Why Not Use Query Logs As Corpora?

Olena Medelyan
University of Freiburg, Germany[1]
medelyan@coling.uni-freiburg.de

Abstract. Generally, every Web search engine logs the user sessions. These records, called query logs, contain valuable information about the behaviour of Internet users and their language. There are only a few experiments on mining query logs, but they confirm that query logs are very useful for designing natural language applications in Web retrieval. This paper shows how lexical and semantic information can be extracted from query logs using statistical methods. I first summarize approaches in query log processing and mining for different purposes. After a short description of the used query logs, I present new domain- and language-independent methods for generating a compound dictionary and extracting semantically similar terms. The evaluation will shed light on the quality of proposed methods and show that the results are good enough to be directly integrated in query processing and improve information retrieval on the Web.

Keywords: Web retrieval, query processing, text mining, information retrieval

1 Introduction

Since the Internet has become one of the most popular information sources and search engines are necessary for navigation in the World Wide Web, query logs are now extremely valuable for information acquisition. Simple statistics about the most frequent queries easily disclose the demands of users in all areas from technology development to the music industry. Deeper mining into queries can reveal more important information about search engine users and their language use. Despite the large interest in pattern extraction and statistics on query logs, results are still dissatisfactory. Interesting approaches like correlation analysis for extracting compounds and collocations (Silverstein et al. 1998), query clustering (Beeferman and Berger 2000), extraction of correlating terms for query extension (Cui et al. 2002) and transforming query phrases for better question answering on the Web (Agichtein et al. 2001) have shown that query logs can compete with conventionally used corpora like newspaper articles. Furthermore, they can be used for the automatic generation of natural language resources, which are perfectly adapted for Web retrieval.

The main part of this paper presents my experiments in query log mining. One of the most important steps of query processing is the segmentation of the query into terms. Many user queries contain compounds (e.g. "*[green card] for a job in the [united states]*"). Tokenizing the query without considering compounds and phrases leads to the loss of the query sense. In Section 4.1, I describe a technique for the automatic construction of a compound dictionary, leaned against decomposition for German and Dutch presented by Chen (2002). This dictionary will be used in the query preprocessing step of the approach in automatic detection

[1] Work partially done during an internship at exorbyte GmbH (http://www.exorbyte.com) in Winter 2003/04

Proceedings of the Ninth ESSLLI Student Session
Paul Egré and Laura Alonso i Alemany (editors)

of query term similarities. I developed a statistical technique, which uses co-occurrences of query terms in the query log to determine their semantic distance. Section 4.2 shows the creation of a similarity dictionary in detail. After the evaluation of the results in Section 5, I will discuss the advantages of query logs and the reasons why they should be better analyzed in the future (Section 7).

2 Approaches in Query Log Mining

Because the search engine operating companies do not want to disclose proprietary information, there are still very few publications about query log analysis. Nonetheless, statistical analysis on query logs is very important not only to understand how human use search engines to find information they are interested in, but also to reveal new information from the search requests. Some approaches have shown that even simple counting of queries and query terms can describe behaviour of search engine users (see Jansen et al. (1998), Silverstein et al. (1998) and Cacheda and Via (2001)). E.g. Silverstein et al. (1998) demonstrate that user queries are very short (in average 2.35 terms per query), which makes sophisticated natural language analysis, as needed in standard IR (e.g. information requests at TREC), unessential for Web retrieval.

