ANALYSING TERMS, PAIRS, TRIPLETS AND FULL QUERIES USED IN INTRANET SEARCHING

Dick Stenmark
IT University of Göteborg, Department of Applied IT, P.O.Box 8718, SE-40275 Göteborg, Sweden

Keywords: Intranet, search engine usage, query term analysis.

Abstract: Web search engines have become often-used tools for many ordinary people today and a growing number of researcher are therefore studying how these lay-persons interact with such tools. Studies of public web search engine usage have often produced term frequency lists to illustrate the information needs the users. This study differs on several aspects from previous work. Firstly, we have analysed the logs of an intranet search engine, since studies of corporate internal search behaviour are in short supply. Secondly, we have not just used search terms but also full queries and show that single terms give a skewed understanding. Thirdly, we have analysed data from three different years - 2000, 2002 and 2004 - to be able to detect shifts and trends in information seeking behaviour.

1 INTRODUCTION

Intranets, i.e., corporate internal webs, have in less than 10 years time gone from being perceived as a spelling error to become one of the most widespread organisational information technologies, and the information available on intranets seems to grow at a higher pace than the web itself (Stenmark, 2005b). Obviously, organisational members need good search tools to find the information they need and since public search engines such as Google are unable to access and index the content of the intranets, organisations have to install and host their own internal search tools.

However, it has been noticed that intranets have their own specific characteristics and that information seeking behaviour seen on the public web not necessarily can be expected to be repeated on intranets (Fagin et al., 2003). Intranet information is narrower in the sense that it is business oriented and more context specific. Intranets provide important business information environments and to understand the information need and behaviour of the organisational members is thus of vital interest for organisations to be able to provide suitable resources and for researchers and developers to be able to design better tools.

In this paper, we contribute to the understanding of intranet search behaviour by providing a longitudinal comparison of the queries submitted to a corporate intranet search engine. Our data covers three different weeks from the years 2000, 2002, and 2004. In particular, we have studied not only the most frequently used search terms (which is otherwise a common approach) but also the actual queries, including term pair and term triplet. We have also studied how these have changed over time and identified both short- and long-term information needs.

The paper is organised as follows. In the next section we account for related research from intranet and public web studies and thereafter we present out research setting and research method. In section four the result of or work is accounted for and we subsequently discuss this in detail in section five. In section six, finally, we draw our conclusions and suggest design implications based on our findings.

2 RELATED WORK

Relatively little work has yet been devoted to intranet searching and practically nothing to the content of intranet searching. Choo et al. (1998) studied corporate employees' use of the web as an information resource to support their daily work activities, and found them engage in a range of complementary modes of information seeking,

Stenmark D. (2007).
ANALYSING TERMS, PAIRS, TRIPLETS AND FULL QUERIES USED IN INTRANET SEARCHING.
In *Proceedings of the Third International Conference on Web Information Systems and Technologies - Web Interfaces and Applications*, pages 122-129
DOI: 10.5220/0001260501220129

varying from undirected viewing to formal searching. Göker and He (2000) examined a week's worth of log file data from Reuter's intranet search engine in order to develop a method for automatic session boundary detection. Hawking et al. (2000) implemented a search engine on a university intranet in order to "reality test" an algorithm, and in a similar vein, Fagin et al. (2003) studied IBM's intranet with a focus on technical matters. Stenmark, finally, reported a time-based analysis of a week's worth of intranet search engine behaviour but he only studied how users interacted with the technology; not what they actually searched for (Stenmark, 2005a). The current study does thus make an explicit contribution to this field, but it also means that there is little previous work on which to build. We have thus had to compare and contrast or results to what is known about public web searching.

On the public web there are two types of search engines – general-purpose engines (such as e.g. Google) and site specific ones (e.g. the one found at www.ibm.com). The most consistent examination of public search engine usage has been carried out by Spink and Jansen, who over the last decade have established a useful research base of web searching behaviour (e.g. Jansen et al., 2000; Spink & Jansen, 2004; Spink et al., 2001; 2002). When it comes to site specific search engines, Chau et al.'s (2005) analysis of the Utah state web site search engine is a useful contribution. Such local web site search engines have much in common with intranet search engines, we argue, and we shall use the results of Chau and colleagues as a point of reference for our own work.

