common than total mismatches between the businesses' genuine category and the category in the index, are more subtle errors in the assignment of categories, for example attributing a manufacturer-type category to a business that only operates as a wholesaler. Another typical categorization problem are businesses that are assigned to many different categories.

The category system used for the data is a further potential source of inaccurate results. Commonly, the category system reflects both the businesses needs stemming from advertisers and the possible heterogeneous data sources. Reconciling these two sources of bias with the goal of creating an consistent, easily accessible and non-redundant category system is not always possible. For example, the category system might contain categories that subsume other categories (e.g. *physicians – nephrologists*) in oder to reflect different levels of granularity of the indexed data (e.g. only some physicians in the data index are known by their specialization).

Moving on to the matching process between users' input and the index fields, reasons for inaccurate answers might lie either in missing a valid match or over-generating matches. The latter occurs for example with an over-permissive spelling correction that produces a correction that was not intended by the user. Another example of over-generating occurs with simple substring or wildcard searches that neglect morpheme boundaries (for example, matching *autor* to *autoreifen*). A missing or too strict approximate matching component will lead to under-generation (for example, failing to match *frisieur* to *friseur*). The more keywords pointing to categories are introduced, the more helpful becomes approximate matching, because there are more terms in the dictionary leading to results that can be used for finding an approximate match.

In general, however, the single most severe drawback of YP search systems lies in the semantic matching component. The lack of term enrichment, both on a per-category basis and on a per-business basis prevents accurate answers for many searches. On a per-category basis, all equivalent terms to a category (for example *attorneys - solicitor*) should be recognized and treated equal than these categories. In addition, terms that reflect needs that are best answered by a category need also to be introduced as pointers to the category. These pointers have to reflect plausible relations between a needs and the types of businesses offering those needs. In this sense, it is plausible that a sports retailer sells tennis sneakers, but it is not plausible that one can find parachutes there.

In general, mappings to categories could be either too broad — resolving for example *wedding dresses* to *fashion stores* if a specific *wedding fashion* -category is available — or too specific — resolving *headache* to *neurologist*, but not to *pharmacies*.

A further challenge is the resolution of ambiguous terminology (see above). Errors can occur here if one meaning is not or not adequately represented in the set of results.

Finally, with regards to the interface and the user interaction it allows, a common issue on YP sites is their reaction to few or zero results. The ideal mechanism offers the user more than just going back to the entry mask again. Recuperative strategies are the broadening of the geographical scope — the business sought for is not present in the specific locality, but it could be present in the vicinity of it — or, to a lesser extent, the broadening of the category scope — for example, there is no *women's fashion shop* in one location, but maybe a general *fashion shop*.

17.4. Creating a top domain thesaurus

A systematic approach to vocabulary enhancements that is re-usable for several YP sites is to set up a comprehensive mapping table that contains the branch in their common variations (including how they appear on different branch lists). The top domain thesaurus also includes the best representatives for each branch, both on the level of Sense-Morphemes and on the level of terms.

The motivation for setting up this thesaurus, the process that populates it and the fields of application are discussed below.

17.4.1. Motivation and goals

The main goals of creating a top domain thesaurus can be summarized by the following three items:

1. Set up an intuitive normalized categories that spring out of users' search habits instead of inconsistent, historically grown category systems

2. Incorporate common variations of these normalized categories

3. Incorporate cross-relations to other category systems and to standardized product and branch classifications

4. Incorporate a broad range of keywords that can serve as pointers to the normalized categories

A recurrent problems with business directories is the variance of terms used for branches. What one directory calls *Actuaries*, another directory lists under *Actuarial Services*, where one can look up *Acupressurists* in the first, she has to resort to *Acupressure* in the second. Not all variances can be detected by edit-distance mechanisms. In some cases, only by knowledge on synonymic variants the corresponding equivalent branch term of another directory may be found: *Laces* vs *Braids*, *Advertising Agencies & Consultants* vs. *Advertising Agencies & Counselors* or *Automobile Fleet Management* vs. emphAutomobile Fleet Maintenance. Moreover, in a considerable number of cases the true equivalent term is lacking. This happens mostly when the granularity differs between the directories. Consider for example *canners* for which in another category system the best hit would be *Food Processors.*

A related issue is the quality of category systems. In many cases, there are cases of intersecting categories, very broad and unspecific categories and a considerable portion of very weakly populated categories. For example, among the ca. 8.080 datagate categories, about 1.000 have currently no business listings to it[11]. There are also many instances of overlapping categories, for example *Zeitarbeitsunternehmen* vs. *Zeitarbeits- und Personalleasingunternehmen.* Another issue that has to be tackled in the TDT are synonymous or quasi-synonymous category descriptors in real-life category systems, for example *Landwirte* and *Landwirtschaftliche Betriebe.*

17.4.2. Headings and sectors

The organization above the level of normalized YP headings on the top domain thesaurus are broad business sectors. The main purpose of sectors lies in maintenance of the normalized YP headings. The selection of sectors also reflects groupings of businesses with a common basis. Empirical facts that corroborate the existence of a sector are for example fairs, associations or — in the case of businesses that require an academic education — also faculties. For example, *chemistry* can be considered to be a sector, as the different types of chemistry-related businesses all share the common ground of belonging to one industry, expressed in faculties, fairs, trade associations, labor unions etc. all dedicated to chemistry[12].

Just as with YP headings, it is also possible to find representative Sense-Morphemes for sectors. These morphemes can help to detect additional YP headings that should be incorporated into a sector. Moreover, they allow to restrict the set of candidate representants for a group of YP category by using only candidates that follow from the set of Sense-Morphemes in the sector.

A sample of sectors and Sense-Morphemes that are highly representative of the sector is provided in the table below.

- reinigung

  reinigen, putzen, säubern, hygiene, desinfektion, waschen, seife, polieren, schmutz

[11] The datagate categories are used by telegate, one of Germany's large directory assistance providers.

[12] See also [Gross/Guenthner 2002]

- sicherheit

  sicherheit, schützen, security, safe, gefahr, personenschutz, bodyguard, wache, patrouille, detektiv, überwachen, polizei, notfall, alarm

Based on experiences with German YP and business directory classifications, the following set of more than 50 sectors had been established:

- agrar
- automobil
- bau
- bergbau
- bildung
- chemie
- computer
- design
- information
- elektro
- energie
- entsorgung
- erotik
- finanz
- foto
- freizeit
- gastronomie
- garten
- gesundheit
- glas
- haushaltselektronik
- holz
- immobilien

- keramik
- nahrungsmittel
- kosmetik
- luftfahrt
- maschinenbau
- medien
- metall
- militär
- möbel
- personalwesen
- recht
- reinigung
- sanitär
- sicherheit
- spielwaren
- sport
- telefon
- textil
- schmuck
- unterhaltungselektronik
- tourismus
- verpackung
- versicherung
- versorgung
- verlag
- büro
- kunst

- soziales
- forschung
- papier
- unternehmensberatung
- persönliche dienste
- musik

Note that the naming of sectors tries to produce the most common representative of a sector and avoids coordinated constructions.

The relationship between branch headings and sectors is not always of 1:1 type, for example *Holzbau* which can be sorted into both *Bau* and *Holz* sectors. In addition, some branch headings are very broadly formulated and it is not clear where to subsume them. For example, *Agenturen* is a reduction form for several branches, such as *Werbeagenturen* or *Presseagenturen.* Such headings can be incorporated as synonyms to normalized branches, though.

Other YP branches only describe the mode of retail, but provide no topical characteristics. There is no sector for branches such as *Versandhandel* if the sectors should be based on topics. Another issue that has to be solved is at what point it is necessary to divide or unify sectors. For example, baby-related categories suggest that it makes sense to divide a sector *toys & children* and single out all baby-related branches, for example *baby swimming* (which is certainly not to be subsumed under the *sports* sector) or *baby equipment.*

In choosing the headings for the TDT, two aspects have to be kept in mind. The heading should represent an intuitive type of business and it should be expressed by a common representative of the variant YP headings that are in use for it. Naturally, some headings will be more popular and wider known than others, but all headings have to be chosen in a way that they spawn distinct vocabulary. For example, discriminating between *clothes wholesalers* and *clothes retail* makes sense when distinguishing businesses, but not when distinguishing vocabulary. Therefore, the TDT headings are stripped of the vendor type facets of conventional YP heading. In addition, if two headings are so similar to each other that the vocabulary they spawn largely overlaps, only one of them may enter the TDT.

