Stenmark, D. (2005). "Searching the intranet: Corporate users and their queries",
in Proceedings of ASIS&T 2005, Charlotte, NC., Oct. 28-Nov.2.

Searching the intranet: Corporate users and their queries

Dick Stenmark
Viktoria Institute, Göteborg, Sweden. stenmark@viktoria.se

Abstract
By examining the log files from a corporate intranet search engine, we have analysed the actual web searching behaviour of real users in a real business environment. While building on previous research on public search engines, we apply an alternative session definition that we argue is more appropriate. Our results regarding session length, query construction and result page viewing confirm some of the findings from similar studies carried out on public search engines but further our understanding of web searching by presenting details on corporate users' activities. In particular, we suggest that search sessions are shorter than previously suggested, search queries have fewer terms than observed for public search engines, and number of examined result pages is smaller than reported in other research. More research on how corporate intranet users search for information is needed.

Introduction
The advent of the World Wide Web (hereafter the web) has radically altered the way individuals find information and transformed information retrieval (IR) in the sense that the field is no longer exclusively populated by trained IR-professionals but open to users with little or no knowledge of non-web-based tools (Spink *et al*., 2001; Jansen & Spink, 2003). Traditionally, IR tools have been designed for IR-professional and it has been argued that web search engines are also based on these principles, but that searching information on the web is very different from information retrieval as performed within the IR discourse (Jansen *et al*., 2000). As a consequence of this observation, a body of research on how casual users interact with web search engines is beginning to form but the topic is far from fully understood (Spink *et al*., 2001; Jansen & Spink, 2003). This paper adds to the growing body of research by presenting another study of how casual users interact with search engines in a real business environment.

Although this study builds on the results of Jansen, Spink and colleagues (Jansen *et al*., 2000, Spink *et al*., 2001, Jansen & Spink, 2003), it adds a new angle by focusing not on public web search engines but on intranet search engines. There are numbers suggesting that three out of every four web servers being installed are intended for intranet usage (Gerstner, 2002), but despite the fact that intranets seem to grow at a higher pace than the web itself, studies of corporate use of IR tools are almost non-existent. Just as web information seekers differ in behaviour from trained IR personnel (and therefore deserve to be better understood), searching an intranet differs from searching the public web (Fagin *et al*., 2003). The influential work carried out by Spink, Jansen and other researchers in recent years therefore need to be repeated in a corporate context and this paper is the result of such an effort.

This paper reports from an ongoing study of intranet search behaviour carried out at a large European company group. Just as with Jansen *et al*. (2000), it involves real users with real information needs submitting real queries to a real search engine. The question we try to answer is basically how corporate users search their intranet and we do this by addressing the following four sub-questions: 1) What is the temporal length of a search sessions?, 2) How many queries per session do they submit?, 3) How many query terms do they use?, and 4) How many pages of result do they examine? These are questions that have previously been examined for public search engines, and we aim to extend that body of research. The paper is structured as follows. In the next section we present a rationale and account for some related work that has influenced the research setup. In section three, we describe the research site and the method used, and our results are thereafter, presented in section four. Section five contains our discussion before closing with conclusions in section six.

Rationale and related work
Although studies of how users interact with traditional IR systems have been presented at the ACM SIGIR conference for many years, it was not until quite recently that scholars began to study how non-professional IR-

personnel interacted with web search engines (Jansen *et al*., 2000). For example, in the late nineties, Hölscher (1998) presented a study of the German web search engine Fireball and Silverstein *et al*. (1999) reported on Alta Vista usage. The most consistent examination of web search engine usage has been carried out by Spink and Jansen, who – alone or in collaboration with others – have established a useful research base of web searching behaviour during the last eight or so years (e.g. Jansen *et al*., 1998; 2000; Jansen & Spink, 2003, Spink *et al*., 1999; 2001; 2002).

Jansen *et al*. (2000) analysed web logs received from the Excite public search engine in 1997. They studied, amongst other things, whether queries were unique, identical or modified, the average number of queries per session, the number of queries submitted per user, if and how subsequent queries were modified, the number of result pages viewed, number of terms per query. In a follow-up study based on a larger sample from the same source, Spink *et al*. (2001) repeated much of the above work and examined the mean number of queries submitted during a session, the mean number of queries submitted per user, if and how subsequent queries were modified, the number of result pages viewed, number of terms per query, and the distribution of the terms. In a subsequent study on data collected from the FAST search engine in 2001, Jansen and Spink examined similar things, e.g. the number of search result pages examined, and the temporal length of a session (Jansen & Spink, 2003). The authors concluded that while web searching still was IR it was a very different sort of IR, and they suggested that designers and researchers of IR tools should pay more attention to this fact. They found web queries to be short, not much modified and very simple in structure, but they reported that, despite short session lengths and short queries, web search engine users seemed to find what they were looking for (Jansen *et al*., 2000; Spink *et al*., 2001; Jansen & Spink, 2003).