Chau et al. (2005) found both similarities and differences when comparing general-purpose search engine users and web site search engine users. The users in Chau et al.'s study used an average of 2.25 terms per query, which is close to the numbers reported for public search engines (Silverstein et al., 1999; Jansen et al., 2000; Spink et al., 2001). The average number of result pages examined (1.47) is also fully in line with what has previously been reported. As far as these aspects were concerned, there was no difference between the two user groups. However, the web site search engine users only submit, in average, 1.25 queries per session, which is only about half the amount reported for public search engine users. Chau et al. suggest that this may be because web site search engine users have more specific information needs. Further, in the Utah study almost 30% of all queries were phrase searches, i.e., contained quotation marks, whereas Spink & Jansen (2001) only found 5% in their study.

However, the most significant difference was, not surprisingly, the content of the queries; Chau et al. compared the most frequently used query terms with those reported by Spink et al. (2001). Chau and colleagues found that web site search engine users submitted terms much more related to the specific domain. Comparing the top 50 terms from Chau et al. and Spink et al., only 9 terms occur in both lists and only two of those are functional words rather than semantic words. This, again suggests that web site searchers have a more specific information need than do users of general-purpose search engines.

In addition, Chau and colleagues also examined the whole queries and found big differences compared to the single term lists. However, they did not present any theory as to why this difference existed. We shall adopt their approach in our study, as explained next, and extend Chau et al.'s study in two ways; firstly by adopting it to the intranet domain and secondly by providing a multiple-year analysis in contrast to Chau et al.'s single year study.

3 RESEARCH SETTING & METHOD

This research is based on analysis of search engine log files from Jupiter's intranet. Jupiter (a pseudonym) is a big Swedish manufacturer group with offices and production plants in many countries around the world that employs some +80,000 people. Jupiter's intranet was established in 1995 and quickly developed into a large information repository. In 1998, Jupiter purchased and implemented a commercial search engine, and when spidering the intranet little over 400,000 documents were indexed from some 450 web servers. These numbers continued to grow; at the end of the millennium the search engine had indexed 750,000 documents and found more than 700 web servers and in 2002 there were over 1,500 known web servers on the intranet, according to Jupiter sources.

The search engine generates a log file where every transaction the users have with the server is recorded. This log file contains the IP addresses of the users' computers, the date and time (datetime) of the transactions (as logged by the server using Central European Time), the query strings as entered by the users, information regarding which result pages the users have requested, and some additional parameters not used in this particular study. The three log files used were collected in 2000, 2002 and 2004, respectively. The 2000 log file contains almost four week's worth of transactions from January 31st to February 24th. The 2002 log file contains one week's worth of transactions from October 21st to October 27th, and the 2004 log file, finally, contains one week's worth of transactions from October 14th to October 20th. In all, the log files contain more than 128,000 activities from more than 23,000 users.

Transaction log analysis (TLA) is a well-established method when examining search engine usage (Jansen, 2006). Still, commentators acknowledge that no standardised metrics have been agreed upon and interpretations and definitions differ between studies (cf. Jansen & Pooch, 2004; Spink et al., 2001). In our study, we extracted all query strings from the log files and sorted and counted all queries. These queries where thereafter split up in individual words and operators, and counted for frequency.

We also counted all term pairs and term triplets. This included both "natural" pairs/triplets where users explicitly had submitted the two/three terms together (such as in human resources or Jupiter golf competition), and "derived" pairs/triplets where these were extracted from longer query phrases (e.g., the query Jupiter golf competition generates the two pairs Jupiter golf and golf competition). All results were thereafter analysed and compared to the results reported by Chau et al. and other related work.

4 RESULTS

We first calculated the absolute frequency for every query term and year. For year 2000 we found 17,390 different terms (hereafter referred to as types). Of these types, 10,376 terms or 59.7% were only used once (hereafter referred to as hapaxes). However, many types were also repeated resulting in a corpus of 69,369 search words (hereafter referred to as tokens) being submitted. For year 2002 we had a corpus of 25,320 tokens containing 8,021 types (31.7%). 4,722 or 59.5% of the types were hapaxes. For year 2004, finally, we had 30,719 tokens consisting of 9,037 types (29.4%) and 5,179 hapaxes (57.3%).