More detailed approaches were made using query data collected from large-scale engines like AltaVista (Silverstein et al. 1998), Lycos (Beeferman and Berger 2000), Encarta (Cui et al. 2002) and (Wen et al. 2002). Besides user queries, some of these query logs contain other components such as "clickthrough data" (query and the URL, which the user selected from among other offered candidates for this query), result screens and submitter information. These data allow graph-based techniques for query and document clustering (Beeferman and Berger 2000), (Cui et al. 2002): related query terms and URLs are identified by their co-occurrance in the clickthrough data. Another interesting approach for query clustering presented in Wen et al. (2002), combines cross-references between users' queries and the documents they clicked on with similarities between query terms. The term similarities are computed with a cosine correlation function, applied to terms weighted with TF*IDF. Agichtein et al. (2001), de Lima and Pedersen (1999) and Silverstein et al. (1998) use pure user queries as a corpus, without considering additional features common corpora do not possess. The latter used a Chi-squared test for correlation analysis of the most frequent 10,000 query terms and yielded phrases such as "*cindy crawford*", "*visual basic*" a. o. The combination of a part-of-speech tagger and a query grammar (a context free grammar with 300 rules) in de Lima and Pedersen (1999) detects phrases like "*free java games*", "*history of stock market*" and "*howard m. dean*". In the next step, these phrases were transformed into a short, more precise form ("*free games*"). The evaluation demonstrated that this technique can improve the average precision of top ranked results.

Thus, elaborated statistics applied on large query logs with millions of different queries can be used to extract linguistic information about the language of the Internet users. These data enable the improvement retrieval methods.

3 Query Logs Used in this Project

In this project I used query logs in three languages from two different search engines. The first query log (henceforth query log EN) comes from a big commercial search engine in Great

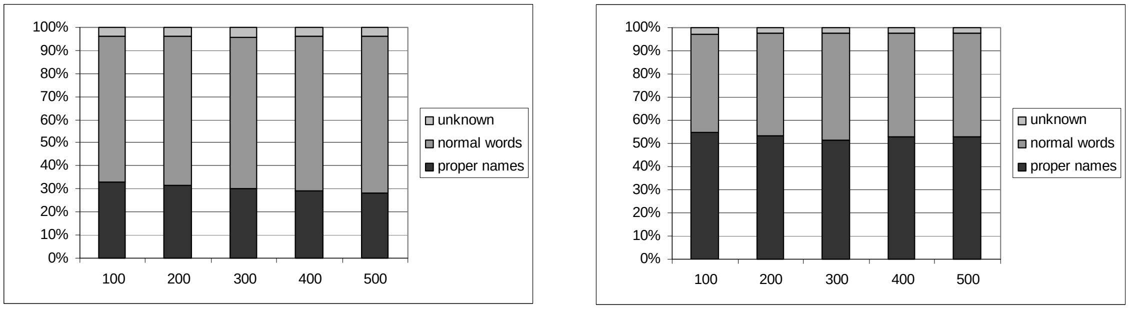

Figure 1.1: Fraction of proper nouns terms in the query log DE (left) and EN (right)

Britain and contains mostly English queries (total number of queries is 71 Mill. where 19,8 Mill. are unique). The other query logs stem from an international paid-listings-provider with different domains for Germany and the Netherlands. The German query log (DE) contains 5,8 Mill. queries where 874,000 are unique. The query log from the Netherlands (NL) contains mostly Dutch queries (totalling 14,7 Mill. queries, 2 Mill. unique). The reason for using query logs in different languages is to show that the presented approaches are language-independent. Unlike the query logs used in Cui et al. (2002), Silverstein et al. (1998), that contain several components (sessions cookies, submitter information etc.), I only had access to pure user queries, which were collected over a period of one month.

On each query log, I applied the statistical methods proposed in Silverstein et al. (1998). The results of these statistics (Table 1.1 and 1.2) demonstrate that the distribution of query length in terms in the query log EN has a similar structure to the Alta Vista query log described in Silverstein et al. (1998), while results for query logs DE and NL are similar to those presented in Wen et al. (2002). These differences can be explained by the nature of the search engines: The main purpose of paid-listings-providers and Encarta is to search for products or word definitions and facts. Common search engines like Alta Vista are more popular and are used to search for every day needs, which are mostly expressed in phrases. These three query logs have in common, that the distribution of query frequencies is in all three query logs very similar (Table 1.2). Furthermore, in all query logs a few queries cover a significant portion of the whole log. The 25 most frequent queries cover 3.07% of all queries, although these are only 0.00013% of all different queries (query log EN). These queries represent the most frequently asked categories or subject areas Internet users are interested in: search engines (*google, yahoo, ask jeeves*), email (*hotmail, msn*), flights (*cheap flights, easyjet*), sex (*sex, porn*), games, news (*bbc, weather*) etc.