Despite being carried out by the same researchers, there are also differences between the above studies that cannot be attributed to the different search engines alone. The variables analysed, the methods applied, and the presentation of the results vary as well. The reported number of queries per session seemed to be decreasing from 2.84 to 2.53 to little over 2 while the portion of single query sessions went from 67% to 48% to 53%. Whether the number is going up or down is difficult to say since the first two measures are from the same year. The number of terms used per query increased from 2.21 to 2.4 and the portion of single term queries dropped from 58% to 26% between the Jansen *et al*. (2000) and the Spink *et al*., (2001) papers. However, again both these values are based on data from 1997 and one can therefore not speak of a trend. The percentage of *unique queries*, i.e., "differing queries entered by one user in one session" (Spink *et al*., 2001, p. 227), submitted was difficult to compare because the two studies measured these values differently, but combining unique and modified queries in Jansen *et al*. (2000) makes it comparable with the numbers in Spink *et al*. (2001). Approximately 57% of the queries seemed to be unique queries, and roughly 43% seem to be identical queries, i.e. requests for new result pages.

Session length – both in terms of queries per session and in time – is another issue that varies between the studies. Jansen *et al*. (2000) and Spink *et al.* (2001) define a session as the entire set of queries submitted by a user over time without trying to measure this variable. Jansen and Spink (2003) also present session duration and measure it as the time from the user's first query until the user leaves the search engine for the last time, ending up with a mean session length of 2 hours, 21 minutes and 55 seconds. However, they also note that 52% of all sessions have duration of 15 minutes or less and that 26% of all sessions are less than 5 minutes. In these studies, Spink, Jansen and others are implicitly suggesting that all user interaction with a search engine (during a single day, presuming, since they all look at one day's worth of data) is to be understood as part of one continuous information seeking session, regardless of the time that may pass between consecutive interactions. Although explicitly suggested to be the common understanding of a search session (Jansen & Pooch, 2001), this definition is problematic. Instead, it seems at least equally likely that web users interact with the search engine several times during a day; perhaps with different information needs. In other words, users may have not merely one but multiple search sessions during a single day, and this needs to be acknowledge in the analysis of users' search behaviour.

In traditional IR systems, a session is clearly determined by login and logout times, but on the web such timestamps are not available (Han *et al*., 2001). An alternative way to identify and separate individual sessions is to find periods of inactivity between interactions and when the length of such an interval exceeds a threshold to regard that as a session delimiter. Some previous studies have used an idle interval heuristic of approximately 30 minutes (*cf*. Catledge & Pitkow, 1995; Choo *et al*., 2000; Liu & Zhang, 2004). However, the seminal work of Göker and He with colleagues has shown that although this approach makes it possible to determine session

boundaries with little or no manual effort, two types of error may occur. Firstly, related activities could wrongly be assigned to different sessions. This – referred to as a Type A error – occurs if the threshold is set too tight. Secondly, unrelated activities could wrongly be allocated to the same session. This – called a Type B error – happens if the threshold is too loose (Göker & He, 2000; Han *et al*., 2001; He & Göker, 2002). Spink, Jansen and colleagues, who assign all user queries from the same day into a single session, generate Type B errors. To avoid this problem it seems more useful to define a session as all queries from a single user pertaining to one particular interest and with a close proximity in time (Göker & He, 2000; Han *et al*., 2001; He & Göker, 2002). Analysing seven days of intranet search engine usage from March 1999, Göker and He (2000) found that the optimal session interval should be in the range of 11-15 minutes and when doing a similar analysis of 30 minutes of search activities from Excite logs from March 1997, Han *et al*. (2001) found the optimal interval be around 9 minutes. These findings suggest that also the 30 minute heuristic previously used is too long, resulting in Type B errors.

