[Cardell-Oliver et al.,1992] Rachel Cardell-Oliver, Roger Hale, and John Herbert. An embedding of timed transition systems in HOL. In Higher Order Logic Theorem Proving and its Applications, pages 263--278, Leuven, Belgium, Sept 1992

[Fikes & Nilsson, 1971] R.E.Fikes, N.J.Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving, Artificial Intelligence, 2(3/4), 1971

[Gogolla & Parisi-Presicce,1998] Gogolla, M. and Parisi-Presicce, F., 1998, "State Diagrams in UML - A Formal Semantics using Graph Transformation", Proceedings ICSE'98 Workshop on Precise Semantics of Modeling Techniques (PSMT'98),

[Handschuh & Staab, 2002] S. Handschuh and S. Staab. Authoring and annotation of web pages in CREAM. In The Eleventh International World Wide Web Conference (WWW2002), Honolulu, Hawaii, USA 7-11 May, 2002

[Hendler, 2001] J. Hendler, Agents and the Semantic Web, IEEE Intelligent Systems, vol.16, no.2, Mar./Apr. 2001, pp.30-37.

[Kalianpur, 2001] SMORE - Semantic Markup, Ontology, and RDF, Editor Aditya Kalyanpur A. V. Williams Building College Park, Maryland 20742

[Manna & Pnueli,1991] Manna, Z., Pnueli, A. The temporal logic of reactive and concurrent systems: Specification. Springer Verlag, 1991

[McGuinnes & van Harmelen, 2004] D.L. McGuinness F.van Harmelen, OWL Web Ontology Language Overview W3C Recommendation 10 February 2004

[McGuinnes et al., 2004] D.L. McGuinness, R. Fikes, J. Hendler, and L.A. Stein. ``DAML+OIL: An Ontology Language for the Semantic Web''. In IEEE Intelligent Systems, Vol. 17, No. 5, pages 72-80, September/October 2002

[Milani & Ghallab,1991] A.Milani, M.Ghallab eds. "New Direction in AI Planning", IOS Press 1996

[Vardi,1991] M. Vardi. An automata-theoretic approach to linear temporal logic. In F. Moller and G. Birtwistle, editors, Logics for Concurrency, pages 238-266. Springer Verlag, 1996

Authors' Information

Alfredo Milani – Department of Mathematics and Informatics, University of Perugia, Via Vanvitelli, 1, 06100 Perugia, Italy; e-mail: suriani@dipmat.unipg.it

Silvia Suriani – Department of Mathematics and Informatics, University of Perugia, Via Vanvitelli, 1, 06100 Perugia, Italy; e-mail: suriani@dipmat.unipg.it

EFFECTIVENESS OF TITLE-SEARCH VS. FULL-TEXT SEARCH IN THE WEB

Peretz Shoval and Tsvi Kuflik

Abstract: Search engines sometimes apply the search on the full text of documents or web-pages; but sometimes they can apply the search on selected parts of the documents only, e.g. their titles. Full-text search may consume a lot of computing resources and time. It may be possible to save resources by applying the search on the titles of documents only, assuming that a title of a document provides a concise representation of its content. We tested this assumption using Google search engine. We ran search queries that have been defined by users, distinguishing between two types of queries/users: queries of users who are familiar with the area of the search, and queries of users who are not familiar with the area of the search. We found that searches which use titles provide similar and sometimes even (slightly) better results compared to searches which use the full-text. These results hold for both types of queries/users. Moreover, we found an advantage in title-search when searching in unfamiliar areas because the general terms used in queries in unfamiliar areas match better with general terms which tend to be used in document titles.

Keywords: Indexing, Information retrieval, Precision of search results, Search engines, Title search, Web search.

