Automatically Adapting the Context of an Intranet Query

Deirdre Lungley
Dept. of Computing and Electronic Systems
University of Essex
Wivenhoe Park, Colchester, CO4 3SQ, UK
dmlung@essex.ac.uk

Abstract

In this paper we propose the use of an interactive interface to allow user exploration of the context of an intranet query. The underlying domain model is that of a Formal Concept Analysis (FCA) lattice. Understanding the difficulty of achieving optimum document descriptors, essential for a browsable lattice, we propose harnessing implicit user feedback in learning document/term associations.

Keywords: information retrieval, implicit user feedback, query refinement, context enhancement, formal concept analysis, collaborative indexing

1. INTRODUCTION

In an age when companies and institutions are becoming ever more reliant on their intranets, when successful management of their information base is key to efficiency and competitive advantage, the tools they use to access this information are being seen as ever more important. It is reasonable to assume that intranet users are generally very familiar with the traditional search interface, they enter their query term(s) in the text box and scan the resulting document snippets. Many are aware of its fundamental "bag-of-words" model and have a well-defined search need, which they represent carefully with chosen keywords. The power of modern search technology is such that their need is met in a high proportion of cases. We, in this research, wish to argue the advantages of improving this interface, of going beyond this standard list of results, of providing an interface which illustrates the contexts within which their query term can be found. This enhanced interface is targeted in particular at multi-context and ambiguous queries. Here an interface which guides the user in his choice of terms can reduce the user effort required.

Context enhancement is not new to Information Retrieval (IR). As far back as the early 1970s the use of clustering was promoted as a form of context enhancement (Jardine & van Rijsbergen, 1971). Since then there have been many examples within IR of a query context model being used to enhance the user experience (Sanderson & Croft, 1999, Carpineto & Romano, 2004). In each case researchers have recognised the advantages of offering an enhanced search interface, particularly to less experienced users and users with an information need which is not clearly defined. These enhanced interfaces, illustrating the context of a query, can guide the user in selecting the keywords in use within the collection.

The question then arises as to how accurately the document contexts are reflected in these automatically generated models? These systems attempt to fully automate the context modelling process - a non-trivial task. An alternative would be to use a human defined ontology. We and the above researchers have avoided this option due to its setup and ongoing maintenance costs (Maedche et al., 2003). However, the success of the commercial systems, Aquabrowser¹ and Clusty², and evaluation studies of the research systems (Sanderson & Croft, 1999, Carpineto & Romano, 2004) show the potential of query enhancement. Therefore, we set out with this

¹<http://aqua.obbeindhoven.nl/>

²<http://clusty.com>

research to build on this success, but harness machine learning techniques to achieve further query enhancement - to tune the context model of a user query.

Our research is to be contained within the field of intranet search. We would argue that there is a particular case, both for providing context enhancement and for harnessing user population feedback, in an intranet setting. Since:

- Intranets are relatively controlled environments, consisting of a relatively confined document collection, often with imposed annotation standards and are also far more unlikely to suffer from spamming, making indexing components such as inlinks and metadata more reliable (Fagin et al., 2003). These properties help to ensure the success of the non-trivial task of creating and tuning a query specific context model.
- The user population of an intranet search system is a particular community of users, e.g., the staff and students of a university or the management and employees of a company. The cohesive nature of these communities and their search needs aids the viability of harnessing user population feedback.

A motivating factor in our research has been our analysis of query log data recorded on the University of Essex intranet search engine from the past year. One of the more striking findings of our analysis is the brevity of user search queries. This is in-line with published findings (Jansen et al., 2000). An analysis of the results returned by even such brief queries, shows that the information need of the user, in as far as can be ascertained by the query term, appears to be generally met within the top five results, e.g. graduation, library and timetable. This, however, is not the case when the term is more general, e.g. parking, sport or printing. These more general terms and ambiguous terms like "union" highlight the potential for context enhancement:

- Printing - Within the university the term printing can have several contexts, e.g. printing centre, printing credit or laboratory printing facilities
- Sport - Again the many contexts of this term could include, sports clubs for children, information on facilities at the Sports Centre or sport science courses. There can be a temporal dimension to a particular context, e.g. in January many parents book summer sports clubs for their children
- Parking - This particular term is often amended to "parking permit", a query where the seemingly most relevant document is accessed indirectly through a link in a returned document

The aim of our research is to meet the shortcomings of present search facilities for these types of queries: to provide context information, to adapt this context to temporal requirements and to override inadequate document annotation, by learning a document's context where necessary. This is to be achieved by learning the optimum association between a document and its terms from past user behaviour.

