

Context in Enterprise Search and Delivery

David Hawking Cécile Paris, Ross Wilkinson, Mingfang Wu

CSIRO ICT Centre, Australia

Email: {Firstname.Lastname@csiro.au}

1. Introduction

Context plays an important role in information processing process. Context has always been elicited and exploited when librarians helped us to look for information, but current information retrieval systems proceed in blissful ignorance of who we are and what we are trying to achieve. The continuing series of workshops in context attempts to redress the problem. In this paper we look at how context can help when finding information in the enterprise.

This paper aims not to fully describe context, or to fully exploit context, but to address the issue of whether a little bit of context can make a big difference. We ask this question in three phases of retrieval of enterprise information seeking, namely query formulation, matching/ranking and delivery. In each phase, we compare what is current, what is possible, and a simple effective improvement.

2. Context in an Enterprise Search Environment

The presenters of the SIGIR 2004 IRIX Workshop [1] discussed context and how it could be used in information retrieval in general. The types of context identified included users' familiarity with the search topic, search system and search collection; users' interaction history; the time, place and device; and the task in hand.

All these discussions on context could apply to enterprise information search environment. However, as a subset of a broad information search, enterprise search brings its own typical context characteristics: users are mainly the employees, search tasks are usually related to its business, and information sources are dominated by those used or generated in its business operation.

Users – Generally, we are aware of the roles of each user in an enterprise search environment. These users are characterized by their responsibilities, skills, interests and experiences. Their roles also represent their general and long term information needs. For example, a scientist and an IT support person may send the same query “network”, but they may in fact look for different types of information.

Task – Knowing the users' tasks can help to locate the information source and understand the requirements. Information seeking and retrieval is embedded in a task context [5]. This is more significant within enterprise search environment, as an enterprise search tool is usually used for searching job-related information.

Document collection – Enterprise searches are usually conducted in an intranet environment. Documents in this environment are usually unstructured, of various types and from heterogeneous

information repositories, e.g. Email systems, client relationship management systems, content management systems, etc.. Each individual system may provide a local context. However, documents from these heterogeneous repositories typically do not cross reference each other.

In addition to the contents of a document, the properties of a document, such as its author, its creation and modification time, can also provide a useful context in information search and delivery.

What of these context can be captured in enterprise search? What can be exploited effectively? In this paper, we look at “what gives bang for buck”!

3. Context in Enterprise Search System

Query Formulation

Simple queries are overwhelmingly used to formulate users' information needs, with no indication of whether the need is for some information, a fact, a home page, or a service. It has been observed that about 70% of web users typically type in only one keyword or search term [2]. Our own analysis of a query log for our email archive also shows that about 33% queries are one word. This type of simple queries provides little information on what a user wants.

Understanding the user query and building the connection between what a user asks for and what the user wants is a challenge for any search engine. In the extreme we might describe all elements of context to substantially improve query specification. Nordlie showed that knowing why information is wanted made a very big difference between human intermediated searching, compared to OPAC intermediated searching [11]. However, typically, the cost of acquiring full context is simply too high, compared to the benefits, let alone possible privacy issues.

Knowing a user's task may help us to understand what a user wants. In an enterprise there exists a variety of information sources used for various using purposes, a single search across all sources is useful to have, but when we know even a little bit more about the task at hand we can do better. For example, if you want a business document, you might use a standard enterprise search.

If you want to find experts in an area, you might use an expert finding tool that returns a list of experts and their profiles based on the evidences such as personal webpage and publication list, and some elements of the organization's corporate data such as the projects that person participated and the person's role within the organization [9].

Or if you need to find the name of a business contact, it is likely to be buried in a corporate email, then an email search tool is more appropriate here.

Thus, we see in this situation, that we can leverage tasks by specifying the type of search, and hence search engine, at query time, probably leading to substantial improvement of search. This is hard to quantify, given very different desired outcomes, and thus possibly different measures, depending upon context.

Matching and Ranking

The purpose of matching and ranking is to identify, based on the search query, those documents that are most relevant, so that the human user can look at the most promising documents first. However, the relevance may be different for web search and enterprise search. On the internet, there are a large number of documents that are typically relevant to query, a user is often looking for the “best” or most relevant documents. On an intranet, the definition of “best” answer may be less clear cut. There may be no authoritative website dedicated to the topic of their query. On the other hand, users might more often know or have previously seen the specific document(s) that they are looking for. Intranets may have a small set of “correct answers” (often a single page) for any given query [10]. So the matching and ranking algorithms work for the web may not have the same effect for the enterprise search.

Matching algorithms typically determine the overlap of the content in the query with the content of the document. On web data, matching algorithms have been improved to take into account factors such as anchor text matching, link graph characteristics, click popularity and so on [6]. However, it is a definite limitation that current search engines return the same ranking to all users despite major differences of preference and purpose. Factoring in user context seems to offer the greatest potential for the next big step forward in ranking quality.

It might be possible to do much better with full context, but this is not generally available, and might be too complex for the search engine to productively exploit. Apart from the user’s interaction history with system [7], what is generally available in enterprise search is the role of the users, as the user is known within an organization, and has a specific role and position. Fagin et al. found that the correct answer to a query is often specific to a site, geographic location, or an organizational division [4]. This information could be helpful for ranking search results from heterogeneous collections. For example, a matched document from client relationship management repository would be ranked higher than that from supplier profile repository for a business manager, while it would be opposite for a service staff.