In our work we shall try to follow the methodological approach of Spink *et al*. (2001) (to produce results that are comparable) but use the more accurate session definition suggested by Göker and He (2000). Compared to other studies of users' search behaviour, our work differs in that we have studied not public search engine usage but intranet searching, since this context is even less understood. Compared to other studies of intranets, we have showed in a previous account (see Stenmark, 2005b) that the little work that has been carried out on intranet searching has not been aimed at understanding user behaviour but on aspects such as session boundary detection (Göker & He, 2000), performance tests (Hawking *et al*., 2000), rank aggregation algorithms (Fagin *et al*., 2003), or query expansion (Stenmark, 2005a). How intranet users interact with their search tools is yet unknown and the primary aim of this work is therefore to explore and describe, and thereby to establish a baseline for other studies of intranet searching. We shall also compare our findings to what has been reported for the public web. If intranet searchers are very similar to public search engine users, the body of research gathered can be applied to intranet tools as well. However, if intranet searches are different just as web searches differ from traditional IR searches, then we must conclude that IR tools for corporate webs must be studied and developed separately. This work therefore is of interest for both academia and search engine vendors.

Research site and method

In the following section, we account for the search engine and the context in which it operates, and for the research approach used for this work.

The SwedCorp intranet

This study is based on data obtained from the SwedCorp intranet. SwedCorp is a Swedish manufacturer of commercial vehicles with offices and factories in many countries around the world. In 2002 there were approximately 60,000 employees in the company group, which consisted of nearly a dozen individual companies. All there companies had shared access to the intranet and we shall in this paper treat SwedCorp as one company. A substantial portion of the employee was blue collar workers without access to individual computers. Instead, they were reduced to use information kiosks located in the assembly plants, whereas the white collar workers typically had individual computers at their desks. The SwedCorp intranet was started in 1995 and did in 2002 consist of more than 1,500 web servers. The exact amount of documents (or web pages) available on the intranet was impossible to determine, but the search engine reported to have indexed 743,826 documents. This corpus consisted of HTML documents (approx. 80%), PDF documents (~15%) and MS-Office documents (~5%). Content was typically work-related and provided in a top-down fashion, i.e., a relatively small group of informants were assigned the responsibility to publish official or semi-official information. Very few employees had individually homepages and very little information was shared horizontally or on a peer-to-peer basis. Where homepages existed, they typically contained official information such as name, address and phone number of the employee, his or her official title, and the primary area of responsibility. No personal information would appear.

Since 1998 SwedCorp uses Ultraseek[1] as their intranet search engine (see figure 1 for an image of the interface). Ultraseek does not accept Boolean operators such as AND or OR but instead allows the use of + (plus) and – (minus) to indicate that a term MUST or MUST NOT appear in the document. Quotation marks are used to indicate a string search and all these features may be combined. For example, the query *apple –mac "fruit salad"* would mean a search for the word apple, but not the word mac and the phrase "fruit salad". Results are returned in chunks of 10 where the user may access the next chunk by clicking the "next" button.

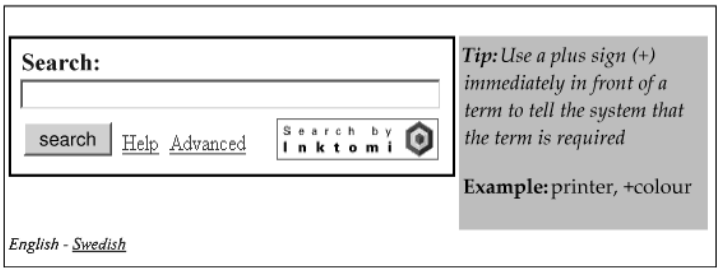

Figure 1. The search interface

Each submitted query resulted in an entry in the search engine log as illustrated below, where the IP-address, the date and time, and the query could be identified. The log also contained additional parameters allowing us, amongst other things, to determine whether the user submitted a new query or was examining another result page

157.171.141.113 15/Oct/2004:10:38:52 apple –mac "fruit salad" [*&parameters*]

Research method

The methodological approach has been chosen to be consistent with previous research in order to allow comparison and to extend the work already carried out by others. However, this is not easily achieved. As already observed (cf. Jansen & Pooch, 2001; Spink *et al*., 2001), there are no standardised metrics to collect, and methods, data definitions and analysis differ from study to study. The fact that our study is concerned with intranet users is yet another source of diversity. Spink *et al*. (2001) therefore suggest that comparisons should be looking at *similarities in trends* rather than comparing actual numbers, and we shall here account for our approach so that others can better understand and evaluate our observations.