In practical applications, the delivery of business categories by keywords from the TDT would use the normalized headings of the TDT as an intermediate step. Starting by locating a term (such as *polo shirt*) in the TDT, the corresponding normalized TDT heading is determined (for example, something like *Men's Fashion* and *Women's Fashion*) and finally a mapping from normalized TDT headings to the YP site's heading system (for example *Men's clothing wholesalers*, *Men's clothing retailers*, *Women's clothing wholesalers*, *Women's clothing retailers*) is applied.

In a first round, the headings of the TDT can be iteratively gathered by combining several YP categorization systems. In a preliminary step, only distinct categories of the external YP systems are kept and categories that only differ by vendor type are folded into one. When comparing the different category systems, categories can either appear in more than one system, requiring a de-duplication, or they appear only in one system. In the latter case, these headings should be carefully examined whether they in fact represent a genuine new concept or are just a variation of a heading already seen. Through this comparison, both a ranking of categories into central and peripheral categories and a first accumulation of category variants are created.

Moving on, other YP category systems can be incorporated successively. For monitoring the coverage of the TDT and adding more vocabulary to it, a pass through a YP site query log is helpful. Passing through the top entries, these terms are either possible good representatives for the headings or synonyms to a heading. Below are the top ranking terms from GY and how they can be associated with sectors. Lines without parentheses stand for individual sectors.

restaurant
friseur
arzt
zahnarzt
gaststätten (belongs to *restaurant*)
hotel
immobilien
rechtsanwalt
computer
möbel
zeitarbeit
autohaus
steuerberater
spedition
psychotherapie
kindergärten
schule
gaststätte (belongs to *restaurant*)
pension (relates to *hotel*)
masseurin
lebensmittel
pizza (relates to *restaurant* and *lebensmittel*)
zahnärzte (belongs to *zahnarzt*)

The majority of these lines are good representatives of headings. However, it becomes apparent even from this small sample that edit rules have to be enforced in order to create consistent headings. While most headings would have representatives in the singular, *Gaststätten* and *Kindergärten* deviates from this pattern.

17.4.3. Populating the TDT

Choosing the proper seed headings for extended YP vocabulary is the first step in setting up a Top Domain Thesaurus. The further steps increase both the breadth of the TDT by identifying branches that have not been covered before and at the same time deepening the knowledge of one branch by adding more vocabulary to it. Additional external heading systems can be incorporated one after another via cross-references.

These enrichment processes becomes easier as more and more vocabulary is structured into the TDT. Through setting up a Sense-Morpheme representation of the headings additional vocabulary can in many cases be auto-mapped via the Meta-Match process into the TDT with only a very limited need for manual inspection and correction.

Not all branches that are conventionally discriminated in YP heading systems can be equally differentiated by terms. Vendor type facets such as wholesalers or retailers do usually not have any impact on the vocabulary of a category[13].

The columns of the TDT — normalized branch, Sense-Morphemes, Products / Services / Keywords, Branch synonyms, Cross-references to norm classifications — are supposed to give a full picture of the vocabulary used in an intuitively understandable category.

For example, an entry in the TDT for the category *Fahrschulen* could look as follows:

Fahrschule

Sense-Morphemes

fahrschule
fahrausbildung
fahrprüfung
fahrlehrer
führerschein
fahrerlaubnis
fahreignung

Products / Services / Keywords (in selection)

fahrausbildung
fahrschule
führerschein
fahrerlaubnis
fahrlehrer

[13] Cross-relations, such as assigning a *dvd wholesaler* to *entertainment electronic retail* have to be avoided, though. This can be achieved by adding the orthogonal vendor facets at a later stage and only list neutral product and service categories in the TDT.

nachschulung
fahrschulen
führerscheinausbildung
führerscheinklassen
fahrprüfung
fahrschüler
ferienfahrschule
fahrstunde
motorradausbildung
eu-führerschein
sicherheitstraining
verkehrsübungsplatz
klasse b
führerscheinprüfung
testbögen
einparken
autoführerschein
klasse a
fahrtipps
motorradführerschein
lernsystem
fahranfänger
fahrertraining
fahren lernen
fahrstunden
fahrunterricht
fahrschulunterricht
fahrlehrerverband
stras̈s enverkehr
theorieprfung
theorieunterricht
fragebogen
führerscheinklasse
verkehrszeichen
ausbildungsfahrzeuge
klasse a1
klasse m
fahrtraining
verkehrspraxis
verkehrsamt
führerschein auf probe
pkw-ausbildung
gefahrgutausbildung
verkehrsinstitut

fahrpraxis
verbandsfahrschule
verkehrstheorie

Branch synonyms

fahrausbildung
fahrlehrer
fahrunterricht
führerscheinausbildung
kraftfahrschulen

Cross-references

WZ03: 80.41.1 Kraftfahrschulen
UNSPSC: 86131701 Vehicle driving schools services – Wagenfahren Schuledienstleistungen[14]
NACE (Rev. 1.1): 80.41 Fahr- und Flugschulen
CPA: 80.41.11 Dienstleistungen von Kraftfahrschulen

17.4.4. Applying the TDT

After having illustrated the vocabulary content of the TDT, how might the set of enriched normalized YP categories help to achieve better search results on YP sites?

The basic scenario is based on an index containing business records with category information. Often, this set of categories is build up of several sources and contains legacy elements as well as modifications made over time — for example due to advertisers asking for a separate category for their business. Another reason for inconsistencies in the index categories lies in the different level of quality of the address data. For example, a YP provider might purchase addresses of physicians that are very detailed and contain information about the specialization of each physician (ophthalmologist, nephrologist, podologist etc). These new addresses have to be integrated with an old address repository in which physicians have no detailed categorization to them. This leads to categories working in parallel such as *opththalmologist* and *specialized physician* in which one category subsumes the other one.

It is therefore required when setting up a correspondence table between the normalized TDT categories and the external heading system to acknowledge the different level of granularity of categories. Moreover, in some cases the optimal corresponding cannot even be deducted from the category label, but has to determined by looking

[14] While the English descriptors are often not the most common names for a category, the quality of UNSPSC translations are in a considerable number of cases totally unacceptable (consider also the example “Herbal medicine or herbalists services”, translated to “Kräuter Medikamente oder Kräuterkennerdienste”).

at how many and what businesses are registered to a category. Consider for example a category system that contains *fitness clubs.* This seems on first glance as an ideal place for inserting all the *fitness club*-related taxonomy into the system. However, it might turn out that in fact most of the fitness clubs are registered as *sport clubs.* Such cases — which happen not too seldom with real YP systems — have to be resolved by taking a look at the actual business data in a category.

Another case that was touched upon above is that sometimes categories subsume other categories. In this case, it can be decided whether to use an inheritance mechanism that accumulates all the vocabulary specific to the subcategories — for example, all ethnic restaurants — to the supercategory — for example, a category labeled *restaurant*, or alternatively to use only vocabulary of equal generality for the supercategory.

Depending on the indexed business data, both alternatives might make sense. If choosing the first alternative, then a specific search for an ethnic cuisine (for example *chicken korma*) brings out not only restaurants that are registered to this specific category, but also restaurants that are registered to the broad category *restaurant.* This makes sense if for most restaurants in the index, their specific type of cuisine is not known, i.e. they are registered to the broad category *restaurant.* In order to avoid zero hits for most locations, the specific dish would then pull out all restaurants, even if no indication that these restaurants would serve the particular dish by virtue of their cuisine is available. In contrary, if for most restaurants in the index their particular cuisine is known, it is no use to pull out hits from the small set of restaurants for which no subcategory is available. Conversely, however, a general search such as *eating out* would then be mapped not only to the broad category *restaurant*, but to all of its subcategories.

As the categories of the TDT are based on types of business and the different vocabulary they spawn, the TDT largely provides the same vocabulary for different vendor types such as wholesalers or retailers. For example, all the subtypes to *footwear* — such as *sneakers, mokassins, high heels* — can correspond to wholesaling or retailing footwear. The TDT therefore is organized by having a neutral, product-based category that contains all the vocabulary that is applicable to all vendor types of a product or service. In addition, specific vocabulary for wholesaling or retailing can be introduced in a separate category of the TDT.

