INTRANET USERS' INFORMATION-SEEKING BEHAVIOUR: AN ANALYSIS OF LONGITUDINAL SEARCH LOG DATA

Dick Stenmark & Taline Jadaan

Viktoria Institute, Göteborg, Sweden

Abstract

Today's knowledge workers rely increasingly on information to get their job done, and the availability of search engines to locate relevant information is thus essential. Understanding how users interact with search engines is a prerequisite for the successful design of useful systems and a body of knowledge has in recent years begun to compile. However, all previous studies have focused on the public web, not acknowledging the fact that much business-related information seeking occur on corporate internal networks. In this exploratory study, we have collected and analysed intranet search engine log files from three different years – 2000, 2002, and 2004 – enabling us to detect shifting trends in intranet search behaviour. Comparing our data to what has been reported from the public web we conclude that intranet searchers are both similar to and different from searchers on the public web. In sum, it appears that intranet users are more extreme in their behaviour and that qualitative studies are needed to understand the motives and rationales governing their actions.

Keywords: intranet search engine usage, log file analysis.

Introduction

The growing amount of electronic information that has become available over the last decades, in particular with the advent of the Web, has underlined the need for efficient and effective information retrieval (IR) tools. Spink (2003) notes that people are spending increasing amount of time working with various sorts of electronic information and that web-based IR tools such as public search engines have become everyday tools. This is particularly so for knowledge workers, since a knowledge worker is someone who interacts knowledgeable with information and sees information not only as something derived from knowledge but as something that changes knowledge (Schultze, 2000; Sellen *et al*., 2002).

Alongside the development of the technology per se, another stream of research is focusing on the users' behaviour when interacting with technology. In resent years, studies of users interaction with Web search engines has begun to gain momentum and a useful body of knowledge is beginning to compile (e.g., Silverstein *et al.*, 1998; Jansen *et al.*, 1998; 2000; Jansen & Spink, 2003, Spink *et al.*, 1999; 2001; 2002), although Sarasevic in the foreword of Spink and Jansen's 2004 book point out that research on information behaviour on the Web is still in its infancy. To better understand users' behaviour is important since it helps the users form better strategies, it helps designers construct better interfaces, it informs developers when constructing new IR tools and it has implications for the design of the Web.

It has been established that web searchers differ from users of traditional IR tools (Spink *et al.*, 2001; Jansen & Spink, 2003). Web users in general have limited understanding of search engines and of the structure of the web and they have developed a "quick and dirty" approach to searching, where they submit few and short queries and seldom browse beyond the first two result pages (Spink, 2003). In

our research, we have focused on a neglected subset of the Web searchers – the intranet users. Intranets, i.e., corporate internal webs, are at the same time both very similar to and very unlike the public Web (Fagin *et al.*, 2003). The technologies used, i.e., web servers, browsers and the HTTP protocol, are identical but the intranet's much smaller size enables approaches that cannot be applied to the Web. In addition, searching becomes very different due to the different social forces that affect the two environments. Fagin and colleagues argue that search strategies that work on the Web may not be successful on an intranet and vice versa. It is therefore important to study intranet searching in its own right and this paper is the result of such an effort.

The results reported herein are part of an ongoing research project. As with many other studies of search engine usage, we have analysed log files. However, where others often have used data from a single day or part of day and hence given a snapshot of the situation at a given moment in time, we have collected and studied log file data from a much longer time period, thus enabling us to identify changes in user behaviour over time. This will yield an important contribution to our understanding of search engine users' behaviour.

In the next section we shall account for some related work that we use as a baseline for our own data. As we shall show, very few longitudinal studies of web search behaviour have been published and little is known about intranet search engine usage so the results reported in section four shall primarily be compared internally to detect changes and possible trends and secondarily to data from studies of public search engines. This discussion is found in section five. Section six concludes the paper with some implications for further research.

Related work

Examining the use of web search engines, our work is obviously related to the body of knowledge on the use of public search engines that has complied over the last eight or so years. As opposed to the many studies of information seeking behaviour that have been conducted in controlled laboratory settings, this is a naturalistic study where we have examined real users' real queries in their real business environment, inspired by Spink *et al.*'s (2001) work. Over the years, a number of interesting log file variables has been identified. Recognising that different tools produce slightly different log file data, we have tried to find a common set of variables reported in many papers on web search behaviour that are present also in our log files. These are session length (temporal and in terms of number of activities), number of search terms per query, and number of result pages viewed.