1. Introduction

Search engines generally use the full text of Web pages for searching. The search of full text may be costly in terms of computing resources and time. A possible way to save such resources is by conducting the search on the titles of the documents rather than on their full text. The title of a document is supposed to provide a concise representation of its content. Kwok (1984) used the titles of cited academic publications to improve the indexing of the documents which cite them. He did so by adding, in the indexing process, the content of the cited titles to the content of the documents. Drori (2003) showed that in many cases title terms can be identified by analyzing the content of a document. Belaïd and David (1999), Taniar et al. (2000) and Schenker et al. (2003) used the titles for document representation to help users find relevant documents in search results. Lam-Adesina and Jones (2001) explored the intuitive assumption about the importance of terms appearing in the titles for increasing the weight assigned to such terms while generating document summaries. They generated summaries by extracting sentences out of the documents; sentences containing title terms were scored higher than sentences without title terms.

Obviously, a title of a document cannot provide much detail; it tends to be general and contain general terms, while more specific terms appear in the text itself. Given this, it seems reasonable that a query used for a title-search should include mostly general terms, while a query used for a full-text search should include mostly specific terms.

Research has shown that familiarity of users with the area in which they seek information has an impact on the quality of their search queries. Users who are familiar with the search area know the relevant terminology; therefore it is reasonable to assume that they are able to define precise search queries. On the other hand, users seeking information in areas with which they are not familiar do not know the relevant terminology, and therefore are likely to define imprecise queries that would yield many irrelevant results. With respect to the earlier discussion on generality or specificity of terms appearing in titles vs. the full text, it may be assumed that unfamiliar users would be better of using title-search because they are likely to use general terms in their queries, while familiar users would be better of using full-text search because they are more likely to use specific terms.

The purpose of this study is to compare the effectiveness of title-search and full-text search, and to determine how user familiarity with the search area interacts with the type of search. Section 2 presents related studies on search habits of Web users and the impact of user familiarity with the search area on search results; Section 3 outlines our hypotheses and describes the research; Section 4 presents the results, and Section 5 concludes and suggests further research.

2. Related Studies

Studies on how users behave while searching the Web reveal that they most often tend to define short queries, having an average of 2.35 terms (Jansen et al., 1998), 3.34 terms (Spink et al., 1999) and 2.4 terms (Spink et al., 2001). Web search queries are significantly shorter than queries in classical information retrieval systems, which consist of between 7 to 15 terms (Jansen et al., 2000). Jansen et al. (1998) found that users tend to explore less than three pages of results; the average is 2.21 pages, while half of the users examine only one page, and three quarters examine only two pages or less. Users also tend to perform short search sessions: they pose a query, look at the first page (or two) of results and explore only a few Web sites listed on that page. If they do not find relevant information, they may reformulate the query and repeat the search once or twice, and then abandon the search. On the average they reformulate a query 2.84 times in a search session; two-thirds of the users submit only a single query. These findings indicate how important it is that Web users will get the most relevant information already in the first few pages of the search results. This also explains why a common measure of performance of search engines is "precision at 10" (Jin and Dumais, 2001; Craswell et al., 2001; Eastman, 2002; Plachouras et al., 2003), which means the precision of the 10 top documents (usually presented by search engines in the 1st page of results).

Some search engines have "advanced" search options which allow users to define and run search queries using specific options that extend beyond the "simple" (common) option. For example, an advanced search may enable Boolean operators, or limit the search to specific file types, or to specific attributes of Web documents, such as the title. But users usually do not use these options (Jansen, 1998). Eastman (2002) claimed that the use of

advanced options does not improve the search results because the performance of current search engines is good anyhow (as measured by "precision at 10"). The author evaluated the benefit of using advanced search options and found that in 50% of the cases there was no difference in performance between a simple search and an advanced search; in 25% of the cases advanced searches yielded better results than simple searches, and in 25% of the cases simple searches yielded better results than advanced searches. (It should be noted that Eastman's research made no distinction between different advanced search options, so there is no way to discern if any of those options, such as a title-search, is better or worse than another.)

