Full citation: Limbu, D.K., Pears, R., Connor, A.M., & MacDonell, S.G. (2006) Contextual and concept-based interactive query expansion, in Proceedings of the 19th Annual Conference of the National Advisory Committee on Computing Qualifications (NACCQ). Wellington, New Zealand, NACCQ, pp.151-155.

Contextual and Concept-Based Interactive Query Expansion

Dilip Kumar Limbu, Russel Pears, Andy Connor, and Stephen MacDonell

SERL, Auckland University of Technology
Private Bag 92006, Auckland 1142, New Zealand
{dilip.limbu, russel.pears, andrew.connor, stephen.macdonell}@aut.ac.nz

Abstract

In this paper, we present a novel approach for contextual and concept based query formulation in web-based information retrieval, which is an on-going PhD project being undertaken at the Software Engineering Research Lab (SERL) at Auckland University of Technology (AUT). Various query formulation approaches have been studied for a long time with varying degree of success. To the best of our knowledge none of the existing approaches offer a similar service to the one discussed in this paper.

Our novel approach centres on the formulation of a high quality search query using a user's contextual profile, a shared contextual knowledge based, lexical databases and domain-specific concepts. A user's contextual profile is constructed by monitoring and capturing user's implicit and explicit data. A shared contextual knowledge based is built by consolidating various users' contextual profiles. A machine learning technique is employed to learn user's specific information needs and support the iterative development of a search query by suggesting alternative terms/concepts for query formulation. Early results indicate that the system has the potential to not only aid in the formulation of high quality search queries but also contribute towards the long term goal of intelligent contextual information retrieval from the WWW.

Keywords: Contextual retrieval, Contextual profile, Query formulation, Concept-based query formulation, Knowledge acquisition.

1. INTRODUCTION

Search engines are the most commonly used resources on the Internet to find relevant information. Despite the recent advances in these search engines features *such as related searches, clustering, find similar, search within, search by language, sort by date, advanced search pages, help pages and so on*, due to the exponential growth of information on the Internet has introduced new challenges for finding relevant information on the Internet. The introduction of approaches that increase the ease of finding information has the potential to improve the web experiences of individuals, businesses and educational institutions. Businesses who embrace new search technologies will reap the rewards of improved competitiveness through faster and more efficient searches and retrieval of information. The approach outlined in this paper, whilst primarily aimed at the public web, can be equally applied to corporate intranets and when used in conjunction with alternative approaches can be used to search semi-structured and unstructured data sources on a corporate network.

Today's search engines results largely depend on the user specified/formulated search query (Taksa, 2005). According to Fonseca et al. (2005), in general consensus, search engine users frequently specify short queries with little or no context information associated with these queries. In addition, according to Challam (2004), today's search engines results are based on simple keyword matches without any concern for the information needs of the user at a particular instance in time *(or in a particular context)*. For example, if a user submits a keyword *(e.g. "surfing")* to search for information from the WWW, the search engine searches through the indexed Web pages, filters and returns a list of those documents that contain the specified keyword *(i.e. surfing)*. However, the keyword "surfing", could have completely different meanings – such as *Internet surfing*, *beach surfing*, *surfing lesson*, *surfing shop* and so on – depending on the context it is used in. The user can include additional search terms that could help to refine the search queries, but it is difficult for even experienced users to select the optimum query terms so that the desired subset of information is retrieved (Leake & Scherle, 2001). As a result, even the most experienced users find it difficult to find relevant information from the WWW (O'Hanlon, 1999).

In this regard today's search engines are lacking a personalization mechanism as well as the capability to 'understand' the search query in terms of the information needs of a user at a particular instance in time, thus enabling them to return customized results (Challam, 2004). The combination of these two factors short queries and keyword based search results present great research challenges related to the query formulation process. Query formulation has been suggested as an effective way to resolve the short query and word mismatching problem (Cui, Wen, Nie, & Ma, 2002).

In this paper, we present an alternative query formulation approach using a user's contextual profile, a shared contextual knowledge based, lexical databases and domain-specific concepts. The remainder of this paper is organised in the following way. Section 2 presents related work. Section 3 describes the approach overview and design - contextual and concept-based query formulation. Section 4 offers some concluding remarks as well as future research directions.