Our work is particularly called for since it is targeted towards intranet users' search behaviour. There are two reasons for this; firstly, intranets are becoming increasingly ubiquitous and implemented at most if not all larger organisations, and, secondly, intranets are largely neglected as a research field. Fagin *et al.* (2003) and Stenmark (2005a) have argued that there are significant differences between how intranet users and public Web users behave. These differences stem from the different social forces at play and the different expectations fostered by the different cultures. Since these two environments serve different purposes, it seems likely that the nature of the queries submitted differs, Fagin *et al.* suggest. Therefore, intranets should be examined in their own right but there is surprisingly little research devoted to intranet seeking and how intranet users seek information using search engines.

Fagin *et al.* (2003) studied the problem of intranet search on IBM's intranet, by implementing and testing an intranet search engine. However, their focus was technical matters; the use of rank aggregation and the effects of different heuristics on ranking of search results. They did therefore not explicitly study the employees' use of the search engine or on how they searched for information.

Göker and He (2000; He and Göker, 2000) examined a weeks worth of log file data from Reuter's intranet search engine. Their primary concern was to develop a methodology to automatically detect search session boundaries in web log files. Although their work is useful (and applied in our research

as explained in the following section), they did not study users' information seeking behaviour *per se*, and their paper does not teach us anything about intranet users' search engine usage.

Hawking *et al.* (2000) implemented a search engine on a university intranet in order to "reality test" a document retrieval software that had performed well in laboratory test at TREC (Text REtrieval Conference). Apart from general observations such as queries are short and spelling errors common, their work did not reveal any details about the intranet users or how they searched.

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, varying from undirected viewing (without a specific information need) to formal searching (for action or decision making). They engaged in multiple methods of qualitative and quantitative data collecting, including questionnaires, interviews and the use of a client-side application, and found that different seeking modes are distinguished by the nature of information needs, information seeking tactics, and the purpose of information use. However, they did not tell us anything specific about intranet search engine usage.

Due to the lack of relevant literature on intranet search engine usage, we shall compare our results to what has previously been reported on public web searching. In particular we shall use the work of Spink, Jansen and colleagues (e.g., Jansen *et al.*, 1998; 2000; Jansen & Spink, 2003, Spink *et al.*, 1999; 2001; 2002; Spink and Jansen, 2004) who stand out as those who most consistently have contributed to this research field.

Context and method

This research is based on analysis of a set of search engine log files from MechCo's intranet. MechCo is manufacturer of commercial vehicles with offices and production plants in many countries around the world and employs some +80,000 people, although the exact number varies over time. MechCo's intranet, which connects all the companies in the group, was established in 1995 and quickly developed into a large information repository. In 1996, little over 100 web servers were known to exist and MechCo realised that their employees needed help with searching. In 1998 MechCo purchased and implemented a commercial search engine from Infoseek Inc, 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 MechCo 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 address of the user's computer, the date and time (datetime) of the transaction (as logged by the server using Central European Time), the query string as entered by the users, information regarding which result page the users have requested, and some additional parameters not used in this particular study. As with most computer software, the search engine has had several version upgrades since 1998 and this has affected the log files in the sense that the more resent ones have parameters not present in the older logs. In our study, however, we have concentrated on the fundamental data described above, and deliberately avoided to include information unique to a particular log file.

Most previous studies of web search engine usage have focused on public search engines and therefore often been forced, by the sheer amount of data, to limit their analysis to (portion of) a single day. Although quite a large number of such studies have been published, they report discrete snapshots, and Cothey (2002) and Spink and Jansen (2004) observe that longitudinal studies in information retrieval in general and of Web search engine usage specifically are very limited but highly needed. Cothey's (2002) study over a 10-month period is one of the longest study durations reported on Web users' information-seeking behaviour. In this study we attempt to make a substantial contribution in this area and have therefore collected data samples from three different years; 2000, 2002 and 2004. Although

the information environment has evolved, the user population has changed, and the tool has undergone version upgrades during these years, we believe that this study qualifies as a longitudinal study since such changes are inevitable when studying web environments and the data is aggregated and discussed at the collective level and not per individual.

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. Table 1 below shows some basic log file statistics.