The data analysed consisted of a search engine log file containing seven days of data from October 21st 2002 to October 27th 2002. The log files were sorted on IP-address and datetime, and the number of calls from each unique IP-address was calculated. The 20 most active addresses were examined to identify and remove obvious proxies (i.e., servers relaying queries from multiple users). After this modification, the dataset contained 26,205 activities, and this "cleaned" set was used in the subsequent analysis (see table 1 for details).

Table 1. Statistics from the dataset

Measured variable	value
Number of days covered	7
No. of unique IP addresses	5,644
No. of activities (total)	26,205
No. of activities per IP	4.64

Session boundary analysis: To determine session boundaries we set the inter-session idle time threshold to X minutes and ran a script that bundled activities from the same IP-address with a time difference of less than X. An *activity* is understood as either a query or a request for a new result page. By letting X assume all values from 0 to 16 and 20, 30, and 45 minutes, we received the graph illustrated in figure 2.

Figure 2 indicates that the number of sessions drop rapidly as the idle interval length increases from 0 to 3 minutes, but that the inclination falls off between 3 and 10 minutes to become almost flat after the 10 minute mark. We therefore selected an idle time threshold of 13 minutes, since this would be in the middle of the 11-15 minutes interval suggested by Göker and He (2000).

Using a 13 minutes threshold we analysed the data to determine the session length in terms of interactions per session, the temporal session length, and the distribution of the session lengths. Since the temporal length of a single activity session is impossible to determine, we only used sessions of two or more interactions for this part

of the analysis and calculated the mean time between two consecutive activities within a session. This value was then multiplied with the mean number of activities per session to calculate the temporal session length.

As an alternative method, we also calculated the actual time difference between the first and the last activity in each multi-activity session.

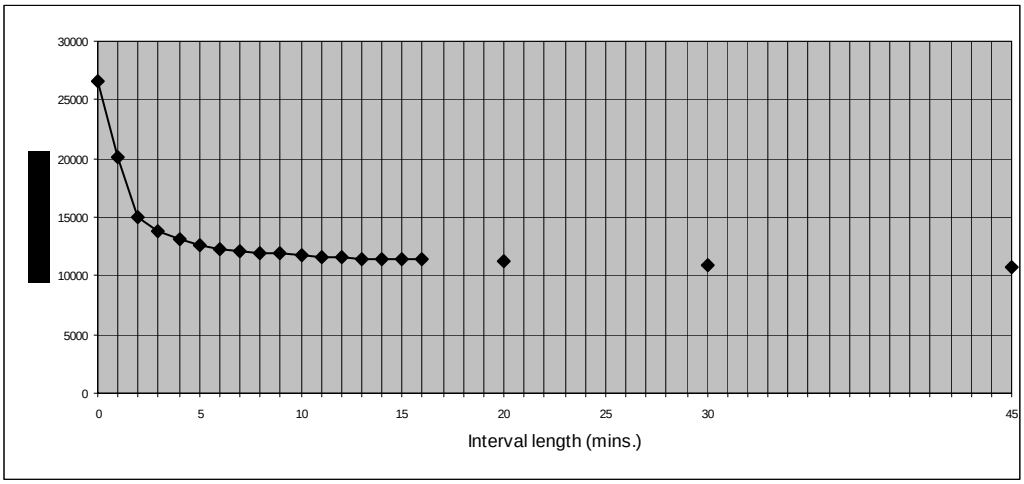

Figure 2. No. of search sessions as a function of inter-session idle intervals.

Query analysis: The Ultraseek search engine treats a request for the next result page as a new query, where the search terms are identical but the result-start (rs) parameter is incremented by 10 (i.e., the default number of displayed results per page). In order to analyse the queries we had to separate the actual queries from the result page requests by identifying all queries where rs was equal to 1. The total number of queries as well as the number of zero term queries were counted. We then analysed the distribution of the terms.

Result page viewing analysis: To analyse the viewing of search results, we returned to the cleaned sets, sorted them on IP-address and datetime, and then counted the number where rs equalled 1. This gave us the number of submitted queries. We thereafter identified series of consecutive queries with identical terms from the same address and counted the number and length of such series. All these manipulations were carried out automatically by scripts and checked for correctness by first being run on smaller subsets and later by manually comparing sub-totals and by checking in detail randomly selected IP-addresses.