The above statistics are summarised in Table 1. The 100 most frequently used terms (the top-100) accounted for between 22.9 and 24.0% of the total terms, as can also be seen in table 1. In addition, table 1 accounts for the portion of the total that the top-50 and top-10 terms result in.

Table 1: Basic statistics for this study.

	2000	2002	2004
Number of tokens	69,360	25.320	30.719
Number of types	17,390	8.021	9.037
top-100	22.9%	24.0%	23.0%
top-50	16.8%	17.6%	16.0%
top-10	8.0%	7.7%	7.3%
Number of hapaxes	10.377	4.772	5.179
out of total	15.0%	18.8%	16.9%
out of different	59.7%	59.5%	57.3%4

We manually analysed the top-100 search terms for each year but due to space limitations we only present the top-25 terms in table 2 below. There were a total of 185 different types amongst the 300 most frequently used search tokens. Thirty-two of these (representing 17.3%) were found across all three years. Another 49 terms (26.5%) were found in two of the years, and the remaining 104 terms (56.2%) were only used in one year.

Table 2: The 25 most frequently occurring search terms for the three years.

pos	2000	2002	2004
1	jupiter	jupiter	jupiter
2	servicebilar	coda	coda
3	servicebil	outlook	rapido
4	and	rapido	tidinfo
5	coda	pc	outlook
6	sif	standard	it
7	standard	tidinfo	service
8	word	mail	gps
9	rapido	servicebilar	ebd
10	it	web	password
11	job	mailforms	gdi
12	class	parma	business
13	service	password	standard
14	eddo	it	parts
15	lift	service	parma
16	ford	eddo	group
17	quality	and	web
18	lediga	parts	tdm
19	competition	gps	reseräkning
20	jbb	forms	and
21	products	business	pbp
22	golf	std	plan
23	product	access	gdp
24	mcs	class	global
25	r70	mcs	of

Looking specifically at the top-10 for each year, we found the distribution to be very similar. Three out of a total of 19 types (representing 15.8%) were amongst the top-10 for all three years (jupiter, rapido, coda), five terms (26.3%) were found in two of the years, and 11 terms (57.9%) were only found in one top-10 set.

The frequencies of the terms appearing in table 2 were left out due to space limitations but to give the reader a flavour of the numbers we here present a few samples. Position #1 for the year 2000 (jupiter) occurred 1,713 times, position #10 (it) 262 times, position #50 (download) 104 times, and position #100 (bus) occurred 70 times. Corresponding frequencies for 2002 were 414, 108, 43, and 27, and for 2004 655, 120, 51, and 35. As can be seen from these numbers, the frequencies drop radically with decreasing rank. This is a since long known phenomenon documented by Zipf, who noted that a double-log rank-frequency plot generates a straight line with a slop of -1 for large (English) texts (Zipf, 1932). Plotting the query words from our log data in such diagrams, our lines were not as steep as Zipf's prediction; the slopes for the three years were -0.8895, -0.8133, and -0.8435, respectively. Figure 1 shows the plot for the year 2000 data.

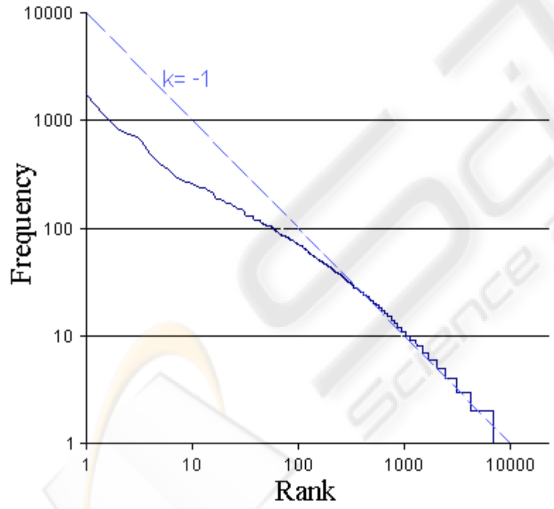

Figure 1: Double-log rank-frequency plot showing the Zipf distribution for the year 2000 (k=-1 indicated).