Only a few studies are concerned about the impact of domain knowledge of users on the results of Web searches. Hsieh-yee (1993) found that owing to their domain knowledge, users are familiar with the relevant terminology and hence can define precise search queries. But users who lack domain knowledge need to search for the right terms first, using various tools such as thesauri. Spink et al. (1998) studied the way Web-users judge relevancy of search results. They concentrated on documents that were defined by the users as "partially relevant" and found that the less users knew about the problem at hand, the more items they assessed as partially relevant; and the more they knew, the more items they assessed as relevant. Hoelscher and Strube (1999, 2000) studied the impact of domain knowledge on search performance combined with Web experience. Their subjects were asked to solve information problems using the Web only. They concluded that in order to succeed, users should have both domain knowledge and Web-search experience. Turnbull (2003) surveyed models to determine how users start to search for information in unfamiliar areas. He observed that users usually start by looking for initial information, learn the general concepts of the domain until they gain enough knowledge to enable them to define precise search queries and then evaluate the search results.

3. The Research

Our research hypotheses are as follows: Web users who are familiar with the research area are able to define search queries that yield high quality (high "precision at 10") results, whether the search is in full-text or in the title only; but in title-search the number of results (documents) they get is smaller than in full-text search because the precise query terms which they tend to use fit less with the more general terms used in titles. Contrarily, Web users who are not familiar with the search area are able to define search queries that yield purer (less precise) results; but in title-search the number of results they get is bigger than in full-text search because the more general terms which they tend to use fit more with the general terms used in titles.

To test the hypotheses we conducted Web searches with thirty-four subjects, all 4th-year students of Information Systems Engineering having several years of computer usage and Web search experience. Each of the participants was asked to define two search queries: one in an area familiar to him/her, and the other in an unfamiliar area. We used the Google search engine (Google, 2003) to run the queries. Google enables users to limit the search to the title field of Web documents. Title field search uses the content of the field enclosed by HTML title tags. (Even documents not in HTML, e.g. PDF files, can be viewed because Google automatically generates HTML versions of such files as it crawls the Web (Notess, 2001; 2002.))

Each user ran each of his two search queries twice: One was a "simple search", i.e. search of the full text, and the other was "advanced search", with the option of searching only in the title field. After conducting each search, the user evaluated the top 10 results appearing in the first page of results (or less, in cases when there were less than 10 results). For this, the user had to access the linked Website, read at least its first page, and decided whether or not it is relevant. The users' decisions were recorded for further analysis.

4. Results

The results of the searches are presented in Table 1. The rows detail the 34 cases (users). The columns show the four search scenarios: full-text search in a familiar area, title-search in a familiar area, full-text search in an unfamiliar area, and title-search in an unfamiliar area. Every column is sub-divided into two: one presents the "precision at 10" and the other presents the number of search results. "Precision at 10" is calculated by counting the number of relevant results (as determined by the user) divided by 10 or by the number of results in cases when there were fewer results. In the following sections we discuss the results according to three issues: a) precision of results; b) number of results; and c) length of queries.