Since English has become the world language, a lot of English words are internationally understood and used (e.g. *weekend* or *software*). It is interesting to see, that about 24% of queries in the query log DE contain English terms ("*small business directories*", "*beauty*", "*steel*"), while only 72% of queries consist of German terms. This ratio was computed on a sample of 500, randomly selected from the 10,000 most frequent queries.

Another interesting data ascertainment concerns the frequency of proper nouns in the query logs, such as brands and company names ("*coca cola*", "*e plus*"), geographical names ("*south african airways*", "*volksbank hoogstede*"), song, movie, tv-show titles ("*once upon a time in mexico*", "*wer wird millionr*", "*gzsz*") etc. These statistics were computed for the German sample set mentioned above and additionally for the sample set from query log EN, created in the same way. Independent from the sample size (100, 200 ... 500 queries), the portion of proper nouns vs. nouns remains constant. As shown in Figure 1.1, query log EN consists of an average of about 45% proper nouns, while the query log DE has noticeably more general nouns (an average of 68%).

Query Log	**Query Frequency**						
	1	**2**	**3**	**>3**	**max**	**avg**	**stdev**
DE	60.8%	17.1%	6.5%	15.7%	14,889	6.62%	63.09%
NL	63.2%	15.7%	6.8%	14.3%	40,689	7.51%	112.22%
EN	61.3%	18.2%	7.2%	13.3%	177,535	3.59%	99.59%

Table 1.1: Query Frequencies in Query Logs DE, NL and EN

Query Log	**Terms per Query**						
	1	**2**	**3**	**>3**	**max**	**avg**	**stdev**
DE	66.3%	26.8%	5.0%	1.9%	31	1.06%	0.94%
NL	59.6%	26.0%	9.7%	4.7%	31	1.53%	1.50%
EN	18.3%	34.2.2%	24.8%	22.6%	177,535	57.00%	1.68%

Table 1.2: Query length in terms in Query Logs DE, NL and EN

4 Description of the Experiments

As shown in Section 2, there are a lot of possibilities for query log mining: query clustering, phrase detection, contextual analysis, pattern mining for information extraction or just simple statistical analysis about, for example, new trends in the mobile phone industry or newly emerging terms. The ways to achieve these aims are also very different. One can use only the query log itself or with help of an additional corpus (e.g. webpages). Statistical analysis can be supported by dictionaries or NLP-tools such as stemmers, taggers and chunkers, but most of them are not disposable and the development of new tools is expensive. The main goal of this work was to develop an automatic method for creating a similarity dictionary, which had to be integrated into the query processing step of a search engine. It had to be done in short terms, and it was also important not to use information sources other than the query log itself. Furthermore, the technique should be domain and language independent. For these reasons I mainly used statistical analysis.

4.1 Compounds Extraction

Almost every analysis of a query log needs a good technique for detecting the terms a query consists of. Simple splitting at blanks is disadvantageous, because query logs contain a lot of compounds or multi-word expressions[2] such as "*david blaine*", "*world cup*", "*bed and breakfast*", "*who wants to be a millionaire*" (cf. Section 3). Furthermore, new compounds like product names or song titles are invented every day. Hence, compounds can not be covered by a fixed dictionary and should thus be computed dynamically. Because query logs frequently contain queries in different languages, the method for automatic compound extraction should be language independent. The first trial was to include in the compounds dictionary all phrases consisting of two or more words, which exceed a choosen co-occurrence frequency threshold. This had several disadvantages:

[2] The definitions for multi-word items used in the literature diverge. A very detailed overview and classification of such sequences is provided in (Guenthner and Blanco). The definition of compounds I suggest is described in Section 5.