In contrast to table 2 above, which lists the most frequently used terms, tables 3 and 4 below show the most frequent pairs and triplets, respectively, found amongst the query terms. The tables contain both naturally occurring pairs and triples and derived occurrences, i.e., pair and triplets extracted from longer text sequences. As we can see, the term jupiter is in tables 3 and 4 combined with other words and appears in about one third of the pairs/triplets, and many of the frequent terms in table 2 (such as coda, sif, rapido, and tidinfo) are not represented in tables 3 and 4.

Table 3: The 10 most frequently occurring query pairs and their frequencies for the three years.

		2000		2002		2004
Pos	Freq	Terms	Freq	Terms	Freq	Terms
1	152	golf competition	51	web access	56	jupiter it
2	131	lediga jobb	43	mail forms	43	jupiter culture
3	104	jupiter servicebilar	42	hem pc	27	jupiter lifts
4	96	jupiter products	40	utbildnings pc	27	the jupiter
5	81	jupiter it	37	jupiter it	26	jupiter products
6	80	jupiter product	34	standard parts	25	jupiter group
7	77	jupiter golf	25	jupiter bil	25	function group
8	63	jupiter lift	23	business plan	25	outlook password
9	59	jupiter nu	20	change password	24	business plan
10	53	jupiter culture	19	jupiter culture	23	business objects

The tables contain both naturally occurring pairs and triples and derived occurrences, i.e., pair and triplets extracted from longer text sequences. As we can see, the term jupiter is in tables 3 and 4 combined with other words and appears in about one third of the pairs/triplets, and many of the frequent terms in table 2 (such as coda, sif, rapido, and tidinfo) are not represented in tables 3 and 4.

We examined the top 25 term pairs for each year (Tables 3 and 4 show only the top-10 due to space limitations). Out of the 75 term pairs, only 4 pairs (jupiter products, jupiter it, jupiter culture and business plan) were present in all three years, which corresponds to 5.3%. Another 11 pairs (14.7%) were present in two of the years, whereas the remaining 60 pairs (80.0%) only ranked amongst the top-25 in one year. Comparing tables 2 and 3, we see that although there are no term pairs in table 2, many of the terms in table 2 can be seen in the pairs of table 3. Many of the highly ranked pairs consist of terms on the top-100 list.

When examining the top-25 triplets from each year we found that only one of the 75 term triplets (the jupiter culture) was represented in all three years, which corresponds to 1.3%. Another 4 triplet (5.3%) were present in two of the years, whereas the remaining 70 triplets (93.3%) only were present in one year. Table 4 shows the top-10 triplets.

Table 4: The 10 most frequently occurring query triplets and their frequencies for the three year.

		2000		2002		2004
Pos	Freq	Terms	Freq	Terms	Freq	Terms
1	71	jupiter golf competition	13	-jbb -jbt jlt	20	the jupiter culture
2	29	word for windows	7	function group index	10	code of conduct
3	25	2000 1999 1998	7	localização das concessionar	10	jupiter lifts plant
4	24	aftermarket and service	6	who is who	9	jupiter do brasil
5	23	no 4 1999	6	design building landscaping	8	i-shift gear box
6	23	cst newsletter , no	6	the jupiter culture	8	regulations and certification
7	23	newsletter, no 4	5	outlook web access	7	engine data sheet
8	20	jupiter servicebilar ab	5	jac quality policy	7	lifts plant in
9	18	jupiter action service	5	-jbt jlt -it	7	welding manual design
10	16	jupiter attitude survey	5	one company vision	6	class for unix

We also examined the most frequently occurring queries as submitted by the users and we found that single term queries dominated; there are only nine multiple term queries amongst the top-100 for the year 2000 and eight and seven for the years 2002 and 2004, respectively. There is only one three-term query (jupiter golf competition) and ten of the multiple term queries contain the word jupiter (the top-25 are presented in table 5).