	2000	**2000 (week #3)**	**2002**	**2004**
Number of days covered	25	7	7	7
No. of unique IP addresses	10,973	4,227	5,644	6,822
No. of activities (total)	75,064	19,992	26,200	27,024
No. of activities per IP	6.84	4.73	4.64	3.96
No. of activities per day	3,003	2,856	3,743	3,861

Table 1. Comparing the numbers from the different log files. Since year 2000 covers more days we have also reported the numbers for the third of the four weeks.

A common approach in our work has been to use scripts to filter and summarise the data. The resulting files have thereafter been fed into MS-excel where data has been sorted and various statistics have been calculated. This process has been repeated in an iterative way. As explained earlier, we have concentrated on sessions (and activities within a session), queries and result pages, and we shall now explain how these aspects have been derived and analysed.

Session boundary analysis: Unlike traditional systems where you log in and out, the web does not provide session boundaries. A simple substitute for well-defined session boundaries is to treat all interactions from the same user (and day) as a session and this is indeed a method adopted by many researchers (*cf.* Jansen *et al.*, 1998; Spink *et al.*, 2001; Cothey, 2002). It is, however, problematic to assume that all activities from the same user belong to the same session, especially if the log file covers different days (He and Göker, 2000). A commonly used alternative is to look at the idle interval between two consecutive activities from the same user and if this interval is "long enough" considered it a session break. This approach, often referred to as the "timeout" method (Huang *et al.*, 2004), is used in our study. The challenge, obviously, is to determine what threshold is "long enough" since this value varies between contexts and between users. As reported in our previous work (Stenmark, 2005b; c), we use the approach advocated by Göker and He (2000) and plotted how the number of sessions decreased with increasing interval length (see Figure 1). This will address the context aspect of setting a threshold value but obviously not handle the individual aspect. We still believe this to be a useful trade-off since it makes possible automatic processing of the log files.

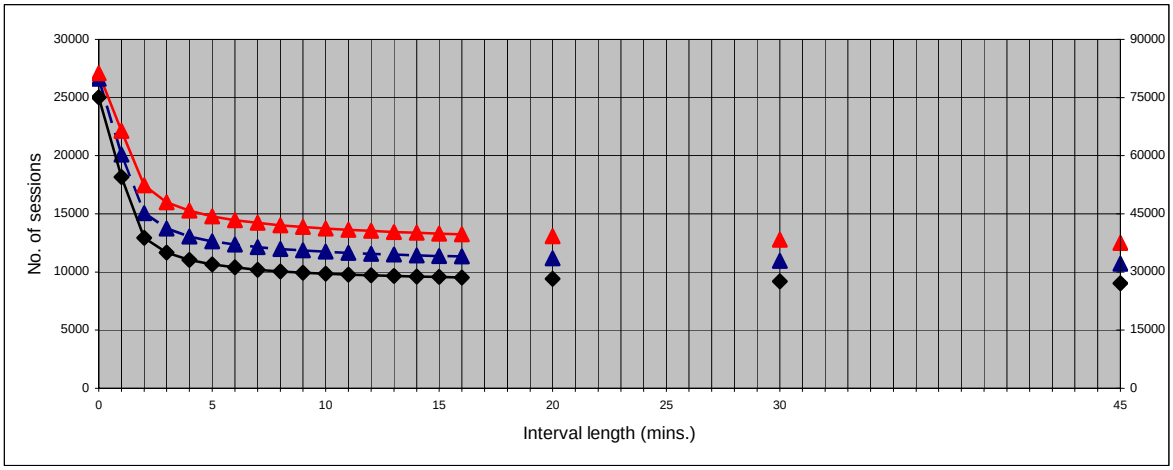

Figure 1. *Number of sessions as a function of idle interval length for the three log files. The top line (2004) and the middle line (2002) use the left-hand axis whereas the low line (2000) uses the right-hand axis.*

As can be seen in Figure 1, the three log files show very similar patterns and we therefore stayed with the 13 minute threshold used in our previous work. Based on the 13 minute threshold, we could identify individual sessions, calculate session length in terms of activities per session, estimate the temporal length and determine the session length distribution. When separating the sessions we detected and corrected a program bug in the script language used, which means that the session-related numbers reported in this paper differs from what has been reported previously (Stenmark, 2005b; c) although the conclusions still hold. We estimated the average temporal length of a session by calculating the average interval length between two consecutive activities and multiplied this with the average number of activities per session. We also recorded the average length of all sessions with more than one activity by calculating the difference between the last and the first activity.

