Enterprise Search Behaviour of Software Engineers

Luanne Freund
Faculty of Information Studies
University of Toronto
Toronto, Canada
luanne.freund@utoronto.ca

Elaine G. Toms
Faculty of Management
Dalhousie University
Halifax, Canada
etoms@dal.ca

ABSTRACT
Technical professionals spend ~25% of their time at work searching for information, and have specialized information needs that are not well-served by generic enterprise search tools. In this study, we investigated how a group of software engineers use a workplace search system. We identify patterns of search behaviour specific to this group and distinct from general web and intranet search patterns, and make design recommendations for search systems that will better serve the needs of this group.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - s*earch process, selection process.*

General Terms
Measurement, Human Factors, Experimentation

Keywords
Enterprise search, software engineers, search behaviour, queries

1. INTRODUCTION
Large scale studies of search engine logs have provided a clear picture of general usage patterns. The average web searcher submits queries slightly over 2 words in length, which is significantly shorter than queries used in earlier types of information systems [5]. Other established patterns of web search behaviour are the use of between 2 to 3 queries per search session, and minimal use of advanced search syntax [6]. Preliminary studies of workplace search suggest that it is qualitatively different from web searching, primarily due to the differing goals of the information community [2]. A recent study of intranet search logs suggests that whereas the pattern of 2-3 queries per session holds true, the average intranet query may be even shorter: 1.4 terms [7]. Subsumed in this data on general behaviour are the search patterns of sub-populations, such as technical professionals using search engines as work tools. In this brief report, we develop a somewhat contrasting picture of how this group searches for and selects documents.

2. METHODOLOGY
The study was conducted as the evaluation phase of a long-term study of the contextual dimensions of workplace search, focusing on a community of software engineering consultants [4]. We implemented an enterprise search engine that crawled and indexed an ~10 GB collection of documents gathered from the Internet, the corporate intranet, and from document repositories heavily used by this group. The crawl was customized for this group based on a targeted set of seed URLS.

The study used a task-based approach to search system evaluation. 32 software engineering consultants working in a large hi-tech company participated, each searching for 4 of a total of 8 simulated work-task scenarios [1] assigned using a Latin square design. The scenarios represent a range of different task types commonly performed by this community, from relatively complex high-level tasks, such as the evaluation of alternative architectural patterns to simple low-level tasks, such as looking up software prerequisites. Sessions were conducted individually and remotely via Windows NetMeeting and lasted approximately 1 hour.

Participants completed four tasks consecutively. Searches took from between 10-20 minutes each. During each search, they were asked to indicate which documents would be useful, to rate them verbally on a simple 1-10 scale for usefulness, and to explain why they thought the selected documents would be useful. Data was collected through the server transaction log, online survey forms, and transcription of verbal responses.

3. RESULTS
On average, this group used relatively long queries (4.38 terms) and identified a small number of useful documents per task (1.7). They issued 2.65 queries on average per task, and visited 3.31 websites. They failed to identify any useful documents in about a quarter (23%) of searches.

Looking more closely at the queries, we found that 66% of all queries contained at least one acronym, and that ~19% of all queries were entirely composed of acronyms. Acronyms were used for product names and various technologies and protocols (i.e. J2EE, SOA, LDAP). Numbers, referring to software version numbers, error codes or years, occurred in 31% of queries. Searchers used phrase syntax in 9% of queries.

Table 1 contains a summary of factors participants used to determine the usefulness of documents. This summary is based on a content analysis of the verbal comments of participants with respect to documents identified as useful. The four broad categories of indicators that emerged from the analysis are: content, format, currency and authority. Participants assessed the content of documents based on the topic, the level of specificity of the information, ranging from broad topical overviews to detailed technical information, and the degree to which the

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

information was generic or situated, i.e. related to a particular case or scenario. For format, they looked for genres suited to the assigned task scenario. They looked for structured documents from which information could be easily extracted, and in some cases preferred documents that allowed for interaction with the author. Currency was assessed with respect to the date of creation and the version of product or technology discussed. Finally, authority was assessed based on acquaintance, reputation or team affiliation of the author and the organizational source of the information (repository or workgroup). These results support and augment previous work on engineers' perceptions of documents [3].