2. RESEARCH QUESTION

We are conducting research designed to answer the following question:

1. Can the context model, the domain specific concept lattice, of an intranet search query be effectively adapted by user implicit feedback?

3. PROPOSED RESEARCH METHOD

Many researchers setting out to improve intranet search through context enhancement have followed a common methodology. They have fed the top 100 to 500 documents returned by an underlying search engine into a domain modelling tool (Carpineto & Romano, 2004, Sanderson & Croft, 1999). We aim to follow this common approach, therefore we are adopting the architecture outlined in Figure 1. This illustrates how documents returned by the underlying search engine, in this case Lucene's Nutch³, undergo some natural language processing (NLP) in order to extract

³<http://lucene.apache.org/nutch/>

Figure 1: System Architecture

initial concept terms. These initial concepts are then merged with the concepts derived from our adaptive element, before being processed by Formal Concept Analysis (FCA) algorithms to produce a browsable concept lattice.

The attraction of the interconnected lattice structure over the more traditional hierarchical tree structure persuaded us to explore FCA as the underlying model of our query context. Toscanaj⁴ an open source FCA tool provides us with the algorithms to generate lattice context models from the search engine results. We are not proposing to visualise the full dimensionality of the lattice, but by showing the intent and extent of a node, we are displaying the various interconnections between the derived concepts. The focus of our research is not concerned with FCA. We purely wish to exploit the navigational advantages of the concept lattice. However, a brief description of the terminology of FCA is necessary and is provided in Section 4.

IR researchers interested in exploiting the multi-dimensional structure of the lattice have discovered challenges in using this technique for domain modelling (Cigarrán et al., 2005):

1. Similarly to LSI (Latent Semantic Indexing) FCA is computationally too expensive for large scale processing
2. Unless great care is taken in the selection of index terms, the resulting lattice can be too big and complex for practical browsing
3. Effective visualisation of the lattice is required to harness the potential of this structure

Our research is taking certain steps to surmount these challenges. Firstly we are not setting out to model the entire document collection, but a selection or 100 or so documents returned as results from an underlying search engine. The NL processing we are doing is our answer to the second challenge. This processing must optimise the initial document descriptors, to create a well interconnected lattice, suitable for browsing. At this early stage of our research we are using relatively unsophisticated NLP to extract initial document derived concepts. We are applying a parts-of-speech tagger to each document snippet and using simple noun phrase patterns to locate noun phrases, e.g., noun-preposition-noun. As our research progresses more sophisticated processing might be found beneficial. With regard to lattice visualisation, we are not intending to

⁴<http://toscanaj.sourceforge.net/>

Figure 2: Initial extent of unadapted lattice following the query "parking"

display the complete underlying lattice, but instead to display the highlighted concept, its intent and the concepts within its extent. Figure 2 is an example of our interface.

The core of our research is to automatically improve the context model of a query by harnessing implicit user population feedback (depicted as "Machine Learning Module" in Figure 1). This will be achieved by learning the document descriptors, the document/term associations. Our initial machine learning processing will utilise Joachims' SVM-Light⁵. Its ranking SVM has been used effectively to re-rank search results (Radlinski & Joachims, 2005). We will explore its potential for tuning our lattice - learning the key documents for a query and the key concepts represented by those documents. Our logged clickthrough data will provide the training data required. Periodic processing of this data will update a model for classification. Figure 2 shows the initial extent of the lattice. The non-adapted lattice appears to give prominence to some terms which appear to bear little relevance to the "parking" query. Through learning the underlying lattice for this query should become more compact, retaining only relevant documents and relevant terms within those documents.

Our interactive environment allows for continual adaptation. Query terms entered, not derived from the underlying documents, allow new user-driven terms to be added to the context. Logging URL clicks beyond the result documents allows for user-driven documents to be added to the context. Temporal aspects of our context e.g., the autumn teaching timetable is most likely to be the required document for "timetable" queries at the beginning of October, can be learnt and can subsequently be overridden. It could be argued that learning such temporal features without introducing incorrect associations is a challenging proposal and may benefit from manual intervention. However, our main aim of full automation may mean we have to tolerate a degree of noise. Users, it appears, can tolerate a degree of noise (White & Ruthven, 2006). The selection of a suitable machine learning threshold will be imperative here.