Activity analysis. Our log files hold records of two types of activities; queries submitted and result pages browsed. We define a query as the activity where the user enters text into the input box and sends this to the search engine. The search engine responds by retrieving a number of links (usually 10) to indexed documents matching the query terms and displaying these to the user. The user may look at these results, decide to click on one of the links, or leave the search engine. Neither of these activities are part of this analysis. The user may also request the next set of (10) links, usually by clicking on a Next button. This activity produces a log file entry which we refer to as a result page request. These two actions are together referred to as activities, and used in our work.

Query analysis. The search engine logs all user queries as they are submitted by the user. Sometimes the user clicks on the submit button without having written anything in the input box. This is referred to as an empty query. After having counted the number of empty queries, these entries were removed from the logs and we used scripts to calculate the average number of terms per query and the distribution of the terms.

Result page analysis. When submitting a query, the web search engine automatically returns a result page with (typically) the 10 highest ranked results. The user may thereafter click on the Next button to get the next bunch of (10) results. The first activity is in our research classified as a query whereas the second activity is classified as an explicit result page request. The files were sorted on IP-address and datetime and we used scripts to determine the distribution of the result page requests. Empty queries, which result in empty and thus uninteresting result pages, were removed from the log files prior to analysis. This was not done for previously reported results and the data in this paper therefore differs (marginally) from what we reported at last year's ASIS&T conference (Stenmark, 2005c)

Results

We shall here present the result of our log file analysis. The results have been organised according to the central themes that we have focused on; session activities, search terms, and result page views.

Session activities

Using the 13 minutes idle interval threshold, we could identify separate sessions. The number of sessions and the calculated duration of a session are presented in table 2. Note that the 2000 log file contained data from 25 days, whereas the other two log files contained data from 7 days. The number of sessions from one week in year 2000 would be approximately 7,361 (26,290 x 7 / 25).

	2000	**2002**	**2004**
No of. sessions	26,290	10,432	12,368
Average no. of activities / session	2.855	2.512	2.185
Average session length for sessions having more than one activity	4:46	4:49	4:41
Average time between two consecutive activities	1:36	1:44	1:52
Calculated session length (row 2 x row 4)	4:33	4:22	4:04
Max. no. of activities in a session	145	81	54

Table 2. Comparing the session related data from the three years.

From Table 2 we see that the average number of activities per session decreases steadily from almost three in 2000 to little over two in 2004. Also the maximum number of activities logged in a single session decreases. At the same time, the average time between two consecutive activities increases equally steady from 1 minute 36 seconds in 2000 to 1:52 in 2004. At the same time, the calculated session length (that we got by multiplying these two variables) decreases from four and a half minute to just over four.

As can be seen from Figure 2, regardless of year most sessions consisted of a single activity, i.e., a submitted query. We can also see that the percentage of single activity sessions increases steadily over the years whereas the percentage of multiple activity sessions decreases, particularly for sessions with more than 10 activities. In 2000, we had 1053 sessions with more than 10 activities, which corresponds to 4% of all sessions. The corresponding numbers for 2002 and 2004 are 301 (2.9%) and 217 (1.8%), respectively.

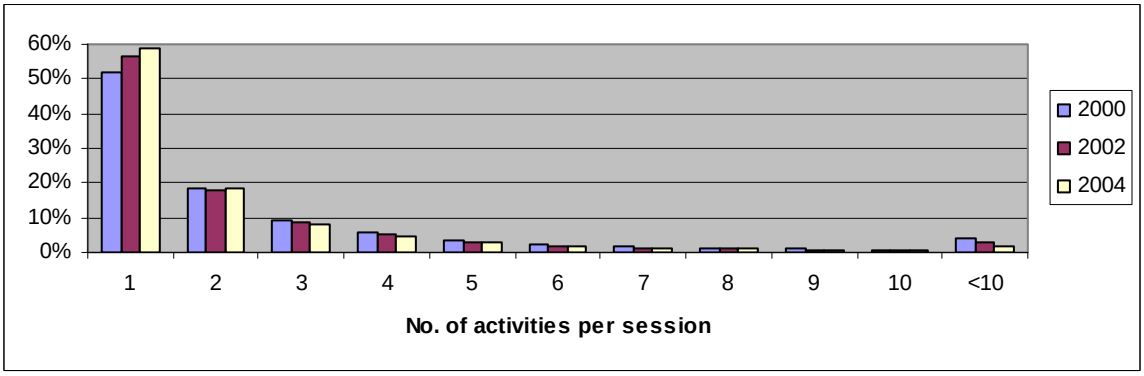

Figure 2. *The percentage of the sessions that contained n activities. As can be seen most sessions consisted of only one activity (a single query)*