Table 1: Factors used to determine information usefulness

Content	Format	Currency	Authority
Topicality Level of specificity Degree of situatedness	Genre Structure Interactivity	Date Version of product or technology	Author Source repository or group

After completing all four tasks, participants provided feedback on features they would like to see added to this search tool. The most common request was for more advanced search syntax options. Participants emphasized that they regularly use Boolean syntax and that they consider it an essential feature. Other features of interest were the ability to group/limit or sort results based on source, genre, date and software product, and to have authors identified in the hitlist.

4. DISCUSSION

Findings suggest that the search behaviour of this group is quite distinct from general web and intranet search behaviour. They use longer and more detailed queries (see Figure 1 for a comparison). They make heavy use of specialized terms and search syntax, and the features they would like added to current search tools reflect the desire for greater control and functionality.

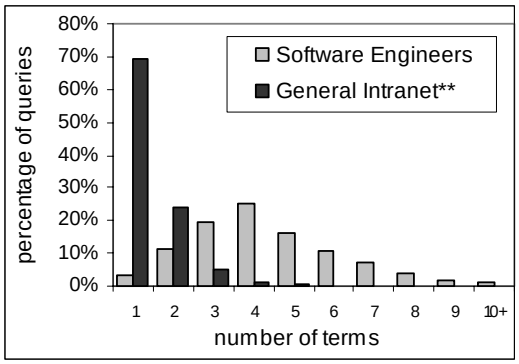

Figure 1: Query length comparison (**intranet data from [7])

The search behaviour of this group likely reflects their domain expertise, which is known to result in longer and more specialized queries [8]. This expertise is also reflected in the non-topical cues they use to assess documents, which rely not only upon technical knowledge, but also on familiarity with their workplace information space. Authorship, genre, and source repository carry strong meaning for them, because so much of the information that is useful to them was created by their own community.

These findings lead to some design implications for workplace search systems for technical professionals. There is a clear need to provide searchers with the tools and support to create complex queries and to make effective use of specialized terminology. Controlled vocabulary look-up lists or query processing tools should be in place to deal with acronyms, product names, and other technical terms. Systems that can extract and highlight non-topical cues (authors, versions, dates, genres) used to determine usefulness will be able to provide benefit. Finally, multi-dimensional means of sorting and limiting the document collection are needed to help searchers find the right information for particular tasks. Automatic methods should be explored to classify collections by genre, level of specificity, degree of situatedness, source repository, etc. These types of specialized, domain-specific features are likely to make workplace search a more effective tool for technical professionals.

5. ACKNOWLEDGMENTS

This research is supported by an IBM CAS Fellowship and a SSHRC and Canada Research Chairs Program grant to the second author. Thanks to all the consultants who took part in the study and to Julie Waterhouse for all her valuable help with this project.

6. REFERENCES

[1] Borlund, P. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research, 8,* 3 (April 2003), paper no. 152.

[2] Fagin, R., Kumar, R., McCurley, K.S., Novak, J., Sivakumar, D., Tomlin, J.A. and Williamson, D.P. Searching the workplace web. In Proceedings of the 12th International World Wide Web Conference *(WWW '03),* (Budapest, Hungary, May 20-24, 2003), 366-375.

[3] Fidel, R. and Green, M. The many faces of accessibility: engineers' perception of information sources. *Information Processing & Management, 40,* (2004), 563-581.

[4] Freund, L., Toms, E.G. and Clarke, C.L.A. Modeling task-genre relationships for IR in the workplace. In Proceedings of the 28th annual international ACM SIGIR conference (SIGIR '05), (Salvador, Brazil, Aug 15-19, 2005), 441-448.

[5] Jansen, B.J. and Pooch, U. A review of web searching studies and a framework for future research. *Journal of the American Society for Information Science, 52,* 3 (2000), 235-246.

[6] Spink, A., Jansen, B.J. and Saracevic, T. Searching the web: the public and their queries. *Journal of the American Society for Information Science, 52,* (2001), 226-234.

[7] Stenmark, D., Searching the intranet: corporate users and their queries. In Proceedings of the 68th Annual Meeting of the American Society for Information Science and Technology (ASIS&T 2005), (Charlotte, NC, Oct.28 -Nov.2, 2005).

[8] Vakkari, P., Pennanen, M. and Serola, S. Changes of search terms and tactics while writing a research proposal: a longitudinal case study. *Information Processing & Management, 39,* (2003), 445-463.