The resulting correspondence table then looks as follows:

TDT

Footwear-Products (for ex. sneakers, mokassins, high heels)

Footwear-Wholesaler (for ex. shoe wholesalers, footwear wholesaling)

Footwear-Retail (for ex. shoe stores, footwear retail)

External heading system and correspondences to the TDT

External: Footwear-Wholesalers → TDT: Footwear-Products and Footwear-Wholesaler

External: Footwear-Retail → TDT: Footwear-Products and Footwear-Retail

Through this approach, a cross-matching (users searching for footwear wholesalers and receiving results from retailers) are eliminated, while general product-type searches produce the maximum number of relevant results.

18. Outlook and Future Work

Corresponding to the analysis of E-Commerce as being in the state of term-driven E-Commerce, the proposed enhancements related to the level of terms and term aggregation. Making matches between the participants of the virtual market place is fundamentally a matching between terms. In analyzing the possible interactions between agents of the virtual market — including immediate participants in transactions and aggregators of participants — it was pointed out how tightly connected these interactions are to a textual representation, and moreover to a term-based representation.

In the model of Term Spaces that was introduced to grasp these term representations, two related spatial metaphors meet. Firstly, a Term Space originates in the variability of terms, leading to the ”corona” of equivalent or related variants that surround a term. Secondly, collecting the terms that are gathered over time (for example query logs) or synchronous (for example accumulated meta keywords from Web crawls) leads to “landscape” of terms that invites to be explored and surveyed. These landscapes of terms exhibit stable and variable stretches. This spatial metaphor illustrates the commensurability of Term Spaces, including the combination, division and structuring of Term Spaces.

The transformation process from Commerce to TE-Commerce still continues. One of the most promising areas for TE-Commerce are mobile devices. Considering that the display size of mobile devices is much smaller than that of PCs or notebooks, mobile Commerce (also called M-Commerce) faces the same push into crisp representations of needs and offers. It is not feasible to present hundreds of result and rely on the same trial-and-error process that is currently employed by many users of Web search engines.

The limited number of input keys in mobile devices increases the need to recognize the users input with as little keystrokes as possible, as interaction (such as reformulating a search input) is time-consuming. Using speech recognition as input device also requires a recognition of the E-Commerce-related meaning of terms and term variants.

Regardless of the device with which E-Commerce transactions can be conducted, typical E-Commerce applications needs a sophisticated term handling. In the sections below, an outlook to what this could comprise for four examples of E-Commerce applications is provided. The four exemplary applications are the following:

1. Optimizing vertical search
2. High precision sentiment analysis
3. High precision product recommendations

18.1. Optimizing vertical search

Optimizing a vertical or topical search requires to know about the entities in a field, their attributes and the relations to each other and how all of this is represented in a topic-specific vocabulary. For users, vertical searches bring the benefit that their searches can be analyzed with a much higher granularity and produce only results from the domain they are interested in. For example, typing in *notebooks* on a vertical search portal for computers avoids to have stationery-related results in the result lists.

For example, setting up a vertical Web search for hotels would entail first to accumulate the entities for this topic which are in this case mainly individual hotels (at least with address information, ideally with a homepage) and hotel chains. In addition, it is conceivable that a limited number of hotel suits are prominent enough to be specifically asked.

Properties of these entities consist on the one hand of standardized features that exhibit low variance in expression, such as the prize per night, the availability for a certain span of time, or the star rating. On the other hand, properties such as individual features of a hotel or a hotel-room (conferencing facilities, Internet connection, concierge service) can be expressed in many different ways. These variant expressions have to be analyzed both on the webpages of hotels (or other descriptive passages on the hotel, for example in printed hotel guides or hotel brochures) and in user queries. For this purpose, specially tailored crawlers and information extraction modules need to be set up[1]

These matching procedure is in general a matching between terms, for example matching the query *hotel conference room* to the phrase in a hotel brochure *conferencing facilities available*. In addition, some queries can be analyzed as containing information which maps to structured and standardized properties. For example, a query looking for a "beachfun hotel" could be brought into connection to a geo database that indicates for a location whether it is close to a beach. Such associations between free text queries and structured information may also be used for ranking results. A vertical search confronted with a broad search such as "luxury hotel" could put hotels with four or five stars in the top positions.

18.2. Sentiment analysis

It was already laid out that exchange of sentiments is an important transaction type in the c2c-world. A feedback based on the viral sentiments spreading on customer protection sites, blogs, forum etc. could often help enormously in product design and marketing. For users, it would be helpful to access the opinions other users expressed without having to look for and read every review published on the Web. In both cases, analyzing sentiments should go beyond a simple count of general praise or critic (thumbs-up vs. thumbs-down) and include the detection of recurrent complaints or

[1] See [Chakrabarti/van den Berg/Dom 1999], [Chau/Qin/Zhou/Tseng/Chen 2005] and [Chau 2002] for an introduction to focussed crawling.

praises about specific product traits.

While some words have a connotation that makes recognizing an evaluation easy (for example: *superb, wonderful, uncomfortable, crap*), determining the attitude of a speaker that is expressed in discourse strategies requires a much deeper textual analysis[2].

Current research approaches use almost exclusively machine learning features to extract if and what users criticize or praise regarding products[3]. While the machine learning frameworks differ greatly (support vector machines, signal decomposition, Bayesian classifiers, decision trees etc), the features used in almost all approaches are almost exclusively based on literal strings and do not incorporate lexical and syntactical variances of expression. Moreover, the lexical resources deployed (if any) are in general not specific to E-Commerce. Gathering adjectives that are connected in WordNet to "good" and "bad" (such as *bang-up, bully, corking, cracking, dandy, great, groovy, keen, neat, nifty, peachy, slap-up, swell, smashing*) cannot be more than a first step towards a comprehensive analysis of evaluative lexis.

While statistical approaches are able to recognize strong lexical connotation, they are cum grano salis unable to grasp complex evaluative patterns. Taking a closer look at these patterns and variants of evaluative expressions, it becomes apparent that they need to be described on a local level in order to acknowledge their richness and repetitive structure.

The following examples were extracted by scraping search engines for patterns based on a clearly evaluative word, *impressed* and grouped based on the intense or other facets (for example temporal facets) of the sentiment:

IMPRESSED - MILDLY
was quite impressed by its
was pretty impressed by its
was generally impressed by its
was rather impressed by its
was somewhat impressed by its
was kinda impressed by its

IMPRESSED - MEDIUM
was genuinely impressed by its
was favourably impressed by its
was reasonably impressed by its
was suitably impressed by its
was pleasantly impressed by its
was really impressed by its

[2] See [Thompson/Hunston 2000] and the appraisal theory, presented in [Martin/White 2005].

[3] See [Turney 2002], [Wilson/Wiebe/Hoffmann 2005], [Wilson/Wiebe/Hwa 2004], [Hu/Liu 2004], [Morinaga/Yamanishi/Tateishi/Fukushima 2002], [Kushal/Lawrence/Pennock 2005] and [Popescu/Etzioni 2005].

IMPRESSED - GREATLY
was greatly impressed by its
was particularly impressed by
was very impressed by its
was thoroughly impressed by its
was especially impressed by its
was highly impressed by its
was mightily impressed by its
was tremendously impressed by its
was duly impressed by its
was most impressed by its
was totally impressed by its
was truly impressed by its
was hugely impressed by its
was profoundly impressed by its
was immensely impressed by its
was heavily impressed by its
was incredibly impressed by its
was enormously impressed by its
was distinctly impressed by its
was overly impressed by its
was absolutely impressed by its
was seriously impressed by its
was dreadfully impressed by its
was severely impressed by its
was singularly impressed by its

IMPRESSED - INSTANTLY
was immediately deeply impressed by
was immediately impressed by
was instantly impressed by its
was quickly impressed by its
was at once impressed by its

CONSTANTLY IMPRESSED
was constantly impressed by its
was always impressed by its
was still impressed by its

IMPRESSED - CONTRAST / NEGATIVE
was nonetheless impressed by its
was unexpectedly impressed by its
was insufficiently impressed by its
was rather routinely impressed by its

POSITIVELY EVALUATING VERBS - ”SURPRISE” CONNOTATION
excited by its
struck by its
baffled by its
surprised by its
stunned by its
stricken by its
awestruck by its

POSITIVELY EVALUATING VERBS - ”TAKEN-BY” CONNOTATION
taken by its
intrigued by its
attracted by its
fascinated by its
hooked by its
drawn by its
seduced by its
delighted by its
charmed by its
entranced by its
riveted by its
enthralled by its
transfixed by its
mesmerized by its
captivated by its
swayed by its
intrigued by its
enthralled by its
overwhelmed by its

The number of these patterns is increased if insertions (such as *still, constantly, very, doubtlessly*) that may occur at different positions are taken into account. All these variants would fit into a compact local grammar.