⁵<http://svmlight.joachims.org/>

Figure 3: Cross-table and related concept lattice

This research aims to evaluate the feasibility of users interactively adapting the context model, the concept lattice, of an intranet query and our final evaluation would therefore involve full interactive IR evaluation. However, our initial explorations of the feasibility of our methodology will involve utilising data from the University of Essex intranet logs. These logs contain the initial query entered and subsequent modifications. We will choose frequently modified multi-context terms e.g., parking, printing and sport, and enter the initial query and subsequent modifications. Our intranet logs do not record the clicked URLs and so our initial investigations will have to make a subjective decision when choosing to click on a URL. The result of these initial investigations will give us some early indications of how effective this combination of Lucene's Nutch search engine, simple NL processing, an FCA domain model and SVM-Light's ranking algorithm are in adapting query contexts.

4. FORMAL CONCEPT ANALYSIS (FCA)

The mathematical foundations of FCA have been detailed in Ganter & Wille (1999). The following brief synopsis highlights those aspects which make FCA an attractive option for our domain model.

4.1. Contexts and Concepts

In FCA, the extension is the set of objects(entities) belonging to the concept, while the intension is the set of attributes (properties) of these objects. Therefore the basic structure of FCA is the **context** (G, M, I) , where I is a binary relation between the set of **objects** G (for the German word Gegenstände) and the set of **attributes** M (for the German word Merkmale), $I \subseteq G \times M$. This context can be illustrated by a cross-table as in Figure 3.

We can define A as a subset of our object set G and A' as the set of common attributes of A . Similarly, we can define B as a subset of our attribute set M and B' as the set of common objects of B . Then the pair (A, B) forms a **concept** of the context (G, M, I) if $A' = B$ and $B' = A$. In other words, in a concept (A, B) the set of objects that the members of B have in common is A and the set of attributes that the members of A have in common is B . In a concept (A, B) , A (the set of objects) is called the concept's **extent** and B (the set of attributes) is called the concept's **intent**. For example, $(\{exam_aut.xls, exam_sum.xls\}, \{exam\})$ is a concept of our context, i.e., $exam_aut.xls$, $exam_sum.xls$ are the members of the extent; $exam$ is the one member of the intent.

4.2. Concept Lattice

We can define a binary relation $\leq$ - "is a subconcept of". If (A_1, B_1) and (A_2, B_2) are concepts of our context, then $(A_1, B_1) \leq (A_2, B_2)$, i.e., (A_1, B_1) is a **subconcept** of (A_2, B_2) if A_1 is a subset of A_2 and B_1 is a subset of B_2 . (A_2, B_2) is then also called a **superconcept** of the concept (A_1, B_1) .

By convention, in the line diagram of a concept lattice, the object names are written below the circle representing the object concept - the object concept being the smallest concept having the object in its extent. The attribute names are written above the circle representing the attribute concept - the largest concept having the attribute in its intent.

5. RELATED WORK

Content enhancement e.g. document clustering or classification, has been explored through the decades of IR research in an attempt to organise the results returned by a search engine and to aid the navigation of the user through the information space (Jardine & van Rijsbergen, 1971). It can help the user to clarify the context of his query. The technique of clustering documents into a hierarchy of concept clusters has proved an effective Information Retrieval mechanism as can be seen by the success of the commercial search engine Clusty⁶, developed by Vivisimo Ltd. However it is not without its problems. Firstly, it relies on an accurate method of determining the concept for a cluster of documents. Secondly, its hierarchical structure does not inherently allow a user to navigate easily between clusters. IR classification generally involves the use of a term hierarchy and much research is currently being done on their use in ordering documents. Some are investigating the use of large external taxonomies such as the Open Directory Project (ODP)⁷. Although the advantage of communities committing to specific ontologies has its appeal, drawbacks also exist. Ontologies are difficult to maintain (Maedche et al., 2003) and also enforce a specific vocabulary on the user community which has not evolved naturally.

Most IR systems require the user to express their information need in the form of concise query terms and therefore a knowledge of the domain vocabulary is useful (Furnas et al., 1987). Query refinement, another prominent aspect of IR research, can address this issue (Rocchio, 1971).