1. Seldom, but high correlated phrases (e.g. "*birmingham international airport*" or "*kingston upon hull*") were not found.
2. Instead, low correlated but frequent phrases (e.g. "*matrix cheats*") were detected by mistake as compounds.
3. The actual length of a phrase was not considered, so that erroneous compounds were extracted:
   (a) "*red hot chilli*" - instead of "*red hot chilli peppers*"
   (b) "*dido life for rent*" - instead of "*dido*" and "*life for rent*"
   (c) "*britney spears me against*" - instead of "*britney spears*" "*me against the music*"

That means that in some cases expanding (a), in some cases decompounding (b) or both (c) is necessary. Due to this, the following steps were applied in the final version of the algorithm:

1. **Expanding:** If a phrase P1, consisting of x words, is a subphrase of another phrase P2, consisting of x+1 words, P1 will be substituted by P2 in a collection of expanded phrases.
2. **Decompounding** (as shown in Chen (2002) and Martnez-Santiago et al. (2003)): All expanded phrases will be split according to the following rules:
   (a) If a phrase is in a base form dictionary, it will not be split.
   (b) The shortest decomposition (with a minimal number of base forms) will be chosen.
   (c) In case of several possible decompositions, the one with the highest probability will be chosen. The probability of a decomposition is the sum of probabilities of all base forms.

The base form dictionary used in the decompounding step consists of all terms of the query log, which utilises a hidden advantage of a query log: some users write high correlated phrases as one word (e.g. "*matrixreloaded*", "*thematrix*", "*enterthematrix*"). Low correlated phrases (e.g. "*matixcheats*") are seldom or never written as one word. Additionaly, one can add known compounds extracted from any electronic dictionary (e.g. WordNet) to the base form dictionary to improve its quality. To keep my method language independent, I didn't use any additional sources.

4.2 Extracting Semantically Similar Word Pairs

In this section I present experiments in statistical extraction of semantically similar terms. The algorithm takes the 10,000 most frequent terms of the query log as input and estimates the similarity value for each term pair, due to their co-occurrence behaviour in the log. The output is a similarity dictionary: a ranked list of term pairs, which exceeded the similarity threshold. The method is based on the idea that semantic distance between two words depends on their contextual interchangeabilty (Miller and Walter 1991). The more contexts two words have in common, the shorter is the semantic distance between these words.

Each query with more than one term can be seen as a 2-tuple $\{$*query term, subquery*$\}$, where a subquery is the remaining part of the query and therefore the context of the given query term. E.g. a query "*cheap car hire london*" contains three tuples: $\{$*cheap, X car_hire london*$\}$, $\{$*car_hire, cheap X london*$\}$, $\{$*london, cheap car_hire X*$\}$, where X is a placeholder for the query term. For each term pair, which co-occurs with a minimum of common subqueries, I analyse the total number of their common subqueries and the characteristics of these terms. This analysis includes estimating of the relevance of each subquery, which is affected by several factors:

a) Number of subquery terms: The more query terms a subquery contains, the higher its contextual information content is.

b) Occurrences in the query log: The more distinctive terms co-occur with a subquery in the log, the lower its relevance is. Very frequent subqueries like *free*, *online* or *download* co-occur with a lot of terms from almost all areas of interest and are therefore irrelevant for similarity estimation.

c) Terms of a subquery: Not all terms definitely indicate to which subject area a query belongs. While a term like "*flights*" indicates an affiliation of the query to the domain [*flights, airlines, planes*], subqueries like "*information on*" or "*find me*" are non-distinctive and do not support the assumption of similarity. The more such irrelevant suqueries could be found, the better the results are.

d) Order of the terms: Though the language used in a query log does not conform to grammatical rules, there are still many queries that form complete or partial noun phrases: "*facts about britney spears*", "*cheap flights*", "*property for sale in france*" etc. Therefore, it is obviously good to consider the order of terms while analysing the contexts.