Search queries

The number of queries found in the log files is reported in table 3, where we also account for the number of empty queries, i.e., queries without any search terms. The queries constitute 74.7%, 78.1%, and 85.7% of the total number of activities for the three years, and show an up-going trend. The maximum numbers of terms used in a query was higher in 2000 than in the latter years, as can be seen in table 3. The percentage of empty queries has gone from 4.9% to 5.0% and then dropped to 0.7%. The average number of terms per query does not vary that much; it made a 3.6% drop in 2002 but recovered to approximately the same value again in 2004.

	2000	**2002**	**2004**
No. of activities	75,064	26,200	27,024
No. of queries	56,041	20,458	23,161
No. of empty queries	2,727	1,025	156
No. of single term queries	46,940	13,445	17,369
Average terms per query	1.4538	1.4019	1.4479
Max. no. of terms in a query	14	9	9

Table 3. *Comparing the query related data from the three years.*

Figure 3 shows the distribution of the query length. The percentage of single word queries is between 67% and 69% and pretty consistent over the years. As with the numbers of terms per query, the data from 2000 and 2004 is almost identical whereas 2002 shows a slightly different pattern; more single term queries and fewer multiple word queries. Figure 3 also reveals that regardless of year very few queries contained more than 3 words. The numbers for the three years are 2.2%, 1.6% and 2.3%, respectively.

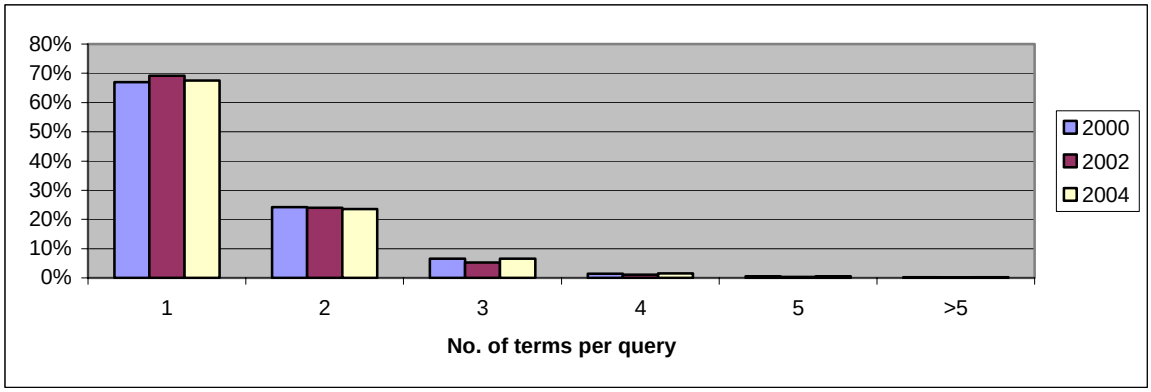

Figure 3. *The percentage of the queries that contained n terms. As can be seen most queries consisted of only a single term*

Result pages viewed

We found that explicit result page requests constituted 25.3%, 20.7% and 13.9%, respectively, of all activities. There is a large span in the number of result pages viewed, though. Although most users, regardless of year, looked at only one result page, the maximum value found for each year was 65, 67 and 24, respectively. Number of result pages viewed, the maximum value and the average value per year is reported in Table 4. As can be seen from Table 4, the mean number of result pages viewed is decreasing over time.

	2000	**2002**	**2004**
No. of activities	75,064	26,200	27,024
No. of (explicit) result pages	19,022	5,420	3,770
Mean no. of result pages viewed	1.34	1.26	1.16
Max. no. of result pages viewed	65	67	24

Table 4. Comparing the query related data from the three years.

In figure 4 below we show the distribution for the result page views. An increasing portion of users only viewing the first result page (i.e., only submitting a query) can be noticed. The number has risen from 89% in 2000 to over 92% in 2004. The portion of users viewing beyond the first result page has dropped consistently for each year examined. In particular, the portion of users viewing more than 5 result pages has dropped from 2.0% in 2000, via 1.3% in 2002, to 0.7% in 2004.

