X-Site: A Workplace Search Tool for Software Engineers

Peter C.K. Yeung
School of Computer Science
University of Waterloo
Waterloo, Canada
p2yeung@plg.uwaterloo.ca

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

Charles L.A. Clarke
School of Computer Science
University of Waterloo
Waterloo, Canada
claclark@plg.uwaterloo.ca

ABSTRACT
Professionals in the workplace need high-precision search tools capable of retrieving information that is useful and appropriate to the task at hand. One approach to identifying content, which is not only relevant but also useful, is to make use of the task context of the search. We present X-Site, an enterprise search engine for the software engineering domain that exploits relationships between user's tasks and document genres in the collection to improve retrieval precision.

Categories and Subject Descriptors
H.3. [**Information Storage and Retrieval**]: Systems and Software

Keywords: Enterprise search, contextual search, task, genre

1. SYSTEM OVERVIEW
The X-Site concept is based on a domain analysis of the information practices among a community of software engineers in a major technology firm, which identified a strong relationship between the tasks they perform and the document genres they use. This analysis enabled us to identify task-dependent patterns of genre preference, which we incorporated into the ranking algorithm of X-Site. X-Site includes the following components:

- a **task profile**, composed of a work task (e.g. installation) and an information task (e.g. find facts), which are elicited from the searcher at query time;
- a **task-genre association matrix**, which specifies known positive, neutral and negative relationships between task and genre pairs;
- a **genre weighting** component in Okapi BM25. Document genre is a weighted field that is used in combination with term frequency to score structured documents. Each weight represents the strength of each task-genre pair[2];
- a **genre classifier**, which uses supervised machine learning (SVM^{light} [1]) and textual features to tag the document collection using a domain-specific genre taxonomy;
- a **language identifier**, which uses n-gram-based text categorization (libTextCat[2]); and
- a multi-user **search engine** (wumpus[3]).

[1] http://svmlight.joachims.org
[2] http://software.wise-guys.nl/libtextcat
[3] http://wumpus-search.org

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

X-Site is currently deployed as a prototype in a real workplace environment. It provides a single point of access to ~8GB of content crawled from the Internet, intranet and Lotus Notes data, using a set of seed URLs tailored to the needs of this group.

To search using X-Site, the searcher types in a query, and selects a work task and an information task from drop down lists (Figure 1). The query is used to retrieve a set of results, and the task profile is used to determine the genre weights associated with the search. The results are then ranked using a modified BM25 scoring method, which incorporates genre weights in addition to length, term frequency and other collection statistics.

Figure 1: X-Site results display with query input sidebar

X-Site provides a customized, flexible and user-controlled means of refining results to suit the task context of a search. This is of particular benefit in enterprise information environments, which need to serve diverse user populations and support a wide range of work and information tasks. Future improvements of the X-Site system will include a component to monitor implicit measures of document preference during system use in order to tune the task-genre associations.

This research was supported by the IBM Centre for Advanced Studies in Toronto, Canada.

2. REFERENCES
[1] Freund, L., Toms, E.G. and Clarke, C.L.A., Modeling task-genre relationships for IR in the workplace. in *ACM SIGIR Conference*, (Salvador Brazil, 2005).
[2] Yeung, P.C.K., Büttcher, S., Clarke, C.L.A., and Kolla, M. A Baysian approach for learning document type relevance. in *ECIR*, (Rome, Italy, 2007).