During the matching and ranking, it is also essential to identify the context within document content. As the enterprise information sources consist of a number of disconnected repositories such as emails, client relationship management system, content / knowledge management system, and so on, the search engine needs to have some knowledge about these applications and their associated document repository, so the context can be used to retrieve key and right information. For example, in an email system, the information contained in the To, CC, and Subject field of a message is probably more important than the information in the message body or attachment. This

information could also be used to relate to the account name and customer problem in retrieving the right information from a customer relationship management system.

Two questions of very considerable interest in the matching and ranking process are how to represent and communicate the elements of user context which will actually make a difference to ranking and how to process the contextualised query. One possible method is *query augmentation* – the addition of extra words or scoping elements to the query to represent the context. Another method modifies the profile of static (query-independent) a priori document probabilities according to context such as role or task. Static probabilities are relied on in modern search engines and are often determined by click or link popularity. In an enterprise one could imagine them being set on the basis of document genre (e.g. policy, media release, technical report, customer email etc.), recency, source etc. One could equally imagine modifying the bias profile according to whether the searcher was a manager, a salesperson, a researcher or an assembly line worker.

We aim to devise a protocol for communicating context to an enterprise search engine which will be simple enough for efficient transmission and general enough to apply to at least 80% of enterprises. Implementation of a practically useful system will also require a generalised method for extracting the relevant context: either requiring the user to specify their own profile or interfacing with the directory service for the organisation. A practical system should also allow for generic search if the user requests it and should allow a user to easily take on more than one role or task profile.

Obviously there is no standard test collection for contextualised enterprise search. We propose using volunteer employees performing their normal everyday task to evaluate the contribution of context in the matching/ranking phase by presenting results from the enterprise’s standard search tool in two side-by-side panes. One pane shows the standard search, the other shows the context-aware version. The two variants are randomly assigned to left or right and the volunteers are asked to rate the relative value of the two panes as follows: prefer left pane, prefer right pane, or both panes of equal value.

Delivery

Display schemes attempt to present a summary of each matching document for the user’s evaluation in such a way that the user can identify the best document at a glance. Most search engines deliver a list of document surrogates with links to the relevant pages. However this is often of little use – suppose my search is over the phone to my corporate search engine, or I wish to take the results of my search from a network printer on the way to a meeting. In the first case, a page of text is way too much, and in the later case, hotlinks are useless. By knowing the delivery device only it is possible to do much better. The two images in Figure 1 show information delivered to two different devices – all as a result of the same information need, but where knowledge of the delivery device makes all the difference.

Again, determining the value of improved delivery is difficult using standard evaluation measures. It is sometimes the case that the value is absolute – delivering in one mode may be completely impractical. In other circumstances, the value is more subtle. The

above delivery is created using rhetorical structure theory [8], attempting to ensure that the information delivered is more easily assimilated than would be the case if the form of delivery was not considered. Possible measures of a delivery form include cognitive load measures and comprehension measures, however, again, the measures depend upon search and use context. Currently we simply know that users prefer delivery that has good rhetorical structure [12].

4. Conclusion

We have seen that modest use of context in enterprise search and delivery can make a significant difference – by knowing what is being searched for, we can invoke different search – this can be achieved at the interface. By knowing who is searching enables us to leverage their role in working out classes of documents that may a priori be more useful. By knowing where the person is, we can deliver information appropriately – paper, desktop, or PDA.

In the enterprise, each of these fragments of context may easily be available, yet each can deliver substantial value.

Consequently we argue for the importance of context, but against the need to fully capture context, as there are practical steps to exploit partial context effectively.

5. References

[1] <http://ir.dcs.gla.ac.uk/context/presentations/>

[2] D. Butler. Souped-up search engines. *Nature*, Vol.405, pp.112-115, May 2000

[3] N. Craswell, D. Hawking, A. Vercoustre, and P. Wilkins. P@noptic expert: searching for experts not just for documents. In *Poster Proceedings of AusWeb'-1*, 2001

[4] R. Fagin, R. Kuman, K. S. McCurley, J. Novak, D. Sivakumar, J. A. tomlin, and D. P. Williamson. Searching the workplace web. In *Proceedings of WWW2003*, Budapest, Hungary, May 2003.

[5] K. Järvelin and P. Ingwersen. Information seeking research needs extension towards tasks and technology. *Information Research*, 10(1) paper212 [Available at <http://inforationR.net/ir/10-1/paper212.html>]

[6] D. Hawking. Challenges in enterprise search. In *Proceedings of the Australasian Database Conference*, Dunedin, New Zealand, pp.15-24, Jan. 2004.

[7] S. Lawrence. Context in Web Search. *IEE Data Engineering Bulletin*, vol23, No.3, pp25-32, 2000

[8] W.C Mann, S. A. Thompson. Rhetorical structure theory: Towards a functional theory of text organization. In *TEXT*, 8(3), pp 243- 281, 1988

[9] A. McLean, A. Vercoustre and M. Wu. Enterprise PeopleFinder: Combining Evidences from Web Pages and Corporate Data. In *Proceedings of the 8th Australasian Document Computing Symposium*, Canberra, Australia, Dec. 2003.

[10] R. Mukherjee and J. Mao. Enterprise search: tough stuff. *Enterprise Search*, vol.2, No. 2, April 2004

[11] R. Nordlie. “User revelation” – a comparison of initial queries and ensuing question development in online searching and in human reference interactions. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*. pp.11-18, 1999

[12] C. Paris, S. Wan, R. Wilkinson, and M. Wu. Generating Personal Travel Guides - and who wants them?. In *Proceedings of the International Conference on User Modelling (UM2001)*; Sonthofen, Germany, July 13-18, 2001



Figure 1: An example of two delivery modes