Results

In the following section we present the results from our study by first accounting for the session analysis, thereafter the query analysis, and finally the result viewing analysis. Where appropriate, we report both mean and median values since the data was highly skewed.

Session analysis results

The dataset contained 11,419 sessions with a mean number of 2.29 activities per session. The number of activities per session ranged between 1 and 80 with a majority (61%) of the sessions being single activity sessions. The median number of activities per session was thus 1. The full distribution of activities per session is accounted for in table 2. Using only multiple activity sessions (i.e., sessions where at least two activities were recorded), the mean time between two consecutive activities was found to be 1 minute 44 seconds. Multiplied by the average number of activities per session this gives a calculated mean session length of 3:59 minutes. The longest multiple activity session actually observed lasted 1 hour 1 minute and 7 seconds and contained 64 activities. The mean actual session length observed amongst the multiple activity sessions was 4:31 whereas the median was 2:21.

Table 2. Distribution of activities per session

Activities per session	Occurrences	Percentage
1	7,025	61.5%
2	1,860	16.3%
3	888	7.8%
4	510	4.5%
5	282	2.5%
6	203	1.8%
7	136	1.2%
8	106	0.9%
9	67	0.6%
10	60	0.5%
> 10	282	2.5%

To allow for comparison with Jansen and Spink's (2003) data we wanted to use their session duration intervals and calculate the number of sessions in each. However, since they define a session as the time from a user's first query to their last, we would have to discard all single activity sessions and use the measured session length for the remaining sessions. This would not be meaningful since 61% of the sessions would have been thrown away.

Query analysis results

The number of zero term queries found amongst the 26,205 logged activities was 1,025 or 5.0%. In the following query analysis, all zero term queries have been discarded, leaving us with 25,180 non-trivial activities. The number of queries found in the dataset was 19,433 or 77.2% of the non-trivial activities. The number of repeat queries (i.e., requests for result pages) was hence 5,747 or 22.8%.

The average number of query terms per query was 1.40 (see table 4). Single term queries dominated with a total of 13.445 or 69.2% of all non-empty queries. Some 24% of the queries used two terms, while 5.3% contained three terms. No query contained more than 9 terms. The distribution is found in table 3 below and as can be seen from the table, the number of queries with five or more terms is almost zero, and there is a big drop in frequency already after one term.

Table 3. Distribution of terms per query

Terms per query	Occurrences	Percentage
1	13,445	69.2%
2	4,650	23.9%
3	1,021	5.3%
4	214	1.1%
5	67	0.3%
6	19	0.10%
7	10	0.05%
8	2	0.01%
9	5	0.03%
10	0	0%
>10	0	0%

Table 4 summarises the results so far and reports the number of unique queries, the portion of zero term queries, the number of repeat queries (i.e., result page requests), the mean number of terms per query, and the number and percentage of single term queries.

Table 4. Summery of query analysis results

Measured variable	value
Total activities	26,205
Zero term queries	1,025
Non-trivial activities	25,180
Real queries	*19,433*
Result page requests	*5,747*
Terms per query	1.4019
Single term queries	13,445
% of real queries	*69.2*

Result page analysis results

To analyse the result page utilisation, we returned to the cleaned set of 26,205 activities. The analysis revealed that there was a large span in how many result pages were examined and the values ranged between 1 and 67. However, the mean number of result pages examined was only 1.35 and our data showed that a vast majority or close to 91% of all users did not bother to check beyond the first result page. Only another 4% checked two pages and less than 5% went on to view three pages or more, as is evident from table 5.

However, in 81 cases the user examined more than 10 pages of results. Looking closer at these cases, we determined that they stemmed from 58 different users. The most active user was responsible for 10 of these 81 cases and had a total of 326 interactions with the search engine during the seven days of logging.

Table 5. Distribution of result pages viewed

No. of result pages examined	Occurrences	Percentage
1	19,145	90.8%
2	923	4.4%
3	400	1.9%
4	218	1.0%
5	121	0.6%
6	64	0.3%
7	56	0.3%
8	39	0.2%
9	26	0.12%
10	13	0.06%
> 10	81	0.4%

Discussion

We shall now discuss our results and, where applicable, compare them to previous findings from the study of public search engine use.