A variety of German evaluative contexts shows the different registers used and the need to keep track of idiomatic evaluative expressions. In the following list of positive German evaluative contexts, the strategies to convey a positive opinion range from use of verbal predicates (*empfehlen, nie entäuscht*), adjectival predicates (*genial, toll, reibungslos*) and nominal predicates (*Wunderteil*) to idiomatic expressions (*Hut ab*):

würde es nur empfehlen
die Qualität ist enorm hoch
hat mich noch nie entäuscht
sprengte alle meine Erwartungen
dieses Gerät ist wirklich eine gute Wahl

ein empfehlenswertes Produkt
bin bisher mit dem Gerät rundum zufrieden
es funktionierte reibungslos
einfach toll das Ding
das Ding ist genial
da kann ich nur sagen, "Hut ab"
bin geilstens zufrieden
hier springt für das Produkt ein ganz klares sehr empfehlenswert heraus
bin absolut begeistert
ein wirkliches Wunderteil
kann ich euch allen dieses Gerät nur ans Herz legen
ein Gerät mit sehr guter Qualität

18.3. Product recommendations

Current approaches to collaborative filtering for product recommendations set up a matrix containing the distinct items to be recommended and the user ratings or purchases[4]. This requires products with an unique identifier that can be determined for each interaction on which the collaborative filtering should be based[5].

If product recommendations should be extended to cover domains where such unique identifiers are not available or cannot be monitored for each interaction, a term handling process is needed to recognize identical products under different textual descriptions. For example, if product recommendations should be made usable for fast moving consumer goods, it will usually only be the bar code that represents the article if purchases are monitored (for example at the point-of-sales terminals). This requires a grouping of bar codes for similar products.

The EAN number used in North America and Europe consists of a country prefix, a maker number and then the article suffix which is managed by the maker[6]. While this allows to discern the maker from the number, finding out if two EANs relate to the same type of product is not possible without additional information that again comes in textual form[7].

Another example of the importance of term handling for product recommendations is the music domain where recommendations and artists, songs and albums could be determined based on purchase histories or opinions extracted from music reviews on the Web[8]. While monitoring purchases on one site can rely on an unique ID per item, the same does not hold if sentiments stemming from different websites should be taken into account. In general, the only available representation of the music item is

[4] The seminal paper of collaborative filter is [Resnik/Iacovou/Suchak/Bergstrom/Riedl 1994]. See also [Sarwar/Karypis/Konstan/Ried 2001] and [Breese/Heckerman/Kadie 1998].

[5] On Amazon, this is provided by the ASIN number.

[6] http://www.gs1.org/services/gsmp/ [Nov. 1, 2006].

[7] This could be avoided if there was one finely granulated universal classification of articles by type that is available for all EAN which is not available today.

[8] In the movie domain, such a system is presented in [Miller/Albert/Lam/Konstan/Riedl 2003].

its textual representation. To illustrate how varying this representation might be, here is a selection of different spellings for *Elvis Presley* in *freedb*, a database containing CD track information[9].

Elvis Presley

elvis presley

ELVIS PRESLEY

Presley, Elvis

Elvis presley

Presley Elvis

PRESLEY Elvis

PRESLEY ELVIS

Evis Presley

Elvis_Presley

ElvisPresley

Elvil Presley

Elivs Presley

Elis Presley

El;vis Presley

ELvis Presley

ELVIS PRESLEY

ELIVIS PRESLEY

Elvis Presly

elvis presly

Elvis Presely

18.4. Epilogue

If TE-Commerce is indeed an intermediate stage of development, what it is that comes after it and what developments are already observable that hint into this direction?

It was already claimed in the introductory passages of this thesis that a genuine textual stage of E-Commerce would be the characteristics after leaving the term-driven stage. The applications touched upon before indicate how a deep understanding of text discourses could boost E-Commerce. It would also be the prerequisite to a true intensional querying that makes searching on free text just as robust and reliable than searching SQL databases.

[9] See `www.freedb.org`

Another general trend that seems to be inherent to the evolution of TE-Commerce are long-tail searches, i.e. leading the user from items she knows about to niche items not known before. With advanced accessing procedures, it becomes possible to search comfortably very large data sets without being overwhelming by noise in the results. The long-tail paradigma is often seen in connection to a decentralization of first level agents (with eBay as a striking example of a drastic increase in inventory and number of vendors). Going hand-in-hand with a decentralization of first level agents, a concentration process in second level agents can be observed (see above, Part A, TE-Commerce Agents).

It seems to be more than just irony that one of the biggest troubles for search engines lies in clickthrough and aggregator sites and that lower the relevance of search results, especially for E-Commerce-related searches with a local scope. The growing concentration within Web search engines makes it attractive for smaller aggregators to put their energy into search engine optimization or even manipulation. This, on the other hand, makes it possible for specialized search agents to step in if the users of generic Web search engines begin to feel discontent about the large number of aggregator sites for searches on a specialized topic. Following from that, the danger of a monoculture in search should not be over-emphasized.

Term handling does not pave the way to a monoculture in second level agents. Even if a comprehensive lists of term variants and robust algorithm to detect them are build, the constant change in terms and term usage prevents that such resources would be an unsurpassable advantage for a search agent. On the contrary, if one agent offers a richer search experience through a better term handling, its competitors will try to follow. As terms as commonly shared descriptors are not the intellectual property of individuals, it will be always possible to try to enhance their handling, and broaden as well as deepen the lexical resources that describe them[10].

[10] See also the roadmap laid out by Wikipedia founder Jim Wales, “Ten things that will be free”, online at http://upload.wikimedia.org/wikipedia/commons/a/aa/Wikimania_Jimbo_Presentation.pdf [Nov. 1, 2006].

Bibliography

[Abiteboul 1997] Serge Abiteboul, “Querying Semi-Structured Data”, Proceedings of the 6th ICDT.

[Aggarwal/Goel/Motwani 2006] G. Aggarwal / A. Goel / R. Motwani, “Truthful auctions for pricing search keywords”, 7th ACM Conference on Electronic Commerce.

[Alfonseca / Manandhar 2002] Enrique Alfonseca / Suresh Manandhar, “Improving an Ontology Refinement Method with Hyponymy Patterns”, LREC: 235-239.

[Anderson 2006] Christian Anderson, *The long tail. Why the Future of Business Is Selling Less of More*, Hyperion.

[Baeza-Yates/Ribeiro-Neto 1999] Baeza-Yates / Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley.

[Barker/Szpakowicz 1998] Ken Barker / Stan Szpakowicz, “Semi-Automatic Recognition of Noun Modifier Relationships”, 36th ACL meeting.

[Beeferman / Berger 2000] Doug Beeferman / Adam Berger, “Agglomerative clustering of a search engine query log”, online at http://www.dsi.unive.it/ orlando/Topic-WSE-Queries/p407-beeferman.pdf [1st of October 2006].

[Biemann / Osswald 2005] C. Biemann / R. Osswald, “Automatische Erweiterung eines semantikbasierten Lexikons durch Bootstrapping auf groen Korpora”, GLDV Frühjahrstagung.

[Blanco/Guenthner 2004] Franz Guenthner / Xavier Blanco, “Multi-Lexemic Expressions: an overview”. *Lexique, Syntaxe et Lexique-Grammaire / Syntax, Lexis & Lexicon-Grammar*, ed. Christain Leclere / Eric Laporte /Mireille Piot / Max Silberztein, John Benjamins.

[Bourigault/Jacquemin/L’Homme 2001] *Recent Advances in Computational Terminology*, eds. Didier Bourigault / Christian Jacquemin / Marie-Claude L’Homme. Amsterdam: John Benjamins.

[Breese/Heckerman/Kadie 1998] J. Breese, David Heckerman and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence, pages 43–52, July.

[Brill/Dumais/Banko 2002] Eric Brill / Susan Dumais / Michele Banko, "An Analysis of the AskMSR Question-Answering System", online at http://research.microsoft.com/ brill/Pubs/EMNLP2002.pdf [1st of October 2006].

[Brin / Page 1998] Sergey Brin / Lawrence Page, "The anatomy of a Large-Scale Hypertextual Web Search Engine", online at http://infolab.stanford.edu/pub/papers/google.pdf [1st of October 2006].