Twenty-five queries (12.6%) were present amongst the top-100 all three years. Almost half of these (12) were to (in-house) systems of various kinds (e.g., coda, rapido, or outlook). Nearly a third (8) were HR-related or link to employee-specific matters, and the remaining concerned organisational matters and miscellaneous. Thirty-nine queries (19.7%) were amongst the top-100 in two years. With only 3 exceptions, it was always from two adjacent years, i.e., 2000-2002 or 2002-2004. Finally, two thirds of the top-100 words or 134 instances were present in one single year only. These terms were difficult to classify since the represented a wide spread of interests.

One noticeable difference when comparing table 2 with table 5 is that the term jupiter has disappeared from the latter. Comparing the top-100 year by year, we found only 46 overlapping terms for the year 2000, 56 terms for year 2002, and 39 for year 2004.

Table 5: The 25 most frequently submitted queries for each year (multiple-word queries coloured).

pos	2000	2002	2004
1	servicebilar	coda	coda
2	servicebil	rapido	rapido
3	coda	outlook	tidinfo
4	sif	tidinfo	ebd
5	rapido	mailforms	gps
6	eddo	parma	parma
7	mcs	servicebilar	gdi
8	metall	eddo	tdm
9	word	gps	pbp
10	standard	mcs	reseräkning
11	class	cats	sox
12	parma	reseräkning	outlook
13	c-bil	servicebil	impact
14	tdm	hempc	cats
15	cf	webmail	gdp
16	blanketter	standard	teamplace
17	lediga jobb	tdm	mailforms
18	bilbiten	utbildningspc	vinst
19	sörredsgården	web access	standard
20	jlt	sbgtools	f2b
21	eifel	mail forms	protus
22	gränna	hem pc	scs
23	jobb	email	alviva
24	jupiter servicebilar	gdp	phoenix
25	job	mail	password

This ends our result section and we shall now discuss these findings and their implications.

5 DISCUSSION

When comparing our tables with results from studies of the public web, we immediately see that the search terms used in public search engines differ significantly from the terms and queries we found at Jupiter. This is not at all surprising and echoes the findings of Chau et al. (2005) who noted that terms used in site searching were very different from those used in general-purpose search engines. For example, neither we nor Chau et al. found many sex related terms, whereas such terms often dominate the ranking list from public search engines. The focus of this work is not on the query terms *per se* since these will vary from setting to setting, but on the method of analysing search behaviour and information needs and on the patterns that can be observed when examining search queries over time.

Studying table 2, one can come to the conclusion that jupiter is a rather common query. This is only partly true; jupiter *is* indeed a frequently used term but not a frequently used query. In fact, “jupiter” as a stand-alone term occurs only in 34 of the 2,782 queries that includes the term jupiter. In 98.78% of the jupiter-related queries, the term jupiter is combined with other terms, which can be seen also from tables 3 and 4. The term jupiter does thus not represent the information need; this can instead be found in the other part of the pair (such as in “jupiter lift”) or triplet (such as in “jupiter golf competition”). So although table 2 is correct in a statistical sense, such listing of individual terms may skew the understanding of the search behaviour. Term frequency lists are presented in much of the published research in this area (cf. Jansen & Spink, 2005; Spink et al., 2001; Jansen et al., 2000), but we argue it may be better to instead list the most frequently submitted queries or to include the most frequently used pairs and triples, as do Chau et al. (2005). Only half of the most frequently used terms overlapped with the most frequently submitted queries. If we see differences between term frequencies and query frequencies already on an intranet where the average query length is 1.44 terms and 69% of the queries are single term queries (Stenmark, 2005b; 2006), this difference would probably be even more evident on the public web where the average query length is closer to 2.5 terms. This further underlines the need to look beyond mere query term analysis when trying to understand the information needs of search engine users.

As in Chau et al.’s (2005) study, our study shows that the frequencies for the highest ranked term pair is considerably lower than the frequency of the highest ranked term, and that the frequency for most sought for triplet is lower still. We also note the drop is much more pronounced in our data than in Chau et al.’s study. In addition, the slope of the Zipf plots in figure 1 is not as steep as theory would have it. These observations suggest that a larger portion of single term queries are used at Jupiter. Referring to Fagin et al. (2003), we suggest that this is because intranets contain more jargon and more acronyms than do the public web. Another possible explanation suggested by Stenmark (2005b; 2006) is the presence of Swedish terms. The Swedish language makes use of compound words, resulting in single terms where e.g. English would have used two terms.