2. RELATED WORK

The need to better target a search on the information that will satisfy a user's information needs is well recognised (Leake & Scherle, 2001). Various query formulation approaches have been studied for a long time to satisfy a user's information needs with debatable success in many instances.

Liu et al.'s (2005) approach uses WordNet to improve retrieval process by adding new terms and phrases to the original query and assigning an additional weight to a feedback term that are related to disambiguated query terms. Fonseca et al.'s (2005) approach uses extracted concepts from a special type of query relation graph to expand the original query. Kraft et al.'s Y!Q (2005) system uses a semantic network for analysing search context and generates a contextual digest comprising its key concepts. Using the digest, the query planner augments a user's search query with relevant context terms to improve the overall search relevancy and experience. Sieg et al.'s ARCH (2004) system uses domain specific concept hierarchies to assist users in formulating an effective search query. The system's query enhancement uses two mutually supporting techniques: semantic and behavioural. The behavioural aspect requires observing the users *"browsing behaviour"* for user profiling and automatic query enhancement, while the semantic aspect supports the use a concept hierarchy for interactive query enhancement.

Billerbeck et al.'s (2003) approach expands query by obtaining expansion terms, based on selecting terms from past user queries that are associated with documents in the collection. Klink et al.'s RUBIC (2002) system uses a user's preferences and search keywords to identify phrases *(i.e. one or more words)*. The system expands each phrase with a concept using existing stored concepts and results as new query are sent is presented to the user to confirm the reformulation. The reformulated queries are sent to the search engines and the user is presented with a hit list with relevant documents. Cui et al.'s (2002) approach expands the query using query logs. This approach extracts probabilistic correlations between query terms and document terms by analysing query logs.

All theses state-of-art query formulation approaches expand the original search query by adding additional new/related terms to it. Various approaches such as Contextual profile based *(i.e. user's behaviour and their explicit preferences)*, Ontology (Gruber, 1993) or Concept-based *(e.g. domain specific "ontology")*, domain knowledge based *(e.g. lexical databases, Web based classification hierarchies, etc.)*, Query logs *(e.g. search query logs, user logs, etc.)* and Collaborative Filtering (Herlocker, Konstan, Borchers, & Riedl, 1999) are being used to extract those additional new/related terms.

Despite the success achieved by these approaches, to the best of our knowledge none of them offer a similar service to the one discussed in this paper. In contrast, our system assists users in the creation of an effective search query using a user's contextual profile, a shared contextual knowledge based, lexical references and domain-specific concepts prior to the initial search task. Our approach is novel in the following aspects. First, it proposes the use of a shareable contextual knowledge based. Second, it introduces a mapping technique that maps users search term with the domain-specific concepts using user's contextual profile and lexical databases. Third, it introduces the query formulation process, which uses a user's contextual profile, a shared contextual knowledge based, lexical databases and domain-specific concepts.

3. APPROACH OVERVIEW AND DESIGN

The summarized functionality of the contextual and concept based query formulation approach is depicted in Figure 1. The approach consists of two main components: the Contextual profile (CP) component and the Knowledge base query formulation (KbQF) component. In this paper, we restrict our discussion of the system functionality to concept-based query formulation.

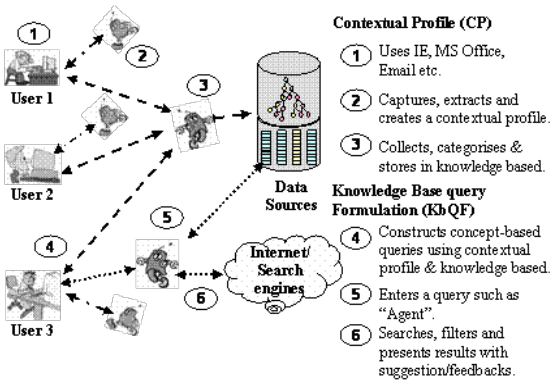

Figure 1: The Approach Functionality

The approach centres on the formulation of high quality concept-based search queries using a user's contextual profile, a shared contextual knowledge based, lexical databases and domain-specific concepts. In the absence of a direct source for a knowledge based, users' contextual profiles are consolidated to form this shared resource.