[Brin 1998] Sergey Brin, "Extracting Patterns and Relations from the World Wide Web", The World Wide Web and Databases: International Workshop WebDB'98. Hrsg. P. Atzeni et al. Valencia: 172-183.

[Bryan/Gershman 2000] Dough Bryan / Anatole Gershman, "The Aquarium: A Novel User Interface Metaphor for Large, Online Stores", Proceedings 11th International Workshop on Database and Expert Systems Applications.

[Broder 2002] Andrei Broder, "A taxonomy of web search", online at http://www.acm.org/sigs/sigir/forum/F2002/broder.pdf [1st of October 2006].

[Buchholz / van den Bosch 2000] Sabine Buchholz / Antal van den Bosch, "Integrating seed names and ngrams for a named entity list and classifier", Proceedings of LREC-2000. Athen: 1215-1221.

[Budanitsky / Hirst 2001] Alexander Budanitsky / Graeme Hirst, "Semantic distance in WordNet:. An experimental, application-oriented evaluation of five measures", online at http://ftp.cs.toronto.edu/pub/gh/Budanitsky+Hirst-2001.pdf [1st of October 2006].

[Carrasco/Fain/Lang/Zhukov 2003] J. Carrasco / D. Fain / K. Lang / L. Zhukov, "Clustering of bipartite advertiser-keyword graph", International Conference on Data Mining.

[Chakrabarti 1999] Soumen Chakrabarti, "Automatic Web Resource Discovery", online at http://http.cs.berkeley.edu/ soumen/doc/acmcs1999/acmcs.pdf [1st of October 2006].

[Chakrabarti/van den Berg/Dom 1999] S. Chakrabarti / M. van den Berg / B. Dom, "Focused Crawling: A new approach to Topic-Specific Web Resource Discovery", Computer Networks 31.

[Chakrabarti 2003] Soumen Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data*, Morgan-Kaufmann.

[Chau 2002] M. Chau, "Spidering and Filtering Web Pages for Vertical Search Engines", Proceedings of The Americas Conference on Information Systems, AMCIS 2002 Doctoral Consortium, Dallas, Texas.

[Chau/Qin/Zhou/Tseng/Chen 2005] M. Chau / J. Qin / Y. Zhou / C. Tseng / H. Chen, “SpidersRUs: Automated Development of Vertical Search Engines in Different Domains and Languages”, Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL’05), Denver, Colorado.

[Chen 1976] Peter P. Chen, “The Entity-Relationship Model - Toward a Unified View of Data”, ACM Transactions on Database Systems 1.

[Choueka 1988] Yaacov Choueka. ”Looking for needles in a haystack or locating interesting collocational expressions in large textual databases”, Proceedings of the RIAO Conference on User-Oriented Content-Based Text and Image Handling, Cambridge, MA, March 21-24.

[Cimiano et al 2004] Philipp Cimiano / Siegfried Handschuh / Steffen Staab, “Towards the Self-Annotating Web”, Proceedings of the WWW-Conference 2004, New York.

[Constant 2001] Matthieu Constant, “On the Analysis of Locative Phrases with Graphs and Lexicon-Grammar: the Classifier/Proper Noun Pairing”, Advances in Natural Language Processing. Proceedings of PorTAL. Berlin: 33-42.

[Coupet 2006] Pascal Coupet, “Searching and Mining”, Search Engine Meeting, Boston, April 24, online at http://www.infonortics.com/searchengines/sh06/slides/temis.pdf [1st of October 2006].

[Cruse 1986] David Alan Cruse, *Lexical Semantics*, Cambridge University Press.

[Cruz / Rajendran 2003] Isabel F. Cruz / Afsheen Rajendran, “Exploring a New Approach to the Alignment of Ontologies”, online at http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-83/int_1.pdf [1st of October 2006].

[Cruz et al 2004] Isabel F. Cruz / William Sunna / Anjli Chaudhry, “Ontology Alignment for Real-World Applications”, in Proceedings of The National Conference on Digital Government Research, online at http://dgrc.org/dgo2004/disc/posters/tuesposters/rp_cruz.pdf [1st of October 2006].

[Cucerzan / Yarowsky 1999] Silviu Cucerzan / David Yarowsky, “Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence”, Proceedings, 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora. University of Maryland: 90-99.

[Cutler / Shih / Meng 1997] Michal Cutler / Yungming Shih / Weiyi Meng, “Using the Structure of HTML Documents to Improve Retrieval”,Proceedings of the USENIX Symposium on Internet Technologies and Systems. Monterey: 241-251, online at http://www.usenix.org/publications/library/ proceedings/usits97/full_papers/cutler/cutler.pdf [1st of October 2006].

[Daum / Brill 2004] Hal Daum III / Eric Brill, "Web Search Intent Induction via Search Results Partitioning", online at http://citeseer.ist.psu.edu/653787.html [1st of October 2006].

[Davies/Fensel/van Harmelen 2003] John Davies / Dieter Fensel / Frank van Harmelen, *Towards the Semantic Web: Ontology-Driven Knowledge Management.* John Wiley & Sons.

[Ding / Gravano / Shivakumar 2000] J. Ding / L. Gravano / N. Shivakumar, "Computing geographical scopes of web resources", Proceedings of the Twenty-sixth International Conference on Very Large Databases (VLDB00).

[Dill et al 2003] Stephen Dill / Nadav Eiron / David Gibson / Daniel Gruhl / R. Guha / Anant Jhingran / Tapas Kanungo / Sridhar Rajagopalan / Andrew Tomkins / John A. Tomlin / Jason Y. Zien, "SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation",Proceedings of the WWW-Conference 2003, Budapest.

[Donalies 2002] Elke Donalies, Die Wortbildung des Deutschen. Gunter Narr.

[Doubleclick 2005] Doubleclick 2005, "Search Before the Purchase", online at http://www.doubleclick.com/us/knowledge [1st of October 2006].

[Etzioni et al 2004] Oren Etzioni et al, "Web-Scale Information Extraction in KnowItAll", WWW Conference.

[Evert/Heid/Lezius 2000] S. Evert /U. Heid / W. Lezius, "Methoden zum Vergleich von Signifikanzmaßen zur Kollokationsidentifikation", KONVENS.

[Fanselow 1981] Gisbert Fanselow, *Zur Syntax und Semantik von Nominalkomposita. Ein Versuch praktischer Anwendung der Montague-Grammatik auf die Wortbildung des Deutschen*, Niemeyer.

[Ferret et al 2001] O. Ferret / B. Grau / M. Hurault-Plantet / G. Illouz / L. Monceaux / I. Robba / A. Vilnat, "Finding an answer based on the recognition of the question focus", online at http://trec.nist.gov/pubs/trec10/papers/qaLIR.pdf [1st of October 2006].

[Fleischer/Barz 1995] Wolfgang Fleischer / Irmhild Barz, Wortbildung der deutschen Gegenwartssprache. Niemeyer.

[Foskett 1982] A.C. Foskett, The subject approach to information. Clive Bingley.

[Friburger / Maurel 2001] Nathalie Friburger / Denis Maurel, "Elaboration dune cascade de transducteurs pour lextraction de motids: lexemple des noms de personnes", Actes de la 8me confrence sur le Traitement Automatique des Langues Naturelles TALN. Tours: 183-192.

[Gal / Modica / Jamil 2003] Avigdor Gal / Giovanni Modica / Hasan Jamil, “Improving Web Search with Automatic Ontology Matching”, online at http://ranger.uta.edu/ alp/ix/readings/webOntologyMatching.pdf [1st of October 2006].

[Gaus 2003] Wilhelm Gaus, *Dokumentations- und Ordnungslehre*, Springer.

[Gilchrist 2000] Alan Gilchrist, “Taxonomies for Business”, TFPL, London, online at http://www.bokis.is/iod2001/papers/Gilchrist_paper.doc [1st of October 2006].

[Gomez-Perez / Fernandez-Lopez / Corcho 2004] Asuncion Gomez-Perez / Mariano Fernandez-Lopes / Oscar Corcho, *Ontological Engineering*, Springer.

[Goldsmith 2001] John Goldsmith, “Unsupervised learning of the morphology of a natural language”, Computational Linguistics 27/2.