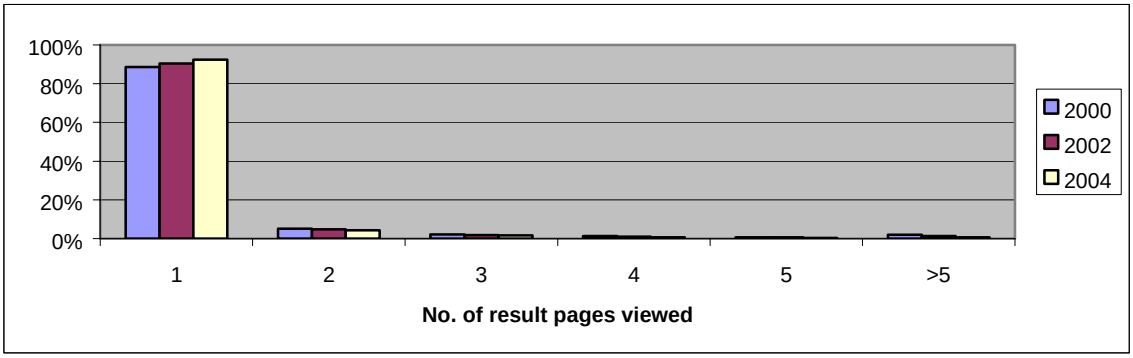

Figure 4. The percentage of users who examined n *result pages. As can be seen most users did not bother to look beyond the first result page*

This concludes our result section and we shall in the next section discuss these results, compare them with one another, and with results from public search engine studies.

Discussion

Having collected search engine log file data from three different years – 2000, 2002, and 2004 – we are able to compare how search behaviour on a corporate intranet has changed over time. To our knowledge, no such longitudinal study has previously been published and our work therefore is a unique contribution to the field. The general statistics in table 1 suggest that that the number of unique users accessing the search engine per week is growing; the numbers have increased by 61%. In contrast, the number of activities *per user* has dropped during the same timeframe, suggesting that usage has changed from a smaller number of more active users to a larger number of not so active users. One could speculate that early users were more influenced by the excitement of the new technology and thus perhaps more exploratory and playful whilst today's users have matured in their

behaviour and searching the intranet has become a mundane part of work and is therefore carried out in a more goal-oriented way.

Although our work is explorative and does at this stage not attempt to address a specific research question we believe our result to be an important contribution. A rich understanding of users' interaction with (intranet) search engines is a prerequisite for the successful development of new technical features, design of useful interfaces, structure of web sites, and improved end-user behaviour. Below, we shall further analyse the changes that can be noted, discuss trends in intranet search behaviour and compare our findings to what is known about public web searching. We shall organise the discussion according to our major themes.

Session activities

It appears that intranet searchers engage in fewer and fewer activities during their sessions. The average number of activities per session has dropped 23% in four years. There are always users who are above average when it comes to number of activities, but we see a trend that these highly active users also engage in fewer activities; the number (and percentage) of users with more than 10 activities dropped significantly. It seems safe to conclude that sessions contain fewer activities over time.

Although the average time between two consecutive activities has increased by 17% during the four years, the calculated temporal length of a session decreases steadily. In contrast, the observed temporal length for sessions with two activities or more remains rather fixed at 4 minutes 45 seconds. Our interpretation of this is that a growing majority of users engage only in single activity sessions, but those who perform multiple activities show a consistent session length.

Jansen and Spink (2003) note that the temporal session length is often not measured when analysing public web logs but it is *assumed* to be short. When analysing 24 hours of log data from FAST's AlltheWeb.com search engine, Jansen and Spink (2003) measured a mean session length of almost two and a half hour and noted that more than 8% of the sessions lasted longer than 4 hours. Such data is difficult to compare to ours, since Jansen and Spink did not identify session boundaries but bundled all activities from one user into a single session. It can be questioned if such a session length definition has any analytic value, particularly so when the authors themselves note that it seems plausible that sessions are short. Jansen and Spink's result is however the only session length reported so we do not know if the trend we have observed on MechCo's intranet can be seen also on the public web.