Table 1: Search results

Unfamiliar Area				Familiar Area				User
Title		Full-Text		Title		Full-Text		
# of results	Precision at 10	# of results	Precision at 10	# of results	Precision at 10	# of results	Precision at 10	
0	no data	> 10	0.6	0	no data	> 10	0.9	1
> 10	0.3	> 10	0.2	0	no data	> 10	0.8	2
> 10	0.8	> 10	0.8	0	no data	> 10	0.9	3
0	no data	> 10	0.7	0	no data	> 10	0.8	4
1	1	> 10	0.6	3	0.667	> 10	0.7	5
> 10	0.8	> 10	0.9	> 10	0.8	> 10	0.9	6
> 10	0.6	> 10	0.7	0	no data	> 10	0.7	7
> 10	0.6	> 10	0.6	> 10	0.6	> 10	0.4	8
> 10	0.3	> 10	0.7	0	no data	> 10	0.9	9
0	no data	> 10	1	0	no data	> 10	0.7	10
2	1	> 10	0.5	0	no data	> 10	0.3	11
> 10	0.8	> 10	0.8	0	no data	> 10	0.9	12
0	no data	> 10	0.4	0	no data	> 10	0.5	13
9	0.889	> 10	0.9	10	0.9	> 10	0.8	14
> 10	0.2	> 10	0.2	0	no data	> 10	0.2	15
> 10	0.3	> 10	0.3	2	0.5	> 10	0.5	16
0	no data	> 10	0.9	0	no data	> 10	0.9	17
0	no data	> 10	0.8	0	no data	> 10	0.9	18
> 10	0.3	> 10	0.5	> 10	0.8	> 10	0.4	19
> 10	0.4	> 10	0.5	0	no data	> 10	0.4	20
> 10	0.8	> 10	0.6	> 10	1	> 10	1	21
> 10	0.7	> 10	0.5	2	0.5	> 10	0.7	22
0	no data	> 10	0.7	0	no data	> 10	1	23
> 10	0.8	> 10	0.8	2	0.7	> 10	0.9	24
> 10	0.7	> 10	0.7	> 10	1	> 10	0.7	25
1	1	> 10	0.7	> 10	0.9	> 10	1	26
> 10	1	> 10	1	0	no data	0	no data	27
0	no data	> 10	0.6	0	no data	> 10	1	28
> 10	1	> 10	1	0	no data	> 10	1	29
0	no data	> 10	0.8	0	no data	> 10	0.3	30
0	no data	> 10	0.7	2	1	> 10	1	31
> 10	0.9	> 10	1	> 10	0.5	> 10	0.4	32
3	1	> 10	0.8	> 10	0.7	> 10	0.6	33
> 10	0.8	> 10	0.7	> 10	0.8	> 10	0.9	34

4.1 Precision of Results

Table 2 presents the average "precision at 10" (as based on the values presented in Table 1) for the four scenarios.

Table 2: Average precision

Title	Full-text	
0.76	0.73	Familiar area
0.71	0.68	Unfamiliar area

As can be seen, title-search yielded better results compared to full-text search in both the familiar and unfamiliar areas. We can also see that search in familiar areas yielded better results compared to search in unfamiliar areas. However, t-tests of differences between the averages of the familiar and unfamiliar areas, for both full-text and title-searches, revealed that the differences are not significant ($p=0.111$ and $p=0.409$, respectively). Similarly, t-tests of differences between the averages of the full-text and title-searches, for both the familiar and unfamiliar areas, also revealed that the differences are not significant ($p=0.248$ and $p=0.151$, respectively). At any rate, it is important to note that the results of the full-text search are not better than those of the title-search.

4.2 Number of Results

Table 3 shows the number of results in all cases. The columns represent the four scenarios; each column is sub-divided into two, distinguishing between the number of cases with 10 or less results, and the number of cases with more than 10 results.

Table 3: Number of results

Unfamiliar Area				Familiar Area			
Title		Full-Text		Title		Full-Text	
# of cases with >10 results	# of cases with ≤ 10 results	# of cases with >10 results	# of cases with ≤ 10 results	# of cases with >10 results	# of cases with ≤ 10 results	# of cases with >10 results	# of cases with ≤ 10 results
20	4	34	0	9	6	33	0

As can be seen, in full-text search there are more than 10 results in all the cases (except for one search in a familiar area search where no results at all were obtained). In title-search the results are different: when searching in a familiar area only 15 cases yielded results, and in only in 60% of them (9) the number of results exceeds 10. When searching in an unfamiliar area, more (24) cases yielded results, and in 83% of them (20) the number of results exceeded 10. These results tell us that full-text search yields a redundancy of results regardless of the level of user familiarity with the search area. On the other hand, title-search yields less results, sometimes too few. However, title-search in an unfamiliar area provided more results than title-search in a familiar area. The reason for the difference may be, as hypothesized, that users in familiar areas are able to defined precise queries yielding good results in any case (full-text as well as title-search); but because their queries are specific, using precise terms, their title-search yield less results (because title terms tend to be more general). Contrarily, users in unfamiliar areas use more general query terms, which correlate better with the general terms used in titles, and therefore they obtain more results.