[Goodman 2004] Andrew Goodman, *Google AdWords Handbook: 21 Ways to Maximize Results.*

[Glover et al 2001] Eric J. Glover / Gary W. Flake / Steve Lawrence / William P. Birmingham / Andries Kruger / C. Lee Giles / David Pennock, “Improving Category Specific Web Search by Learning Query Modifications”, online at http://clgiles.ist.psu.edu/papers/SAINT-2001-learning-queries.pdf [1st of October 2006].

[Gravano/Hatzivassiloglou/Lichtenstein 2003] Luis Gravano / Vasileios Hatzivassiloglou / Richard Lichtenstein, “Categorizing Web Queries According to Geographical Locality”, online at http://www1.cs.columbia.edu/ gravano/Papers/2003/cikm03.pdf [1st of October 2006].

[Gross 1979] Maurice Gross, “On the failure of generative grammar”, Language 55:4: 859-885.

[Gross 1999a] Maurice Gross, “Lemmatization of Compound Tenses in English”, Lingvisticae Investigationes XXII.

[Gross 1997] Maurice Gross, “The Construction of Local Grammars”, Finite-State Language Processing, ed. E. Roche / Y. Schabes. MIT Press.

[Gross 1999b] Maurice Gross, “A Bootstrap Method for Constructing Local Grammars”, Proceedings of the Symposium on Contemporary Mathematics. Belgrad.

[Gross 2002] Maurice Gross, “Les dterminants numraux, un exemple: les dates horaires”, Languages: 21-37.

[Gross/Guenthner 2002] G. Gross / F. Guenthner, “Comment décrire une langue de spécialité ?”, Cahiers de Lexicologie 80.

[Guenthner / Blanco 2004] Franz Guenthner / Xavier Blanco, “Multi-lexemic Expressions: An Overview”, Lingvisticae Investigationes Suplementa. Amsterdam/Philadelphia, online at http://seneca.uab.es/filfrirom/BLANCO/PUBLIC/multilex.pdf [1st of October 2006].

[Guenthner 2002] Franz Guenthner, “Suchmaschinen als Wissensvermittler”, Die Geisteswissenschaften in der Informationsgesellschaft. Hrsg. Venanz Schubert. St. Ottilien: 109-126.

[Harris 1951] Zellig Harris, Structural Linguistics. The University of Chicago Press.

[Harris 1955] Zellig Harris, “From phonemes to morphemes”, Language, 31/2.

[Hearst 1992] Martin A. Hearst, “Automatic acquisition of hyponyms from large text corpora”, Proceedings of the 14th conference on Computational linguistics.

[Hearst 1998] Martin A. Hearst, “Automated discovery of wordnet relations”, WordNet: An Electronic Lexical Database, ed. Christiane Fellbaum. MIT Press.

[Heid et al 1996] Heid Ulrich / Jau Susanne / Krger Katja / Hofmann Andrea, “Term extraction with standard tools for corpus exploration - Experience from German”, in Proceedings of the TKE Conference, Frankfurt.

[Hockett 1954] “Two models of grammatical description”, Word, 10.

[Hotho/Staab/Stumme 2003] Andreas Hotho / Steffen Staab / Gerd Stumme, “Wordnet improves Text Document Clustering”, SIGIR 2003 Semantic Web Workshop.

[Hoelscher/Strube 2000] Christoph Hoelscher / Gerhard Strube, “Web search behavior of Internet experts and newbies”, WWW Conference.

[Hoelscher 2002] Christoph Hoelscher, *Die Rolle des Wissens im Internet. Gezielt suchen und kompetent auswählen*, Klett-Cotta.

[Hu/Liu 2004] M. Hu / B. Liu, “Mining and Summarizing Customer Reviews”, Proceeding of KDD.

[Jackendoff 1990] Ray Jackendoff, *Semantic Structures*, MIT Press.

[Jacquemin 1999] Christian Jacquemin, “Syntagmatic and paradigmatic representations of term variation”, Proceedings of the ACL on Computational Linguistics.

[Jacquemin 2001] Christian Jacquemin. *Spotting and Discovering Terms through NLP.* Cambridge MA: MIT Press.

[Jansen 2000a] Bernard J. Jansen, “The effect of query complexity on Web searching results”, Information Research, Vol. 6 No. 1, online at http://informationr.net/ir/6-1/paper87.html [1st of October 2006].

[Jansen 2000b] Bernard J. Jansen, “An Investigation Into the Use of Simple Queries On Web IR Systems”, online at http://cybermetrics.cindoc.csic.es/cybermetrics/pdf/163.pdf [1st of October 2006].

[Jansen et al 2000a] Bernard J. Jansen / Amanda Spink / Tefko Saracevic, “Real Life, Real Users, and Real Needs: A Study and Anaylsis of User Queries on the Web”, Information Processing and Management. 36(2): 207-227, online at http://ist.psu.edu/faculty_pages/jjansen/academic/pubs/ipm98/ipm98.pdf [1st of October 2006].

[Jansen et al 2000b] Bernard J. Jansen / Amanda Spink / Anthony Pfaff, “Linguistic Aspects of Web Queries”, American Society of Information Science 2000. Chicago.

[Joachims 2002] Thorsten Joachims, “Optimizing Search Engines using Clickthrough Data”, online at http://www.cs.cornell.edu/people/tj/publications/joachims_02c.pdf [1st of October 2006].

[Jones / Fain 2003] Rosie Jones, Daniel C. Fain, “Query word deletion prediction”, in ACM SIGIR Conference: 435-436.

[Justeson / Katz 1995] John S. Justeson / Slava M. Katz, “Technical Terminology: some linguistic properties and an algorithm for identification in text”, Natural Language Engineering 1: 9-27.

[Kamvar et al 2003] Sepandar D. Kamvar / Taher H. Haveliwala / Christopher D. Manning / Gene H. Golub , “Extrapolation Methods for Accelerating PageRank Computations”,Proceedings of the WWW-Conference 2003, Budapest, online at http://www.stanford.edu/ sdkamvar/papers/extrapolation.pdf [1st of October 2006].

[Kiryakov et al 2005] Atanas Kiryakov / Borislav Popov / Damyan Ognyanoff / Dimitar Manov / Angel Kirilov / Miroslav Goranov, “Semantic annotation, indexing, and retrieval”, Journal of Web Semantics, 2, Issue 1, online at http://www.ontotext.com/publications/SemAIR_ISWC169.pdf [1st of October 2006].

[Kleinberg 1998] J. Kleinberg, ‘Authoritative Sources in Hyperlinked Environment”, Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithm.

[Koehn / Knight 2003] Philipp Koehn / Kevin Knight, “Empirical Methods for Compound Splitting”, 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), online at http://people.csail.mit.edu/ koehn/publications/compound2003.ps [1st of October 2006].

[Kuhlen 1999] Rainer Kuhlen, “Die Konsequenzen von Informationsassistenten”, Frankfurt/Main.

[Kushal/Lawrence/Pennock 2005] D. Kushal / S. Lawrence / D. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews”, Proceedings of WWW.

[Lam/Pennock/Cosley/Lawrence 2003] S.K. Lam/ D.M. Pennock / D. Cosley / S. Lawrence, “1 Billion Pages = 1 Million Dollars. Mining the Web to Play “Who Wants to be a Millionaire”, Conference on Uncertainty in Artificial Intelligence.

[Landow 1997] George P. Landow, *Hypertext 2.0: The Convergence of Contemporary Literary Theory and Technology*, Johns Hopkins University Press.

[Langer 1996] Stefan Langer, *Selektionsklassen und Hyponymie im Lexikon. Semantische Klassifizierung von Nomina fr das elektronische Wörterbuch CISLEX*, CIS.

[Langer 1998] Stefan Langer, “Zur Morphologie und Semantik von Nominalkomposita”, Konvens.

[Latour 1991] Bruno Latour, “Technology is Society made durable”,in John Law (Hrsg.) A Sociology of Monsters. Essays on Power, Technology and Domination. London: Routledge.

[Law 1992] John Law, “Notes on the Theory of the Actor-Network: Ordering, Strategy and Heterogeneity”, Systems Practice 5: 379-393.

[Lee 1999] Lillian Lee, “Measures of Distributional Similarity”, in 37th Annual Meeting of the ACL, pp. 25-32, online at http://www.cs.cornell.edu/home/llee/papers/cf.pdf [1st of October 2006].

[Lee/Liu/Cho 2005] Uichin Lee / Zhenyu Liu / Junghoo Cho, “Automatic Identification of User Goals in Web Search”, online at http://www2005.org/cdrom/docs/p391.pdf [1st of October 2006].