A user's contextual profile is constructed by monitoring and capturing user's implicit *(i.e. behaviour)* and explicit *(i.e. preferences)* data. A combination of non-intrusive and intrusive approaches is employed to gather implicit and explicit data from a user. Once the contextual profile is built for a user, the search query is evaluated against the contextual profile and a mapping to domain-specific concepts in the shared contextual knowledge based is obtained. The end result is a query enriched with additional search terms which can then be submitted to a search engine such as Google. The following paragraphs discuss this process in more detail.

3.1 Contextual Profile Component

The Contextual profile (CP) component performs Knowledge acquisition (KA) and Knowledge reasoning (KR) processes. Figure 2 provides a summarized representation of the KA and KR processes as integrated within the umbrella CP component.

The KA process starts with the building of a user's contextual profile by monitoring and capturing his/her desktop activities. The content *(or contextual profile)* captured during this process can be implicit data (*i.e. user's behaviour*) as well as explicit data *(i.e. their stated preferences)* and these data represent a user's interests. A typical user exhibits many patterns when interacting with a computer system and patterns vary from user to user. Implicit user behaviour, such as subject content and frequency of access to web pages browsed is used to complement their explicit preferences in building the user contextual profile.

The KA process utilizes the nearest neighbour technique, to group similar interaction patterns together into classes which are stored in a personal knowledge based *(i.e. the user profile)* and updates the shared knowledge based in the event that a new domain concept was discovered.

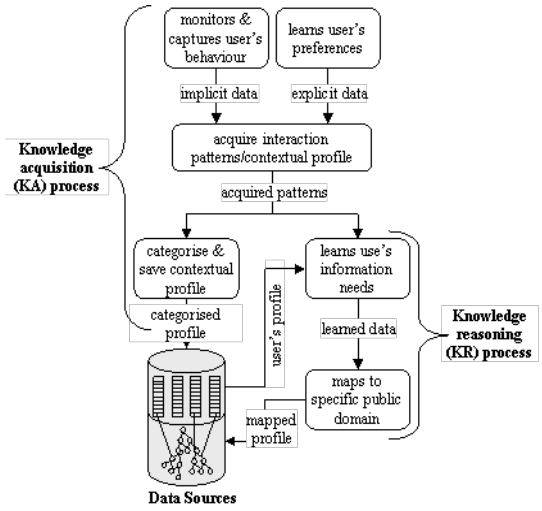

Figure 2: The Contextual Profile Functionality

The KR process takes each user's search query and computes the nearest neighbours in the user's profile to extract candidate domain concepts relevant to the user's query. These candidates are then mapped to concepts in the shared knowledge based.

3.2 Knowledge-base Query Formulation

The Knowledge base query formulation (KbQF) component consists of a Query disambiguation engine (QDE) and Query expansion engine (QEE). Figure 3 provides a summarized representation of the QDE and QEE processes comprising the KbQF component.

The main objective of the KbQF component is to expand a simple keyword query into one or more concept-based queries in order to improve the results of that query. Various query formulation techniques exist, such as user relevance feedback (URF), automatic local analysis (ALA) and automatic global analysis (AGA). The KbQF employs the URF technique to expand a simple keyword query into concept-based search queries. The URF technique has been successfully applied to a wide variety of areas including interactive text-based image retrieval (Zhang, Chai, & Jin, 2005), adaptive Web search (Sugiyama, Hatano, & Yoshikawa, 2004), content-based music retrieval (Hoashi, Matsumoto, & Inoue, 2003), misuse detection in information retrieval systems (Ma & Goharian, 2005), ARCH (Sieg et al., 2004) and so on.