The ratio of single activity sessions measured in our data was high and increasing, going from 51% (2000) to 59% (2004). Meanwhile, the maximum number of activities logged for a single session decreases radically from 145 to 54; a 63% drop over the four years. It seems intranet users are becoming more reluctant to engage in long interactions with the search engine. This trend seems to conflict with Spink and Jansen's (2004) observation for U.S.-based web search engines that the number of single query sessions was trending down. AltaVista went from 77% in 1998 to 47% in 2002 and Excite from 60% (1999) to 55% (2002). However, AlltheWeb – a Europe-based search engine – logged a 6% increase in number of single query sessions from 2001 to 2002, and the number of queries per session dropped from 3.0 to 2.8 (Spink and Jansen, 2004). Again, different methods were used producing these results; the longer sessions reported by Spink *et al.* (2002) was due to the inclusion of result page requests. So, it is difficult to say whether the increasing amount of single activity sessions noted in our study corresponds to trend on the web. A more reliable way of measuring search sessions would be needed, and the use client-side applications to capture this metric would perhaps be a viable approach (cf. Choo et al, 1998; Jansen, 2005; Jansen and Pooch, 2005).

Search queries

What is immediately striking is that the average number of search terms per query is significantly lower in our study than what has previously been reported for the public web. Our average of 1.4

terms is approximately a whole word less than for public search engines. It has previously been suggested that this may be due to different languages used. Although Spink and colleagues do not explicitly report this we can assume that the queries examined in their studies were primarily in English, whereas a sizeable portion of the queries in our study can be expected to be in languages other than English. Another explanation would be that intranets are jargon-heavy and relies heavily on acronyms and abbreviations (Fagin et al., 2003), and this would also affect the number of terms used.

Based on our data, we claim that the number of query terms stay rather stable over the years. The average number of query terms dropped 3% between the years 2000 and 2002, from 1.45 to 1.40, but was back up to 1.45 again in 2004. In comparison, Spink and Jansen (2004) report that the query length for Europe-based search engines appears to be decreasing whilst the opposite seems to be the case for U.S.-based engines. European AlltheWeb.com goes from 2.4 terms per query in 2001 to 2.3 in 2002 and the number of single term queries increases by 11%. U.S.-based Excite shows a small increase in query terms (around 2.4-2.6) and the number of single term queries drops 5% from 26.6% in 1997 to 25% in 2001. With MechCo being mostly a European-based company, one might have expected a decrease in search terms, but we cannot see such a development.

The number of single term queries dominates across the years and our 67% and above is consistently higher than the numbers reported for the public web. For example, Jansen *et al.* (2000) had 31% single term queries in their 1997 data and Spink *et al.* (2001) report 26.6% single term queries in their study, also based on 1997 data. Our data is more recent and we have not been able to find up-to-date reports from the public web. As with the average number of query terms, our data suggests that the ratio of single term queries remains rather stable over the observed four-year period, although small variations can be noticed.

The number of empty queries on the public web has almost disappeared. In 1998, more than 20% of all queries submitted to AltaVista were empty queries and Excite had more than 7% empty queries in 1997. In 2001, Excite's numbers were down to 0.01% and in 2002 the number of empty queries on AltaVista had dropped to 0.03% (Spink and Jansen, 2004). This suggests that public search engine users have learned not to submit empty queries. We can see a similar trend on MechCo's intranet where the percentage of empty queries was around 5% before almost disappearing in 2004. The span in number of query terms ranges from 0 to 14 terms, which is consistent with Jansen *et al.*'s (2000) findings. The drop in maximum number of terms used from 14 to 9 between 2000 and 2002 is a significant 36% and during the same interval the percentage of queries containing more than 3 words decreased from 2.2% to 1.6%, a decline of 27%. However, the number of maximum terms has thereafter remained at 9 and the ratio of queries of 4 words or more has climbed to 2.3% in 2004. Therefore, we cannot speak of a clear trend.

Result pages viewed

Judging from our data, there appears to be an accelerating trend towards intranet users viewing less and less result pages. The overall number of explicit requests for result pages dropped from 25% in 2000 to 14% in 2004 and a similar pattern can be seen for the maximum number of result pages viewed. We do not know why there is a more pronounced drop in 2004 for these variables. In contrast, the mean number of result pages viewed has decreased more evenly with a drop of 6-9% between each logging.

The rapid drop in interest in result pages that occur after the first page is obvious and echoes the behaviour seen on the public web. However, our intranet users are even less tolerant for wading though result pages than are users of the public web. Spink and Jansen (2004) note a significant decrease in interest between the first and second result page but also between the second and the third. In their studies, they report that between 7.5% and 18% of the users look at two pages. Our numbers are in the 5% range.