Search session length

Our study uses a more fine-grained and, we argue, more intuitive session definition than the one used in previous work by Spink and Jansen and furthers our understanding of users' search behaviour. We calculated the average session length to just less than 4 minutes and measured multiple-activity sessions to have an average of approximately 4:31 minutes. These results are difficult to compare to directly to the 2.4 hours mean reported by Jansen and Spink, but, arguing that our session definition is more accurate than those previously used, our results suggest that the search sessions may be much shorter than previously assumed. It remains unknown how much of this difference is due to the session definition and what can be attributed to the organisational context of an intranet. We speculate that web users "know" that the answer is out there and hence are likely to be more persistent, while intranet searchers give up quicker, assuming that the information they are looking for does not exist on their corporate intranet. Additional research is needed to test this hypothesis.

How many queries per session do the intranet users submit? As can be expected, short sessions also mean little activity. Spink, Jansen, and colleagues report the number of queries per session to be between 2 and 3 (and possibly decreasing). Our result of 2.31 is consistent with these findings despite the much shorter session length observed. A possible explanation for this result is the large number of single activity sessions: just over 61% in our study and 67% in Jansen *et al*. (2000). When users submit only one query and then leave, it does not matter what session definition you apply; the session will still only contain one query.

However, in another study, Spink *et al.* (2001) found the percentage of single query sessions to be only 48%, and that as much as 31% of all users entered three or more unique queries during a session. As can be derived from table 2, we found that only some 22% of our sessions contained three or more activities. In Jansen *et al*. (2000) the number is only 14%. The difference between our results and that of Spink *et al*., (2001) may be due to the fact that we calculate queries *per session* whereas Spink *et al*. count queries *per user*. As noted in Spink *et al*. (2001), the two methods are similar but not identical. To further add to the confusion, in Jansen *et al*. (2000), *queries per session* and *queries per user* are used alternately. With the definition of session used in this paper, it seems more logical to report on activities per session since a user may have several sessions, but again it makes comparison more difficult. Nonetheless, the conclusion remains that both intranet and public web users submit only one query before leaving.

Query construction

When it comes to the number of terms used per query, literature tells us that casual searchers use few terms: between 2 and 2.5 terms per query (Silverstein *et al*., 1999; Jansen *et al*., 2000; Spink *et al*., 2001). Our data suggests that intranet users' queries are shorter still. An average of 1.40 terms and more than 69% submitting single term queries is much lower than the numbers presented for public search engine usage. In our study, no query contained more than 9 terms. This is comparable with the 0-10 term range reported in Jansen *et al*. (2000) but much less than what is indicated in Spink *et al*. (2001), where fig. 4 suggests that some queries had over 100 terms.

However, previous studies have used English as the primary query language. In our study, most users were from Swedish (mixed with smaller groups of employees from France, the Netherlands, the United Kingdom, North America, and Brazil). Although an analysis of what languages were used to construct the queries is outside the scope of this paper, it can be assumed that a substantial portion of the queries were not in English. Different languages have different syntax and this may have affected the number of query terms. Terms that require two words in English, e.g., torque wrench, would in Swedish be written as a single compound word. It may therefore not be the intranet *per se* but the use of a particular (set of) language(s) that caused the drop in query terms. Further research is needed to clarify this.

The number of zero term queries found in our study was 5% of the unique queries, which is also what Jansen *et al*. (2000) reported. Other studies have reported numbers as high as 18% (Spink *et al*., 2001) but this may again be due to differences in definitions and analytic methods.

Result pages

It seems obvious that most casual intranet searchers do not bother to look beyond the first few pages of results. This finding is consistent with what Spink and Jansen has reported for web searchers (Jansen *et al*., 2000; Spink *et al*., 2001). However, in Spink and Jansen's data, some 43% of the activities were result page requests, whereas in our data, the corresponding number was less than 23%; a significantly lower amount. Our study has a much higher number of users looking only at one result page. Jansen *et al*. (2000) ask whether this behaviour is because the users indeed find what they are looking for or if they just give up easily. In a subsequent study, however, Jansen and Spink (2003) conclude that approximately half of the web pages actually examined by the users seemed to be relevant, suggesting that users of public search engines actually *do* find the information they were looking for. In our study, we cannot see which of the result pages the users clicked on and are therefore not able to do a similar analysis for intranet users. This would be an interesting issue for further studies, particularly so since it has been suggested that there is a difference between searching the public web and searching an intranet (Fagin *et al*., 2003). As of now, it is not possible to tell whether our users find what they are looking for quicker or give up more easily.