4.3 Lengths of Queries

The average length of the search quarries was 3.29 terms for an unfamiliar area, and 2.94 terms for a familiar area. These lengths are similar to Web query lengths reported earlier. In order to better understand the differences in the results between the two cases of the title-search, we analyzed the lengths of queries by comparing the differences between lengths of queries which yielded results and queries which did not yield results. Table 4 shows the query lengths, distinguishing between searches in familiar and unfamiliar areas.

Table 4: Length of queries

Title-search in Unfamiliar Area		Title-search in Familiar Area		
Results	No results	Results	No results	
2.79	4.5	3.06	4.53	Average
1.14	1.18	0.93	1.23	Standard dev.

For search in familiar areas, the queries that yielded no results are 50% longer than those that yielded results (4.53 compared to 3.06 terms). For search in unfamiliar areas, the difference is even greater, being about 60% longer (4.5 terms compared to 2.79 terms). T-tests reveal that the differences in the query lengths are significant ($p<0.00$ for both types of queries).

The lengths of queries that yielded results are within the range of query lengths reported in the literature (3.34 according to Spink et al., 1999; and 2.35 according to Jansen et al., 2000). But queries that yielded no results are substantially longer. Hence, the reason for fewer results in title-search can be explained by their length, because of the excessive number of specific terms. While specific/detailed queries are good if one wants to reduce the number of irrelevant results and seeks high precision in full-text search, they seem not to be so good in title-search, because – as said - a title consists of a small number of general terms, which do not correlate with the terms in detailed queries. Hence, queries used in title-search should be shorter if the user wants to get a substantial amount of results.

5. Conclusions

Using Google as a search engine, we showed that search queries provide highly precise results, regardless of whether a familiar or unfamiliar area is being searched. The results support the hypothesis that Web users searching in a familiar area are able to define precise search queries that yield high quality results. For searching in an unfamiliar area, this result contradicts the hypothesis (as we were expecting low precision). But the more interesting result is that search in the title filed yielded results which are not worse, and sometimes even better, than searching in the full text. Hence, a lot can be saved in the ways searches are performed and indexes are constructed: searching and indexing of Web pages can be based on titles of documents rather than on their full text.

Although a title-search yielded fewer results, this does not present a problem since users usually do not examine more than one or two pages of search results. At any rate, if a query yields too few results, it may be too specific and include too many terms, so the user can revise the query accordingly.

As any other experiment, this too has limitations, such as the small number of search queries and the use of one search engine only. The results of this study should be further validated by using more queries, different types of users, different search areas, and more search engines (besides Google).

Bibliography

Belaïd, A. & David, A. (1999). The use of information retrieval tools in automatic document modelling and recognition. Proc. of the 10th International Workshop on Database & Expert Systems Applications, 522-526.

Craswell, N., Hawking, D. & Griffiths, K. (2001). Which search engine is best at finding airline site home pages? CSIRO Mathematical and Information Sciences TR01/45.

Drori, O. (2003). Identifying the subject of documents in digital libraries automatically using frequently occurring words – study and findings. Proc. of the 3rd International Workshop on New Development in Digital Libraries, NDDL 2003, 3-12.

Eastman, C. (2002). 30,000 hits may be better than 300: precision anomalies in Internet searches. Journal of the American Society for Information Science and Technology, 53 (11), 879-882.

Google. www.google.com, (accessed September, 2003).

Hoelscher, C. & Strube, G. (1999). Searching on the Web: two types of expertise. Proc. of the 22nd Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, 305-306.

Hsieh-yee I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. Journal of the American Society for Information Science, 44 (3), 161-174.

Jansen, B., Spink, A., Bateman, J. & Saracevic, T. (1998). Real life information retrieval: a study of user queries on the Web. SIGIR Forum, 32 (1), 5-17.

Jansen, B., Spink A. & Saracevic T. (2000). Real life, real users and real needs: a study and analysis of user queries on the Web. Information Processing and Management, 36, 207-227.

Jin, R. & Dumais, S. (2001). Probabilistic combination of content and links. Proc. of the 24th Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, 402-403.