Due to these heuristics, the information content for each subquery in the query log was computed. Then the characteristic of each query term could be estimated as a sum of frequencies of all queries, where the term appear, weighted by the information content of the subquery. This value was used to determine the direction of the semantic relation between terms. In case if the relation is asymmetric (e.g. hypo- and hyperonymy), the more generic term always has a higher characteristic value than its species. I used the Dice-coefficient (Manning and Schütze 1999) to compute the overlap O of two terms A and B in the query log:

$$O_{(A,B)} = \sum_x min(f'(Ax), f'(Bx)) \tag{1.1}$$

$f'(Ax)$ is here the frequency of the query Ax, weighted by the information content of the subquery x. The semantic distance or similarity penalty $PEN_{(A \rightarrow B)}$ is estimated for both terms as:

$$PEN_{(A \rightarrow B)} = \frac{O_{(A,B)}}{F_{(A)}} \tag{1.2}$$

F(A) is here the characteristic of the term A, as described above. This is the final cost function for substitution of the specific term by the generic term usable in a query extension application. The higher the overlap of A and B is, the lower the penalty will be. In case of exact synonymy or equivalence I expect to get $PEN = 0$ in both direction.

5 Evaluation

The most established evaluation criteria in the IR are recall and precision. They describe the ratio of relevant found results to the answer set (precision) or to all relevant results in the document set (recall). Computing recall means extraction of all possible compounds and semantically similar words from the query log manually, which is a very complex task. However, the shortening of the query log would dramatically lower the significance of the

Query log	Compound	R1	R2
EN	air conditioning	1	1
	distance learning	1	0
	self build	0	1
	yu gi oh	-1	-1
DE	lueneburger heide	1	1
	geld verdienen	1	1
	x2 die bedrohung	1	-1
	uni duesseldorf	1	0
	rund um	0	1
NL	vroom en dreesman	1	–
	zonne energie	1	–
	s hertogenbosch	0	–

Table 1.3: Examples from the evaluation table for extracted compounds

statistics. In order to compute the precision, I only need to evaluate word pairs returned by the algorithm, which is easier and more relevant for my purposes.

The complete evaluation was made on sample sets extracted from the top-ranked 1000 results. In order to evaluate the extraction of compounds from the query log EN and DE, two sample sets each composed of 300 compounds were rated by two native speakers. To compute the inter-rater reliability (IRR) in each set the randomly selected compounds were the same. Because the concept *compounds* is vague, all raters got brief instructions based on the definitions in Quirk et al. (1985). The following is the extract from these instructions:

Instructions for the Evaluation of Compounds

- Definition: *A compound is a lexical unit consisting of more than one base and functioning both grammatically and semantically as a single word (concept).* E.g.: *washing machine*, *hot dog.*
- Compounding can take place within any of the word classes resulting above all new nouns and, to a lesser extent, adjectives.
- Although both bases in a compound are in principle equally open, they are normally in a relation whereby the first is modifying the second. E.g.: *pine tree*, *meat delivery*, *language teacher.*
- In contrast to noun phrases, in English, compounds have primary stress on the first constituent. E.g.: *a 'blackbird* vs. the phrase *a ,grey'bird.*
- Consider all multi-word-expressions and proper names (locations, names, brands, song titles etc.) as good compounds. E.g.: *las vegas, st petersburg, michael jackson, windows xp.*
- Don't pay attention to bad spelling. If an expression should be written in one word, it is also a good compound. The main idea is to extract all expressions, which may not be split.
- Judge each compound with 1 (good), 0 (bad) and -1 (unknown).

The evaluation yields precision of 89.2 by IRR of 83.9 for English and precision of 87.9 by IRR=88.3 for German. Only small test with one native-speaker and two samples with 100 results each was made for the Dutch query log, where the precision of 94.6 was reached. To show the typical quality of the results, I present an extract of the evaluation in Table 1.3.