[Leidner 2004] Jochen Leidner, “Toponym Resolution in Text: Which Sheffield is it?”, Proceedings of the 27th Annual International ACM SIGIR Conference (SIGIR 2004).

[Lewandowski 2005] Dirk Lewandowski, *Web Information Retrieval: Technologien zur Informationssuche im Internet*, DGI.

[Lewandowski 2006] Dirk Lewandowski, “Themen und Typen der Suchanfragen an deutsche Web-Suchmaschinen”, in: Lehner, Franz; Nsekabel, H.; Keinschmidt, P. [Hrsg.]: *Multikonferenz Wirtschaftsinformatik 2006* (MKWI ’06), Proceedings 20. - 22. Februar 2006, Universitt Passau. Berlin: Gito [Lecture Notes in Informatics], p. 33-43, online at http://www.durchdenken.de/lewandowski/doc/mkwi2006.pdf [1st October 2006].

[Lin / Zhao 2003] Dekang Lin / Shaojun Zhao, “Identifying Synonyms among Distributionally Similar Words”, online at http://www.cs.ualberta.ca/ lindek/papers/polarity.pdf [1st of October 2006].

[Lucas / Topi 2002] Wendy T. Lucas, Heikki Topi, “Form and function: The impact of query term and operator usage on Web search results”, JASIST 53(2): 95-108, online at http://www.ececs.uc.edu/ annexste/Courses/cs690/formandfunction.pdf [1st of October 2006].

[Machill/Welp 2003] Wegweiser im Netz, ed. Marcel Machill / Carsten Welp. Bertelsmann Stiftung.

[Mallet / Willmott 2003] James Mallet / Keith Willmott, “Taxonomy: renaissance of Tower of Babel?”, Trends in Ecology and Evolution 18: 57-59, online at http://www.ucl.ac.uk/taxome/jim/pap/mallet03tree.pdf [1st of October 2006].

[Mann 2002] Thomas M. Mann, *Visualization of search results from the World Wide Web*, PhD dissertation, Department of Computer and Information Science, Universitat Konstanz, online at http://www.ub.uni-konstanz.de/kops/volltexte/2002/751/ [1st of October 2006].

[Martin/White 2005] J. Martin / P. White, The Language of Evaluation: The Appraisal Framework, Palgrave Macmillan.

[Martinez-Santiago/Montejo-Raez/Urena-Lopez/Diaz-Galiano 2003] Fernando Martinez-Santiago / Arturo Montejo-Raez / L.Alfonso Urena-Lopez / Manuel Carlos Diaz-Galiano, “SINAI at CLEF 2003: Decompounding and Merging”, Working Notes for the CLEF 2003 Workshop 21-22 August.

[Markoff/Hansell 2006] John Markoff / Saul Hansell, “Hiding in Plain Sight, Google Seeks More Power” The New York Times, June 14.

[McDonald 1996] David D. McDonald, “Internal and External Evidence in the Identification and Semantic Categorization of Proper Names”, Corpus processing for lexical acquisition. Hrsg. B. Boguraev/J. Pustejovsky. Cambridge, MA: 21-39, online at http://acl.ldc.upenn.edu/W/W93/W93-0104.pdf [1st of October 2006].

[McGuinness et al 2000] Deborah L. McGuinness / Richard Fikes / James Rice / Steve Wilder, “An Environment for Merging and Testing Large Ontologies”, in Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000), Breckenridge, Colorado, USA, online at http://dit.unitn.it/ accord/RelatedWork/Matching/McGuinnessKR.pdf [1st of October 2006].

[Mehta et al.] Aranyak Mehta / Amin Saberi / Umesh Vazirani / Vijay Vazirani. “AdWords and Generalized On-line Matching”

[Merkel / Andersson 2000] Magnus Merkel / Mikael Andersson, “Knowledge-lite extraction of multi-word units with language filters and entropy thresholds”, in Proceedings of Conference on User-Oriented Content-Based Text and Image Handling (RIAO’00), pages 737–746, Paris, France, online at http://www.ida.liu.se/ magme/publications/merkel-andersson-riao-2000.pdf [1st of October 2006].

[Merz 2002] Michael Merz, *E- Commerce und E- Business. Marktmodelle, Anwendungen und Technologien*, 2nd Edition. dpunkt- Verlag.

[Mikheev 1999] Andrei Mikheev, “Periods, capitalized words, etc”, Computational Linguistics 28(3): 289-318.

[Miller / Walter 1991] George Miller / Charles Walter, “Contextual Correlates of Semantic Similarity”, Language and Cognitive Processes 6(1): 1-28.

[Miller/Albert/Lam/Konstan/Riedl 2003] B. Miller / I. Albert / S. Lam / J. Konstan / J. Riedl, “MovieLens unplugged: Experiences with an occasionally connected recommender system”, Proceedings of the ACM Conference on Intelligent User Interfaces.

[Morinaga/Yamanishi/Tateishi/Fukushima 2002] S. Morinaga / K. Yamanishi / K. Tateishi / T. Fukushima, “Mining product reputations on the web”, Proceedings of KDD.

[Nenadic et al 2002] Goran Nenadic / Irena Spasic / Sophia Ananiadou, “Automatic Discovery of Term Similarities Using Pattern Mining”, in Proceedings of CompuTerm, Taipei, Taiwan, pp. 43-49, online at http://acl.ldc.upenn.edu/W/W02/W02-1408.pdf [1st of October 2006].

[Novak/Raghavan/Tomkins 2004] Jasmine Novak / Prabhakar Raghavan / Andrew Tomkins, “Anti-Aliasing on the Web”, WWW Conference.

[Papa 2006] Steve Papa, “Searching The Long Tail”, Search Engine Meeting, April, online at http://www.infonortics.com/searchengines/sh06/slides/endeca.pdf [1st of October 2006].

[Pazienza 2003] *Information Extraction in the Web Era: Natural Language Communication for Knowledge Acquisition and Intelligent Information Agents*, ed. Maria Teresa Pazienza. Springer.

[Pindyck/Rubinfeld 2004] Robert S. Pindyck / Daniel L. Rubinfeld, *Microeconomics*, 6th Edition. Prentice-Hall.

[Popescu/Etzioni 2005] A. Popescu / O. Etzioni, “Extracting Product Features and Opinions from Reviews”, Proceedings of HLT-EMNLP.

[Radev / Qi / Zheng / Blair-Goldensohn / Zhang / Fan / Prager 2001] Dragomir R. Radev / Hong Qi / Zhiping Zheng / Sasha Blair-Goldensohn / Zhu Zhang / Weiguo Fan / John Prager, "Mining the Web for Answers to Natural Language Questions", online at http://tangra.si.umich.edu/ radev/papers/cikm01.pdf [1st of October 2006].

[Raghavan / Wong 1986] Vijay V. Raghavan / S.K.M.Wong, "A Critical Analysis of Vector Space Model for Information Retrieval", Journal of the Americal Society for Information Science, 37(2):279-287, online at http://www-ufrima.imag.fr/FORMATION/FILIERE/ MASTER/SI/SiteMasterSI/Documents/MORI/raghavan.pdf [1st of October 2006].

[Rashtchy/Avilio 2003] S. Rashtchy / J. Avilio, "The Golden Search: Dynamics of the Online Search Market and the Scope of Opportunity." U.S. Bancorp Piper Jaffray.

[Ravin / Wacholder 1996] Ravin Yael / Nina Wacholder, "Research Report: Extracting Names from Natural-Language Text", Research Report 20338. IBM Corporation.

[Rayson / Garside 2000] Paul Rayson and Roger Garside, "Comparing corpora using frequency profiling", in proceedings of the workshop on Comparing Corpora: 1-6, online at http://www.comp.lancs.ac.uk/computing/users/paul/publications/rg_acl2000.pdf [1st of October 2006].

[Rayson et al 2004] Paul Rayson / Damon Berridge / Brian Francis, "Extending the Cochran rule for the comparison of word frequencies between corpora", in proceedings of the 7th international conference on statistical analysis of textual data (JADT 2004): volume II, Louvain-la-Neuve, Belgium, Presses Universitaires de Louvain: 926-36, online at http://www.comp.lancs.ac.uk/computing/users/paul/publications/rbf04_ jadt.pdf [1st of October 2006].