Formal Concept Analysis (FCA) has been an attempt at combining context enhancement and query refinement: retrieval results organised in a browsable concept lattice, prompt the user in the selection of terms in use within the document collection (Carpineto & Romano, 2004). Over recent years there has been a growing interest among information scientists in FCA and its merits for supporting search have been highlighted by other groups (Cole et al., 2003, Ferré & Ridoux, 2000). Although all these works show the non-trivial nature of using a lattice structure as the basis of a domain model, we are attracted by its navigational capabilities and the prospect of surmounting its difficulties, particularly the choice of index terms, through the use of implicit user feedback.

The use of learning from both implicit and explicit feedback is not new in information retrieval. Relevance feedback has been promoted as an effective method of improving retrieval performance (Rocchio, 1971). Recently, there has been a renewed surge of interest in harnessing the potential of user feedback in an attempt to capture the true information need of the user. Implicit user feedback i.e., user actions such as accessing or printing a document, is seen as an answer to the enormous drawback of explicit feedback, the reluctance of the typical user to rate a document (Jansen et al., 2000). This current interest in using clickthrough data to mine preferences is not without its dissenters. It has been proposed that clickthrough data is not necessarily a good indicator of relevance and one needs to be careful when interpreting this data (Scholer et al., 2008). An answer to this has been to use clicks not as indicators of absolute relevance but of relative relevance, e.g. if a document is clicked on in response to a query, that document is deemed more relevant to that particular query than those listed above it, which have been ignored and the one below it, which has also been ignored (Radlinski & Joachims, 2005). They have had considerable success in using this method to re-rank result documents. Our initial explorations in implicit feedback will also use this concept of relative relevance, not for ranking, but to tune the concept lattice. We are, however, aware of its limitations. Poor content snippets mean that users could repeatedly click on a non-relevant item, so reinforcing invalid relevance and introducing noise. One solution to this is the use of multiple indicators of relevance (Melucci

⁶<http://clusty.com>

⁷www.dmoz.org

& White, 2007). As our research develops it will be necessary to further explore and employ such discriminating methods.

While some have shown the potential of user population feedback in re-ranking search results (Radlinski & Joachims, 2005, Smyth et al., 2004), others use this technique to improve document retrieval through automatic query expansion (Cui et al., 2002). The language modelling IR research community are also showing considerable interest in feedback. One group derive their query language model by combining the current query terms, related query term histories and related clicked document histories and use it to better the ranking of documents (Shen et al., 2005). Another example of the use of implicit user feedback is where query recommendations are made based on the previous query reformulation behaviour of users (Jones et al., 2006).

Using implicit user feedback to aid context enhancement has also been explored. Web search logs have been mined in an attempt to surmount the classic clustering problems: clusters discovered do not necessarily correspond to the interesting aspects of a topic from the user's perspective and cluster labels generated are not informative enough to allow a user to identify the right cluster (Wang & Zhai, 2007). In answer to these problems they learn these aspects from Web search logs and organise search results accordingly. Search queries and their clicked results have also been used to provide valuable feedback about the relevance of documents to queries (Poblete & Baeza-Yates, 2008). They choose the document's features from the terms of the queries from which it was clicked from and so reduce by over 90% the set of features needed to represent a set of documents. They propose this document representation model for clustering and classification. This use of implicit feedback in aiding the choice of document features is similar to our research, however it aims at reorganising the results rather than adapting a domain model to assist users in interactive search.

The question may be asked as to whether users actually want any refinement/context navigation? Previous task-based evaluations on the University of Essex intranet suggest that users do in fact prefer system-assisted search to a standard search engine (Kruschwitz & Al-Bakour, 2005). It is also interesting to note that mainstream search engines are moving in this direction: Google is making context dependent suggestions, Microsoft Live search is offering related searches for ambiguous queries such as "java" and Yahoo! is offering a "Search Assist Settings" tab.

6. ACKNOWLEDGEMENTS

I would like to thank my supervisor Udo Kruschwitz for his help and advice on drafting this paper and two anonymous reviewers for very helpful feedback on an earlier version of this document.