Unfortunately there is no sufficient test base for the field of semantic similarity estimation. Most authors compare their results with human judgements (only 30 word pairs) as published in Miller and Walter (1991). To evaluate my approach, I created an own test set consisting of 100 word pairs, which were randomly selected from the top-ranked 1000 similar pairs extracted

			Question 1				Question 2			
Word 1	Word 2	Query	R1	R2	R3	R4	R1	R2	R3	R4
volkswagen	vw	X campervan sale	1	1	1	1	1	1	1	1
erotic	sexy	X chinese girls	1	1	0	1	1	1	1	1
cds	dvds	cheap blank X	1	0	0	0	1	1	0	0
spares	accessories	kenwood chef X	0	0	1	0	1	0	1	0
revision	coursework	a2 psychology X	0	1	0	0	0	1	0	0
fiesta	escort	X body kit	0	0	0	0	0	0	0	0
accessories	reviews	suzuki jimny X	0	0	0	0	0	0	0	1

Table 1.4: Examples from the evaluation table for extracted similar words

from the query log EN. Four native English speakers (R1 - R4) were then asked to rate this set. Because the similarity depends on context, each word pair was provided with a randomly selected query (cf. Table 1.4). The rater had to answer the two following questions with yes (1) or no (0): "*Is the sense of the query by using both words instead of X the same?*" and "*If you used the query with the first word, would you be interested to get results from the query with the second word?*". This interchangeability test refers to the works described in Miller and Walter (1991). To compute the precision I considered only those pairs as similar, which have passed the test by more than two judgers. The first question yielded 60% precision, and the average inter-rater reliability was here 78.2%. The second question was approved by 87% of word pairs, with the average IRR of 84%. I didn't evaluated the German and the Dutch query logs, because as described in Section 3, only a few queries consist of more than one term. Therefore, they do not provide a good co-occurrences source as the English query log does and will probably yield worse results. Table 1.5 contains some examples in order to show the typical quality of my results as proposed in Lin (1998). Numbers in brackets are penalty scores on a scale from 0 (identical) to 30 (low similarity).

Query log	**Part of speech**	**Query term**	**Extracted similar words**
EN	Noun	hotels	b&b (2), guest houses (3), motels (4), bed breakfast (6), guest house (6), inns (7), resorts (14), accommodation (14), inn (18), flights (28), holidays (28), villas (29)
	Noun	jobs	vacancy (3), careers (7), job vacancies (2), recruitment (16), employment (17), qualifications (17), agencies (20), courses (23), equipment (23), training (24), working (26), recruitment (28)
	Adjective	naked	pictures (11), nude pics (13), topless (14), nude (15), pictures (19), pics (23), sexy (24)
	Verb	apply	application (9), applying (10)
DE	Noun	ferienhaus [*summer cottage*]	ferienwohnung [*holiday flat*] (16), ferienhaeuser [*summer cottages*] (18), urlaub [*holiday*] (18), hotel (20), immobilien [*real estate*] (20)
	Adverb	neuwertig [*as good as new*]	ovp [*still in package*] (10), top (14)
NL	Noun	vakantie [*vacation*]	vakanties [*vacations*] (5), vliegreizen [*air trip*] (6), vakantiepark [*holiday village*] (8), last minute (9), bungalowpark (10), campings (10), vliegen [*travel by air*] (10), reizen [*travel*] (11), hotels (15)
	Adjective	naakt [*naked*]	bloot [*nude*] (6), playboy (10), sex (10), naked (12), nude (13), naakte (15), geile [*horny*] (16)

Table 1.5: Examples for extracted similar words from all query logs

Another interesing way to evaluate the proposed approach for extraction of similar words would be by applying the same algorithm to a standard corpus, after it was tagged according to part of speech and split into phrases with a Chunker. This experiment is planned for further research.