Kwok, K. (1984). A document-document similarity measure based on cited titles and probability theory, and its application to relevance feedback retrieval. Proc. of the 7th Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, 221-231.

Lam-Adesina, A. & Jones G. (2001). Applying summarization techniques for term selection in relevance feedback. Proc. of the 24th Annual Int'l ACM SIGIR Conference on Research & Development in Information Retrieval, 1-9.

Notess, G. (2001). Tracking title search capabilities. http://www.onlinemag.net/OL2001/net5_01.html.

Notess, G. (2002). Review of Google. http://www.searchengineshowdown.com/features/google.

Plachouras, V., Ounis, I., Amati, G. & Van Rijsbergen. (2003). University of Glasgow at the Web track of TREC 2002. Proc. of the 11th Text Retrieval Conference, TREC 2002.

Schenker, A., Last, M., Bunke, H. & Kandel, A. (2003). Graph representations for Web document clustering. Proc. of the 1st Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2003).

Spink, A., Graisdorf, H. & Bateman, J. (1998). From highly relevant to not relevant: examining different regions of relevancy. Information Processing and Management, 34 (5), 599-621.

Spink, A., Bateman, J. & Jansen, B. (1999). Searching the Web: a survey of EXCITE users. Internet Research: Electronic Networking Applications and Policy, 9 (2), 117-128.

Spink, A., Wolfram, D., Jansen, B. & Saracevic, T. (2001). Searching the Web: the public and their queries. Journal of the American Society for Information Science and Technology, 52 (3), 226-234.

Taniar, D., Jiang, Y., Rahaya, J. & Bishay, L. (2000). Structured Web pages management for efficient data retrieval. Proc. of 1st Int'l Conference on Web Information Systems Engineering (WISE'00), 2097-2104.

Turnbull, D. (2003). Augmenting information seeking on the World Wide Web using collaborative filtering techniques. http://donturn.fis.utoronto.ca/research/augmentis-toc.html (accessed June 2003).

Authors' Information

Peretz Shoval – Department of Information Systems Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel; e-mail: shoval@bgumail.bgu.ac.il

Tzvi Kuflik – Department of Management Information Systems, Haifa University, Haifa, Israel;
e-mail: tsvikak@mis.hevra.haifa.ac.il

EMPIRICAL METHODS FOR DEVELOPMENT AND EXPANDING OF THE BULGARIAN WORDNET

Pavlina Ivanova, George Totkov, and Tatiana Kalcheva

Abstract: Some basic points from the automated creation of a Bulgarian WordNet – an analogue of the Princeton WordNet, are treated. The used computer tools, the received results and their estimation are discussed. A side effect from the proposed approach is the receiving of patterns for the Bulgarian syntactic analyzer.

Keywords: Empirical Methods in NLP, WordNet

1. Introduction

WordNet is developed in the Princeton University [2,4] as a lexical database of English. The first multilingual database to realize such approach is *EuroWordNet* (EWN) ([11], [12]) consisting of eight European languages. The monolingual databases are related to the *Princeton WordNet* (PWN) (and in this way to each other) via an interlingual index (ILI).

The *Bulgarian WN* (BWN) has been developed as a cooperative task involving the Plovdiv University and the Department for Computer Modelling of Bulgarian Language at the Bulgarian Academy of Sciences (DCMB). The work is part of an EC funded project (IST-2000-29388) *BalkaNet* [7] for the creation of a multilingual lexical database (like EWN) for 6 Balkan languages (Bulgarian, Greek, Romanian, Serbian and Turkish, Czech).

2. Forming of a BWN

The main stages in the automatic creation of a BWN (A_BWN) are presented in [8]. We discuss further the tools and the results received in this process – namely the extraction of synsets from an English-Bulgarian dictionary (EBD) and the receiving of A_BWN.

Our starting point is the transformed EBD [6] with more than 160,000 entries. Each different meaning of an English word is placed on a different row. Each row contains the English word (entry) and its translation