In addition, whilst we notice a consistent trend towards more users examining only the first result page, findings from the public web show mixed results. On the one hand Spink and Jansen (2004) report an increase in users viewing only the first result page for Excite; the number has gone from 66% in 1997 to over 84% in 2001, a growth of almost 28%. On the other hand, they notice a drop from 85% in 1998 to 73% in 2002 for AltaVista and similar numbers (83% and 76%, respectively) for AlltheWeb from 2001 to 2002. Nonetheless, the authors conclude that there is an increase in users viewing only one result page (Spink and Jansen, 2004), and this concurs with our findings.

The reason for the decreasing viewing of result pages remains unknown but one explanation suggested is that the search engines ability to relevance-rank the results have improved, thereby eliminating the need to browse beyond the first set. This hypothesis is still untested. Another possible explanation is that the users spend more time examining each result page and thereby may be able to find information satisfying their needs without going beyond page one. Although not explicitly tested, this hypothesis has some support by the fact that users now on average spend almost 2 minutes on a result page before taking new actions.

Limitations and future studies

Although our data covers a period of 4 years, it is limited to a single intranet in a single organisation. Without similar studies from other organisations we cannot tell whether these findings are unique to MechCo or if they apply to intranets in general, even though we do not believe MechCo to stand out in any way. Nonetheless, we encourage others to copy our study in other companies and in other parts of the world and we hope our study will serve as an inspiration and a useful baseline for such studies.

A quantitative method such as log file analysis can only answer *what* and *when* questions, but not *why*. The motives and reasons behind the behaviours that we report here are important to understand before search engine developers, interface designers and web site owners can derive useful design implications from our work. Consequently, we are currently in the process of setting up a second phase of more qualitatively oriented studies to follow up our findings.

Conclusions

In this paper we have analysed and compared three intranet search engine log files from the same organisation but from three different years. This has enabled us to detect longitudinal changes in intranet users' search behaviour and these changes have been compared to findings reported on public web search engine users, whenever possible.

When it comes to activities our data clearly shows that our intranet users engage in fewer and fewer activities, resulting in shorter sessions. Those who engage in multiple activities, however, have a rather stable session length at just less than five minutes. Unfortunately, not much is known about the session lengths on the public web since this variable is seldom measured.

Although the number of search query terms fluctuates somewhat over the years, it remains relatively stable at 1.4 terms per query. What we can therefore conclude that our intranet users submit significantly shorter queries than do users of public search engines, and that the queries can be expected to continue to be short in years to come.

Regarding result pages, finally, we conclude that an increasing portion of the intranet users view only the first set, and that the viewing of result pages beyond the first set is rapidly decreasing. This appears to be true also for the public web, albeit the pattern is more obvious on the intranet.

In all, we suggest that intranet search engine usage is changing from a smaller number of more active users, who submit more queries and look at more result pages, in to a larger number of less active users, who instead spend more time on reading the result pages. Whether or not this behaviour is beneficial needs to be qualitatively examined.

References

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, pp. 290-302.

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

Cothey, V. (2002). A longitudinal study of World Wide Web users' information-searching behavior. Journal of the American Society for information Science and Technology, 53 (4), 67-78.

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.

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.

He, D. and Göker, A. (2000). Detecting Session Boundaries from Web User Logs. In Proceedings of 22nd Annual Colloquium on IR Research, Cambridge, UK, pp. 57-66.

Jansen, B. (2005). Seeking and implementing automated assistance during the search process. Information Processing and Management, 41 (4), 909-928.

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., 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.

Sellen, A.J., Murphy, R. and Shaw, K.L. (2002). Web Behavior Patterns: How knowledge workers use the web, in Proceedings of CHI 2002, Minneapolis, USA, 227-234

Schultze, U. (2000). A confessional account of an ethnography about knowledge work. MIS Quarterly, 24 (1), 3-41.

Silverstein C., Henzinger, M., Marais, H. and Moricz, M. (1998). Analysis of a Very Large AltaVista Query Log, Digital SRC Technical Note #1998-014, October 26.

Spink, A. (2003). Web search: emerging patterns. Library trends, 52 (2), 299-306.

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

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). How intranets differ from the web: organisational culture's effect on technology. In Proceedings of ECIS 2005, Regensburg, Germany, 26-28 May 2005.

Stenmark, D. (2005b). One week with a corporate search engine: A time-based analysis of intranet information seeking. In Proceedings of AMCIS 2005, Omaha, Nebraska, August 11-14, 2005, 2306-2316.

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