6 The Advantages of Using Query Logs

The evaluation of methods for compounds extraction and creating a similarity dictionary shows that even simple techniques can yield good results in information extraction from a pure query log. This fact can be explained by considering the structure and the characteristics of query logs. Each query is a compressed formulation of the information request of the search engine user. People represent their questions as concisely as possible and use casual terms. Applying context based methods, as described in Subsection 4.2, does not need any preparatory work to extract good and correct contexts: they are already provided by the query log itself. The next advantage is the good quality of resulting words, because they contain only common terms used in everyday language. After integrating methods in the query processing stage of a search engine, the updating of automatically created dictionaries using the most recently recorded query log is always possible. Furthermore, the results are language and domain independent, which is the most important advantage. Statistical methods can be easily used for query log data from different search engines. Using query logs from a domain-specific search engine will also return results from this particular domain. Although the query logs in other languages seem to contain about 25% English queries (as shown for the German query log in Section 3), the method yieds usable results (cf. Section 1.5, where English and Dutch similar words were found for the Dutch word "*naakt*" [*naked*]).

7 Conclusion

In this paper I presented known approaches for query log mining and experiments in applying simple statistical methods for the extraction of important natural language resources, which yielded remarkably good results. Advantages of using query logs as corpora, summarized in previous section, demonstrate that query logs possess features, which are unusual for normal corpora, but very beneficial for information extraction. Unfortunately there are still very little research in the exploration of these features and using query log as corpus in general. Search engines are still restricted to a simple full text search, without making use of any natural language methods. Including automatically created dictionaries with frequently used phrases, similar words or other useful information in query processing, would be a step towards improving Web retrieval and making it more sophisticated.

8 Acknowledgements

I would like to thank all human testers, who helped me to evaluate the sample sets. A special thanks to my mentor Benno Nieswand and the company exorbyte GmbH, where I learned so much.

Bibliography

Agichtein, E., S. Lawrence, and L. Gravano (2001). Learning search engine specific query transformations for question answering. In *Proceedings of World Wide Web Conference*, Hong Kong, pp. 169–178.

Beeferman, D. and A. Berger (2000). Agglomerative clustering of a search engine query log. In *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining*, pp. 407–416. ACM Press, Boston.

Cacheda, F. and A. Via (2001). Understanding how people use search engines: a statistical analysis for e-business. In *Proceedings of the e-2001 (e-Business and e-Work Conference and Exhibition)*, Volume 1, Venice, Italy, pp. 319–325.

Chen, A. (2002). Cross-language retrieval experiments at CLEF 2002. In *Advances in Cross-Language Information Retrieval*, Volume 2785/2003 of *Lecture Notes in Computer Science*, pp. 28–48. Heidelberg, Germany: Springer Verlag.

Cui, H., J.-R. Wen, J.-Y. Nie, and W.-Y. Ma (2002). Probabilistic query expansion using query logs. In *Proceeding of the Eleventh World Wide Web conference (WWW 2002)*, Hawaii.

de Lima, E. and J. Pedersen (1999, August). Phrase recognition and expansion for short, precision-biased queries based on a query log. In *Proceeding of the SIGIR'99*, Berkeley, CA.

Guenthner, F. and X. Blanco. Multi-lexemic expressions: an overview. Linguisticae Investigationes Supplementa.

Jansen, B., A. Spink, J. Bateman, and T. Saracevic (1998). Real life information retrieval: a study of user queries on the web. In *SIGIR Forum*.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of International Conference on Machine Learning*.

Manning, C. and H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Martnez-Santiago, F., A. Montejo-Rez, L. Urea-Lpez, and M. C. Daz-Galiano (2003). Sinai at clef 2003: Decompounding and merging. In *Working Notes for the CLEF 2003 Workshop*, Trondheim, Norway.

Miller, G. A. and G. C. Walter (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes 6(1)*.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvick (1985). *A Comprehensive Grammar of the English Language*. Longman.

Silverstein, C., M. Henzinger, J. Marais, and M. Moricz (1998). Analysis of a very large altavista query log. Technical Report 014, Compaq Systems Research Centre, Palo Alto, CA.

Wen, J., J. Nie, and H. Zhang (2002). Query clustering using user logs. *ACM Transactions on Infromation Systems 20*(1), 59–81.