Additional comments

Although our results suggest that intranet users in general submit few and short queries, are reluctant to revise and resubmit their questions, and examine few result pages, there are also obvious exceptions. Our data shows traces of "superusers" whose sessions last over an hour and contains dozens of activities and who wade through tens of result pages. It is unclear, however, whether these heavy searches spend so much time with the tool because they are good at searching or if it is because they are unable to find what they are looking for. Future research may help us identify and categorise different types of intranet users and understand these groups' needs and preferences.

There are limitations to this research that needs to be recognised. Firstly, not only are intranets different from the public web but they can also be assumed to be very different from one another. Previous research (Spink *et al*., 2002) has pointed to the fact that there are regional differences on public web and it seems very likely that this holds also for intranets. This paper reports only from one intranet and we cannot know how much (if any) of this work that is representative for corporate searchers in general. Secondly, this study does not describe the context wherefrom the users operate, and we know nothing of their reasons for engaging the search engine. Log file data sometimes only allows us to form new hypothesis rather than produce reliable answers.

Conclusions

In this study we have explored and described the web search behaviour of corporate intranet users and tried to contrast this to what is known about search engine users on the public web. We have focused on sessions, queries, and result pages and conclude that there are both similarities and differences between how corporate users search their intranet and how public search engines are used. Regarding session length, intranet search sessions are similar to those previously observed for public web searchers in the sense that they are short. However, intranet search sessions are also different in the sense that they are much shorter than previously believed.

Query construction amongst intranet users is similar to how users of the public web go about, meaning that both groups use few search terms. However, the users in our study differ from the public users since they use much fewer terms than previously reported.

Finally, intranet searchers' use of search results is similar to what has previously been reported on web searchers inasmuch as they both look at very few result pages. However, intranet searchers differ from web searchers since a much larger portion seems to be examining only the very first result page.

A conclusion on a more general level is that more intranet studies are needed to fully understand the differences between searching the public web and searching a corporate intranet. Some new and open research questions have been suggested in this paper. We also conclude that standardised metrics and methods would be helpful since it would make comparisons easier and simplify the analysis work. We hope that such agreements will emerge as the field of intranet search studies matures.

Acknowledgments
We thank SwedCorp for giving us access to their search engine logs and Swedish Council for Working Life and Social Research for funding this research via grant #004-1268.

References

Catledge, L. and Pitkow, J. (1995). Characterizing browser strategies in the world-wide web. *Computer Networks and ISDN Systems*, *26*, *6*, 1065-1073.

Choo, C. W., Detlor, B., and Turnbull, D. (2000). *Web Work: Information Seeking and Knowledge Work on the World Wide Web*. Kluwer Academic Publishers, Dortrecht.

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, 366-375.

Gerstner, J. (2002). Intranets mean Business, *Communication World, 19*, *2*, 14-17.

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, 319-322.

Han, S., Göker, A. and He, D. (2001). Web user search pattern analysis for modeling query topic changes. In *Proceedings of User modelling for context-aware applications, a workshop at the 8th International Conference on User Modeling*, Sonthofen, Germany, July 13-17.

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

He, D. and Göker, A. (2002). Combining evidence for automatic web session identification. *Information Processing and Management*, *38*, 727-742.

Hölscher, C. (1998). How Internet experts search for information on the Web. In *Proceedings of WebNet '98*, Orlando, FL.

Jansen, B. and Pooch, U. (2001). A review of web searching studies and a framework for future research. *Journal of the American Society for Information Science*, *52*, *3*, 235-246.

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

Jansen, B. and Spink, A. (2003). An Analysis of Web Documents Retrieved and Viewed. In *Proceedings of ICIC´03*, Las Vegas, NE, 65-69.

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.

Liu, J. and Zhang, S. (2004). Characterizing web user regularities with information foraging agents, *IEEE Transactions on Knowledge and Data Engineering*, *16*, *5*, 566-584.

Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. (1999). Analysis of a very large web search engine query log. *ACM SIGIR Forum*, *33*, *1*, 6-12.

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

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.

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

Stenmark, D. (2005a). Query expansion on a corporate intranet: Using LSI to increase relative precision in explorative search. In *Proceedings of HICSS-38*, Big Island, HI., January 3-6.

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

---

[1] Ultraseek is a commercial product now provided by Verity, Inc. (see http://www.verity.com/products/ultraseek/). At the time of the study Ultraseek was still owned by Inktomi.