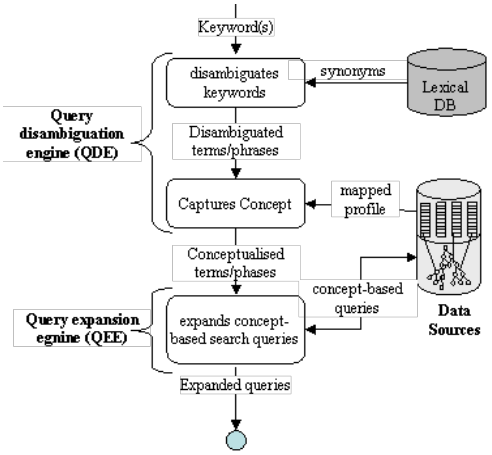

Figure 3: The Knowledge Base Query Formulation Functionality

The QDE starts by accepting the user's input *(i.e. one/more keywords)* as a search query *(e.g. "Surfing")*. The crucial challenge in the QDE process is how to reduce keyword *(i.e. search query)* ambiguity in order to select effective search terms/phrases to formulate concept-based search queries. The QDE uses lexical databases *(e.g. WordNet)* to reduce the ambiguity of the entered keyword(s) *(i.e. user's input)* by presenting the user with potentially relevant terms/phrases. The user selects the terms/phrases that best describe the subject of their query. In addition, the QDE uses the associated domain-specific public domain concept hierarchy *(e.g. computer science ontology)* to conceptualize these disambiguated terms/phrases. By default, the base domain-specific concept or ontology is selected for the user using his/her initial search term and disambiguated search term. However, the user has the option to alternatively select a more relevant ontology. Once the ontology is selected, the QDE dynamically presents a list of classes extracted through the KR process outlined earlier. In addition, the user may select a class that describes their information needs. Figure 4 provides a simple search scenario.

In this search scenario, a user enters a query *"surfing"* and the system assists the user saying *"SURFING may mean: surfing, surfboarding, surfriding"*. The system then presents the user with suggested domain *(i.e. travel)* and related concepts *(i.e. sports and surfing)*. The user may select additional concepts and click on search button.

Next, the QEE uses a Boolean query expansion model to formulate search queries using all of the above information for submission to a search engine *(e.g. Google)*. The simple formula for the Boolean query expansion is as follows:

$$q_m = q_0 \; AND \; (d_1 \; OR \; .. \; d_n) \; AND \; o_n \; AND \; (c_1 \; OR \; .. \; c_n)$$

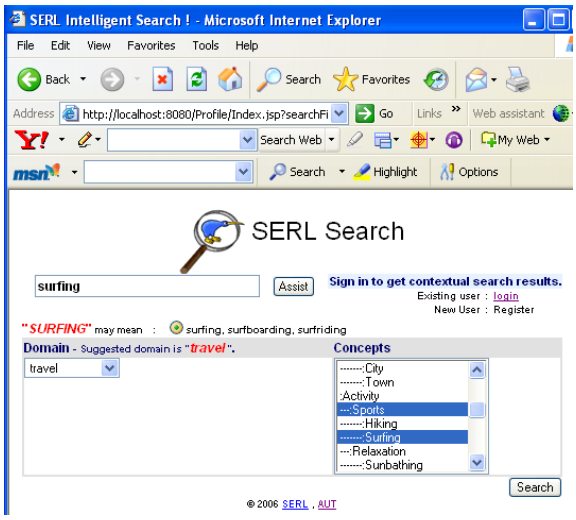

Figure 4: A Simple Search Scenario

In the above formula, q_m = modified query; q_0 = original query; d_1 to d_n = disambiguated term(s); o_n = selected domain name, c_1 to c_n = selected concept(s). Using this simple formula, the QEE generates an enhanced (*or expanded*) query and submit it to search engine. The enhanced query is said to represent the user's search intent more accurately and potentially improves recall and precision. Figure 5 provides the "surfing" search results.

Figure 5: The "surfing" Search Results

In this search results, only travel and sport related surfing results are presented and others results are discarded. Early results indicate that the system has the potential to not only improve the formulation of high quality search query but also contribute towards the long term goal of intelligent search query formulation in web-based information retrieval. The results have shown that the system significantly improves the effectiveness of the search query and improves precision and recall in search results. Precision is improved since ambiguous query terms are disambiguated using lexical databases. Recall is improved since additional concept-based query terms are added in original search query that would not be retrieved by using only the original search query.

However, there are still some unresolved issues in our work. First of all, the system requires a large number of contextual profiles as training data has both technical and ethical challenges. Second, the precision of concept-based search query formulation is directly proportionate to the availability and number of shared contextual profiles. Third, security and privacy are major issues which warrant separate and extensive consideration. Once in the place, the system will be a significant contribution in query formulation research as well as enhancing information retrieval in general.