We were expecting there would be more unique search terms on a general-purpose search engine than on a site-specific one, but Jansen et al.’s (2000) slope of -0.975 for Excite terms is very close to Chau and colleagues’ slope of -0.9533 for the Utah search engine. A single web site can be expected to be more narrow in coverage and thus have a more limited vocabulary, and we were expected this to show in the distribution of search words. We had originally been expecting the Zipf plot of an intranet search engine to fall somewhere in between the Utah and the Excite plots but now our slopes of around -0.85 are less steep than both the other. We posit that the Swedish way of constructing compound words make the number of terms grow quicker than the frequency, hence producing these results. Additional (linguistic) analysis is required to fully understand this issue. It would be interesting to compare our findings to those from other intranet using other languages, say Finnish or English, to try to establish what is intranet dependent and what dependents on the language.

As was evident from table 1, the top terms portions of the total are pretty consistent over the years, i.e. a relatively small subset of the terms is used again and again. The portion of hapaxes (i.e., not repeated words) is not equally stable, although the variances are rather small. Close to 60% of the query terms are used only once, but since the repeated words are sometimes used very frequently, the hapaxes only make up some 15-19% of the total corpus. Still, 15-19% is a significant portion and it indicates that the information need is focused on quite a narrow field. When studying the top-100 terms, we noted that although more than half of the terms were present only in one year, some 17% of the terms reappeared every year. This distribution holds also for the top-10 terms. The corresponding numbers for the top-100 queries are similar; some 12% of the queries are found across all years. Apparently, there are things that the Jupiter employees continue to search for year after year, indicating what we mean is *a long-term information need*. Information about such needs would be useful to information providers and site designers within the organisation. Chau et al. (2005) argue that such frequently sought-for information should be made accessible via prominently placed links.

However, we see that the portions of terms and queries *not* repeated are bigger and we posit that the large portion of unique terms and unique queries indicate that there is a shift in information seeking behaviour from year to year. These queries may indicate *the short-term information needs*. These needs may be further be seasonal, as suggested by Chau et al. (2005). It seems plausible the information about the Jupiter golf competition will be more attractive closer to the actual event. The shift in information needs that this data suggest may also stem from a re-organisation of the available information or a re-make of the intranet. We suggest qualitative studies be carried out to explore this issue in more depth.

Our study also shows that a large international organisation may have a multi-lingual intranet, despite an official corporate language (English in this case). This stresses the importance of multi-language information retrieval research. Search engine vendors aiming for the intranet market should closely follow this development and preferably form joint ventures with multi-lingual retrieval researcher to help push the frontier further. In addition, the large number of indeterminable terms also point to the need for research on how to correctly deal with synonyms and homonyms in information seeking.

There are several organisational implications to be drawn from this study. Some information needs appear to be persistent and time-independent and organisations should adjust their information provision accordingly. This means that adding information, updating it, highlighting it, adding metadata to it and linking to it from many places are important activities for the organisation once these needs are identified. Search engine log file analysis may thus be a useful tool when assessing the effects of information architecture remakes and new web site designs. Other information needs are more short-term; they emerge and disappear in short cycles, but may still be very important to the business. To be able to respond to such shifting information needs, organisations must closely monitor the queries and be quick to provide the required information. As we have illustrated, it is not enough to study the most frequently used terms, but the whole query.

There are also obviously limitations to this study. Although we have used data from three different years and thus been able to follow the development of the queries, our study is limited to one intranet. This is understandable, since a lot of work is required to analyse this amount of data, but our findings still have to be replicated and tested elsewhere before any far-reaching conclusions can be drawn. In our qualitative analysis of the data we have restricted us to the most frequently used terms from each year. It is possible that this has skewed the outcome of the analysis and that our findings do not represent the corpus as a whole. This also has to be taken into consideration.