References

- Carpineto, C. & Romano, G. (2004), 'Exploiting the potential of concept lattices for information retrieval with credo', *Journal of Universal Computer Science* **10**(8), 985–1013.
- Cigarrán, J., Peñas, A., Gonzalo, J. & Verdejo, F. (2005), Automatic selection of noun phrases as document descriptors in an fca-based information retrieval system, *in* 'Formal Concepts Analysis. Third International Conference, ICFCA 2005', Springer.
- Cole, R., Eklund, P. & Stumme, G. (2003), 'Document retrieval for e-mail search and discovery using formal concept analysis', *Applied Artificial Intelligence* **17**(3), 257–280.
- Cui, H., Wen, J., Nie, J. & Ma, W. (2002), Probabilistic query expansion using query logs, *in* 'WWW '02: Proceedings of the 11th international conference on World Wide Web', ACM Press, New York, NY, USA, pp. 325–332.
- Fagin, R., Kumar, R., McCurley, K., Novak, J., Sivakumar, D., Tomlin, J. & Williamson, D. (2003), Searching the workplace web, *in* 'Proceedings of the Twelfth International World Wide Web Conference', pp. 366–375.
- Ferré, S. & Ridoux, O. (2000), A file system based on concept analysis., *in* 'Proceedings of the 1st International Conference on Computational Logic', pp. 1033–1047.

- Furnas, G., Landauer, T., Gomez, L. & Dumais, S. (1987), 'The vocabulary problem in human-system communication.', *Comm. of the ACM* **30**(11), 964–971.
- Ganter, B. & Wille, R. (1999), *Formal Concept Analysis. Mathematical Foundations.*, Berlin: Springer.
- Jansen, B., Spink, A. & 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**(2), 207–227.
- Jardine, N. & van Rijsbergen, C. J. (1971), 'The use of hierarchic clustering in information retrieval', *Information Storage and Retrieval* **7**, 217–240.
- Jones, R., Rey, B., Madani, O. & Greiner, W. (2006), Generating query substitutions, in 'Proceedings of the 15th international conference on World Wide Web', pp. 387–396.
- Kruschwitz, U. & Al-Bakour, H. (2005), 'Users want more sophisticated search assistants - results of a task-based evaluation.', *Journal of the American Society for Information Science and Technology (JASIST)* **56**(13), 1377–1393.
- Maedche, A., Motik, B., Stojanovic, L., Studer, R. & Volz, R. (2003), An infrastructure for searching, reusing and evolving distributed ontologies., in 'Proceedings of the Twelfth International World Wide Web Conference', pp. 439–448.
- Melucci, M. & White, R. (2007), Utilizing a geometry of context for enhanced implicit feedback, in 'Proceedings of the Conference on Information and Knowledge Management', pp. 273–282.
- Poblete, B. & Baeza-Yates, R. (2008), Query-sets: Using implicit feedback and query patterns to organize web documents, in 'Proceeding of the 17th international conference on World Wide Web', pp. 41–50.
- Radlinski, F. & Joachims, T. (2005), Query chains: learning to rank from implicit feedback, in 'Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, Chicago, Illinois, USA', pp. 239–248.
- Rocchio, J. (1971), *The SMART Retrieval System: Experiments in Automatic Indexing.*, Prentice Hall, Englewood Cliffs, NJ., chapter Relevance Feedback in information retrieval.
- Sanderson, M. & Croft, B. (1999), Deriving concept hierarchies from text, in 'SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval', ACM Press, New York, NY, USA, pp. 206–213.
- Scholer, F., Shokouhi, M., Billerbeck, B. & Turpin, A. (2008), Using clicks as implicit judgements: Expectations versus observations., in 'Proceedings of the 30th European Conference on Information Retrieval (ECIR'08)', Glasgow, pp. 52–64.
- Shen, X., Tan, B. & Zhai, C. (2005), Context-sensitive information retrieval using implicit feedback, in 'Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval', pp. 43–50.
- Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M. & Boydell, O. (2004), 'Exploiting Query Repetition and Regularity in an Adaptive Community-Based Web Search Engine.', *User Modeling and User-Adapted Interaction* **14**(5), 383–423.
- Wang, X. & Zhai, C. (2007), Learn from web search logs to organize search results, in 'Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval', pp. 87–94.
- White, R. W. & Ruthven, I. (2006), 'A study of interface support mechanisms for interactive information retrieval', *Journal of the American Society for Information Science and Technology (JASIST)* **57**(7), 933–948.