[Resnik/Iacovou/Suchak/Bergstrom/Riedl 1994] P. Resnick / N. Iacovou / M. Suchak / P. Bergstrom / J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews", Proceedings of ACM Conference on Computer Supported Cooperative Work.

[Rieh / Xie 2006] Soo Young Rieh / Hong Iris Xie, "Analysis of multiple query reformulations on the web: The interactive information retrieval context", Information Processing Management 42(3): 751-768 (2006).

[Riloff 1993] Ellen Riloff, "Automatically constructing a dictionary for information extraction tasks", Proceedings of the 11th National Conference on Artificial Intelligence.

[Riloff/Jones 1999] Ellen Riloff/R. Jones, "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping", 16th AAAI Conference.

[Roche/Schabes 1997] , E. Roche / Y. Schabes, *Finite-State Language Processing*, The MIT Press.

[Rose/Levinson 2004] Daniel E. Rose / Danny Levinson, “ Understanding user goals in web search”, WWW Conference.

[Sarwar/Karypis/Konstan/Ried 2001] B. M. Sarwar / G. Karypis / J. A. Konstan / J. Riedl, “Item-based collaborative filtering recommendation algorithms”, WWW Conference.

[Scholer / Williams 2002] Falk Scholer / Hugh E. Williams, “Query Association for Effective Retrieval”, online at http://www.seg.rmit.edu.au/research/download.php?manuscript=17 [1st of October 2006].

[Schonfeld 2005] Erick Schonfeld, “The Flickrization of Yahoo”, *Business 2.0*, online at http://www.business2.com/b2/web/articles/0,17863,1129448,00.html [1st of March, 2006].

[Senellart 1998] Jean Senellart, “Locating Noun Phrases with Finite State Transducers”, COLING98.

[Seuren 1997] Peter Seuren, *Western Linguistics: An Historical Introduction*, Blackwell Publishers.

[Shen / Zhai 2003] Xuehua Shen / ChengXiang Zhai, “Exploiting Query History for Document Ranking in Interactive Information Retrieval”.

[Silverstein/Henzinger/Marais/Moricz 1998] Craig Silverstein / Monika Henzinger / Hannes Marais / Michael Moricz, “Analysis of a Very Large Alta Vista Query Log”, SRC Technical Note, 1998-014, online at http://www.hpl.hp.com/techreports/Compaq-DEC/SRC-TN-1998-014.pdf [1st of October 2006].

[Sinha 2005] Rashmi Sinha, “A cognitive analysis of tagging”, online at http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html [1st of October 2006].

[Spink et al 2002] Amanda Spink / Bernard J. Jansen / Dietmar Wolfram / Tefko Saracevic, “From E-Sex to E-Commerce: Web Search Changes”, IEEE Computer, 35(3): 107-109, online at http://blog.namics.com/2006/IEEEComputer2002.pdf [1st of October 2006].

[Staab et al 2001] Steffen Staab, Alexander Maedche, Siegfried Handschuh, “An Annotation Framework for the Semantic Web”, in Proceedings of the First Workshop on Multimedia Annotation, Tokyo.

[Stainton 2006] Robert Stainton, *Words and thoughts*, Oxford University Press.

[Stevenson / Gaizauskas 2000] Mark Stevenson / Robert Gaizauskas, “Using Corpus-derived Name Lists for Named Entity Recognition”, ANLP-NAACL, Seattle: 290-295.

[Stumme et al 2006] Gerd Stumme / Andreas Hotho / Bettina Berendt, “Semantic Web Mining: State of the art and future directions”, Journal of Web Semantics, 4(2): 124-143, online at http://www.kde.cs.uni-kassel.de/stumme/papers/2006/stumme2006semantic.pdf [1st of October 2006].

[Surowiecki 2004] J. Surowiecki, The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations, Little, Brown.

[Thompson/Hunston 2000] “Evaluation: An Introduction.” Evaluation in Text: Authorial Stance and the Construction of Discourse, ed. S. Hunston, and G. Thompson, Oxford University Press.

[Turney 2002] P. D. Turney, “Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews” Proceedings of ACL.

[Turney 2006] Peter D. Turney, “Similarity of Semantic Relations”, online at http://cogprints.org/5098/01/NRC-48775.pdf [1st of October 2006].

[Tzoukermann et al 1997] Evelyne Tzoukermann, Judith Klavans, Christian Jacquemin, “Effective Use of Natural Language Processing Techniques for Automatic Conflation of Multi-Word Terms: The Role of Derivational Morphology, Part of Speech Tagging, and Shallow Parsing”, SIGIR 1997.

[Valiente 2001] Gabriel Valiente, “An Efficient Bottom-Up Distance between Trees” Proceedings of the 8th International Symposium on String Processing and Information Retrieval.

[Volz et al 2004] Raphael Volz / Siegfried Handschuh / Steffen Staab / Ljiljana Stojanovic / Nenad Stojanovic, “Unveiling the hidden bride: Deep Annotation for Mapping and Migrating Legacy Data to the Semantic Web”, Journal on Web Semantics. 2. 187-206, online at http://www.uni-koblenz.de/ staab/Research/Publications/2004/final-hidden-bride.pdf [1st of October 2006].

[Wang 2003] Jiying Wang, “Information Discovery, Extraction and Integration for the Hidden Web”, online at http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-76/jwang.pdf [1st of October 2006].

[Wanner 1996] *Lexical Functions in Lexicography and Natural Language Processing*, ed. Leo Wanner, John Benjamins.

[Weeds/Weit/McCarthy 2004] J. Weeds / D. Weir / D. McCarthy, “Characterising measures of lexical distributional similarity” Proceedings of CoLing 2004.

[Wilson/Wiebe/Hoffmann 2005] T. Wilson / J. Wiebe / P. Hoffmann, “Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis”, Proceedings of HLT-EMNLP.

[Wilson/Wiebe/Hwa 2004] T. Wilson / J. Wiebe / R. Hwa, “Just how mad are you? finding strong and weak opinion clauses”, Proceedings of AAAI.

[Wirth 1999] Selektion im Internet. Empirische Analysen zu einem Schlüsselkonzept, ed. Werner Wirth / Wolfgang Schweiger, Westdeutscher Verlag.

[Witten/Paynter/Frank/Gutwin/Nevill-Manning 1999] I. H. Witten / G. W. Paynter / E. Frank / E. Gutwin / C. G. Nevill-Manning, “Kea: Practical automatic keyphrase extraction”, Proceedings Fourth ACM Conference on Digital Libraries.

[Yun/Chen 2000] C.H Yun / M.S. Chen, Mining Web Transaction Patterns in an Electronic Commerce Environment, in Proceedings of the 4th Pacific-Asia Conf. on Knowledge Discovery and Data Mining, April 2000.

[Zeng/He/Chen/Ma 2004] H. Zeng / Q. He / Z. Chen / W. Ma, “Learning To Cluster Search Results”, The 27th Annual International ACM SIGIR Conference.

[Zhang / Lee 2005] Dell Zhang / Wee Sun Lee, “Learning to integrate web taxonomies”, Journal of Web Semantics, 2(2): 131-151, online at http://www.comp.nus.edu.sg/ leews/publications/dellzhang_ws2004.pdf [1st of October 2006].

Lebenslauf Gerhard Rolletschek

Adresse:

Jahnstr. 12
85221 Dachau
Tel: 08131 52592
E-Mail: gerhard@rolletschek.com

Geburt	10. Juli 1978 in München
Nationalität	Deutsch
Familienstand	verheiratet, eine Tochter

Ausbildung:

1984-1988	Grundschule Markt Indersdorf
1988-1997	Josef-Effner-Gymnasium Dachau, Abitur
1997-2002	Studium der Germanistik, Geschichte und Informatik an der LMU München und der TU München
Apr 1999	Zwischenprüfung
Juli 2000	Wechsel zum Magisterstudium (HF: Deutsche Sprache und Literatur des Mittelalters, NF1: Mittelalterliche Geschichte, NF2: Neuere deutsche Literatur)
Februar 2002	M. A.
2002 - 2003	Aufbaustudium Computerlinguistik am Centrum für Information und Sprache CIS, LMU München
Seit WS 2003/2004	Promotionsstudium Computerlinguistik Nebenfach: Deutsche Sprache und Literatur des Mittelalters Thema der Dissertation: „Term-driven E-Commerce“
Seit Okt 2004	Wissenschaftlicher Angestellter am Centrum für Information und Sprache CIS, LMU München

München, den 2. Oktober 2006