6 CONCLUSIONS

We have studied three log files from a corporate intranet search engine; one file from 2000, one from 2002, and one from 2004. Having extracted the actual queries and the query terms we have been able to analyse what the organisational member have sought for and how their information needs have shifted over time.

It is common practice to use query term frequency lists to illustrate information needs. In this paper we have shown that this may produce misleading conclusions since single words in isolation carry very little information. More useful is to present the most frequently used *queries* or the most frequently used term *pairs* or term *triplet*, since this approach allows for more context.

The Zipf plots from our intranet study show slopes that are less steep than those produced by both public search engines and web site search engines. This means that new terms are used more often than expected and further research is needed to show if this holds for intranet search in general.

The majority of the queries and query terms are replaced from year to year. This suggests that short-term information needs fluctuate and are time-dependent. Organisations must thus continuously keep track of the current and emergent needs and be ready to provide the corresponding information. However, we also conclude that certain information needs are rather persistent and time-independent and organisations should focus on providing content in these areas. The Zipf-like distribution means that only a fraction of the queries need to be catered for in order to cover much of the information needs.

ACKNOWLEDGEMENTS

The author is grateful to the Jupiter corporation for providing access to their log files, to Artur Foxander, Richard Wallmark and Taline Jadaan for help during the data processing, and to the reviewers for constructive critique. This work was sponsored by the Swedish Council for Working Life and Social Research (FAS) via grant #004-1268.

REFERENCES

Chau, M., Fang, X., and Sheng, O. R. L. (2005). Analysis of the Query Logs of a Web Site Search Engine. *Journal of the American Society for Information Science and Technology*, 56(13), 1363-1376.

Choo, C. W., Detlor, B., and Turnbull, D. (1998). A Behavioral Model of Information Seeking on the Web: Preliminary Results of a Study of How Managers and IT Specialists Use the Web. In *Proceedings of ASIS Annual Meeting*, Pittsburgh, PA., Oct 24-25, 290-302.

Fagin, R., Kumar, R., McCurley, K., Novak, J., Sivakumar, D., Tomlin, J. and Williamson, D. (2003). Searching the Corporate Web. In *Proceedings of WWW2003*, Budapest, Hungary, pp. 366-375.

Göker, A. and He, D. (2000). Analysing Web Search Logs to Determine Session Boundaries for User-Oriented Learning. In *Proceedings of Adaptive Hypermedia and Adaptive Web-based Systems*, Trento, Italy, pp. 319-322.

Hawking, D., Bailey, P. and Craswell, N. (2000). An intranet reality check for TREC ad hoc. *Technical report: CSIRO Mathematical and Information Sciences*.

Jansen, B. (2006). Search log analysis: What is it; what's been done; how to do it. *Library and Information Science Research*, 28(3), pp.407-432.

Jansen B. and Pooch U. (2004). Assisting the searcher: utilizing software agents for Web search systems. *Internet Research: Electronic Networking Applications and Policy*, 14 (1), 19-33.

Jansen, B. and Spink, A. (2005). An analysis of Web searching by European AlltheWeb.com users. *Information Processing and Management*, 41, 361-381.

Jansen, B., Spink, A., and 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.

Spink, A. and Jansen, B. (2004). *Web Search: Public searching of the web*. Kluwer Academic Publisher.

Spink, A., Ozmutlu, S., Ozmutlu, H. and Jansen, B. (2002). U.S. versus European Web Searching Trends. *ACM SIGIR Forum*, 36(2), 32-38.

Spink, A., Wolfram, D., Jansen, B. and 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.

Stenmark, D. (2005a). One week with a corporate search engine: A time-based analysis of intranet information seeking. In *Proceedings of AMCIS 2005*, Omaha, NE, 11-14 August.

Stenmark, D. (2005b). Searching the intranet: Corporate users and their queries. In *Proceedings of ASIS&T 2005*, Charlotte, North Carolina, October 28-November 2, 2005.

Stenmark, D. (2006). Intranet users' information-seeking behaviour: A longitudinal study of search engine logs. In *Proceedings of ASIS&T 2006*, Austin, Texas, November 3-6, 2006.

Zipf, G. K. (1932). *Selected studies of the principle of relative frequencies in language*. Addison-Wesley.