4. CONCLUSION

This paper has presented the implementation of contextual knowledge-based and concept based query formulation system in web-based information retrieval.

Preliminary experiments have shown that the system significantly improves the effectiveness of the search query and improves precision and recall in search results. Precision is improved since ambiguous query terms are disambiguated using a thesaurus method/linguistic approach. Recall is improved since additional concept-based query terms are added in original search query that would not be retrieved by using only the original search query. These results indicate that the system has the potential to not only improve the formulation of high quality search query but also contribute towards the long term goal of intelligent search query formulation in web-based information retrieval. We believe query formulation is a very promising and challenging research direction and has the potential to improve the quality of search on both the public web and corporate intranets. Improved, more efficient and more relevant searches will benefit businesses who embrace such new approaches.

Our future work will involve further testing of the functionality of the system, analysing the search results and enhancing the concept based query expansion using Boolean method.

5. REFERENCES

Billerbeck, B., Scholer, F., Williams, H. E., & Zobel, J. (2003). *Query expansion using associated queries.* Paper presented at the 12th international conference on Information and knowledge management, New Orleans, LA, USA.

Challam, V. K. R. (2004). *Contextual Information Retrieval Using Ontology-Based User Profiles.* Unpublished Master's Thesis, University of Kansas.

Cui, H., Wen, J. R., Nie, J. Y., & Ma, W. Y. (2002). *Probabilistic query expansion using query logs.* Paper presented at the 11th international conference on World Wide Web, Honolulu, Hawaii, USA.

Fonseca, B. M., Golgher, P., Pôssas, B., Ribeiro-Neto, B., & Ziviani, N. (2005). *Concept-based interactive query expansion.* Paper presented at the 14th ACM international conference on Information and knowledge management, Bremen, Germany.

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition, 5*(2), 199-220.

Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (1999). *An algorithmic framework for performing collaborative filtering.* Paper presented at the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, California, United States.

Hoashi, K., Matsumoto, K., & Inoue, N. (2003). *Personalization of user profiles for content-based music retrieval based on relevance feedback.* Paper presented at the 11th ACM international conference on Multimedia, Berkeley, CA, USA.

Klink, S., Hust, A., Junker, M., & Dengel, A. (2002, 19-23 August). *Collaborative Learning of Term-Based Concepts for Automatic Query Expansion.* Paper presented at the 13th European Conference on Machine Learning, Helsinki, Finland.

Kraft, R., Maghoul, F., & Chang, C. C. (2005). *Y!Q: contextual search at the point of inspiration.* Paper presented at the 14th ACM international conference on Information and knowledge management, Bremen, Germany.

Leake, D. B., & Scherle, R. (2001, January 14 -17). *Towards Context-Based Search Engine Selection.* Paper presented at the International Conference on Intelligent User Interfaces, Santa Fe, New Mexico, United States.

Liu, S., Yu, C., & Meng, W. (2005). *Word sense disambiguation in queries.* Paper presented at the 14th ACM international conference on Information and knowledge management, Bremen, Germany.

Ma, L., & Goharian, N. (2005). *Query length impact on misuse detection in information retrieval systems.* Paper presented at the 2005 ACM symposium on Applied computing, Santa Fe, New Mexico.

O'Hanlon, N. (1999). Off the shelf & onto the Web: Web search engines evolve to meet challenges. *Reference & User Services Quarterly, 38*(3), 247.

Sieg, A., Mobasher, B., Lytinen, S., & Burke, R. (2004, February). *Using Concept Hierarchies to Enhance User Queries In Web-Based Information Retrieval.* Paper presented at the IASTED International Conference on Artificial Intelligence and Applications, Innsbruck, Austria.

Sugiyama, K., Hatano, K., & Yoshikawa, M. (2004). *Adaptive web search based on user profile constructed without any effort from users.* Paper presented at the 13th international conference on World Wide Web, New York, NY, USA.

Taksa, I. (2005, 4-6 April). *Predicting the Cumulative Effect of Multiple Query Formulations.* Paper presented at the International Symposium on Information Technology: Coding and Computing, Las Vegas, Nevada.

Zhang, C., Chai, J. Y., & Jin, R. (2005). *User term feedback in interactive text-based image retrieval.* Paper presented at the 28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil.