Full citation: Limbu, D.K., Connor, A.M., & MacDonell, S.G. (2005) A framework for contextual information retrieval from the WWW. Bulletin of Applied Computing and Information Technology 3(3), December, on WWW.

A Contextual Information Retrieval Framework

Dilip K. Limbu, Andy M. Connor, and Stephen G. MacDonell

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

Abstract

The amount of information on the Internet is constantly growing and the challenge now is one of finding relevant information. Contextual information retrieval (CIR) is a critical technology for today's search engines to facilitate queries and return relevant information. Despite its importance, little progress has been made in CIR due to the difficulty of capturing and representing contextual information about users. Numerous CIR approaches exist today, but, to the best of our knowledge, none of them offer a similar service to the one proposed in this paper. This paper proposes an alternative framework for CIR from the World Wide Web (WWW). The framework aims to improve query results (or make search results more relevant) by constructing a contextual profile based on a user's behaviour, their preferences, and a shared knowledge base, and by using this information in the search engine framework to find and return relevant information.

Keywords: Contextual information, contextual retrieval, contextual model, contextual search.

1. INTRODUCTION

The Internet is, in the simplest terms, a huge, searchable database of information reached via a computer (McQuistan, 2000). It makes available an enormous amount of information, the challenge then being one of finding relevant information (Fan, Gordon, & Pathak, 2004) . Search engines are the most commonly used type of tool for finding relevant information on the Internet. However, even the most experienced searchers/users are finding it increasingly difficult to retrieve relevant information from the World Wide Web (WWW) (Fan et al., 2004; O'Hanlon, 1999). Contextual information retrieval (CIR) is introduced to address these challenges. Numerous CIR approaches – employing contextual user profiles, concept-based query formulation and relevance filtration and relevance feedback/suggestion – already exist today. However, the problem is far from solved. CIR has distinct challenges when compared to either general Information Retrieval (IR) or non-contextual information retrieval from the web.

This paper discuses an alternative framework for CIR from the WWW in the context of the shortcomings of existing search engine technology. The framework aims to improve query results (or make search results more relevant) by constructing a contextual profile based on a user's behaviour, their preferences, and a shared knowledge base, and using this information in the search engine framework to query, filter and return relevant information. This paper also briefly describes the problems and challenges faced in this area, outlines the expected contribution of this work and presents an outline of the proposed research method.

2. PROBLEMS AND CHALLENGES

As useful as they are, today's search engines are far from perfect. Typical search queries are short and are often ambiguous, potentially returning inappropriate results (Leake & Scherle, 2001). Including additional search terms can help to refine the search queries, but it is difficult for even experienced searchers to select the optimum query terms so that the desired subset of information is retrieved (Leake & Scherle, 2001). Moreover, these search engine results are based on simple keyword matches without any concern for the information needs of the user at a particular instance in time (Challam, 2004). A critical goal of successful information retrieval on the web, then, is to identify which pages are of high quality and relevance to a user's query (Sahami, Mittal, Baluja, & Rowley, 2004).

The need to better target a search on the information that will satisfy a user's information needs (Leake & Scherle, 2001) is well recognised. In this regard today's search engines are lacking a personalisation mechanism that can ‘understand' the query or reflect the information needs of a user at a particular instance in time and return customised results (Challam, 2004). To provide the desired information to the user requires effective methods for identifying the user's task context based on available information (Bauer & Leake, 2003). CIR has been and remains one of the major long-term challenges in information retrieval (Allan et al., 2003).

There has been significant research in this area to date that has attempted to overcome the major challenges of CIR, and current research continues to improve the methods used. The key research in this area includes the development of PRISM (Leake & Scherle, 2001), Letizia (Lieberman, 1995), the Wisconsin Adaptive Web

Assistant (WAWA) (Rad & Shavlik, 2003) and Syskill & Webert (Pazzani, Muramatsu, & Billsus, 1996).

Leake's PRISM (Leake & Scherle, 2001) uses *Watson* (Budzik & Hammond, 2000) to monitor user behaviour in standard applications (such as word processors and Web browsers) and predict the type of information likely to be of interest to the user. Search queries are dispatched to special purpose search engines tailored towards the user's particular needs.

Lieberman's *Letizia* (1995) monitors a user's browsing behaviour and develops a user's contextual profile. The system uses the user's contextual profile to search and recommend potentially interesting pages to the user.

WAWA (Rad & Shavlik, 2003) constructs a Web agent by accepting user preferences in the form of instructions and adapting the agent's behaviour as it encounters new information. The system uses machine-learning methods to retrieve and/or extract textual information from the Web.

Pazzani et al.'s (1996) Syskill & Webert asks the user to rank pages on a specific topic. Based on the content and rating of the pages, the system constructs a user profile and predicts whether pages encountered subsequently are likely to be of interest to the user.

Despite the achievements of these approaches, there remains no comprehensive model to describe the CIR process (Wen, Lao, & Ma, 2004) due to the difficulty of capturing and representing knowledge about users, context, and tasks in a general Web search environment (Allan et al., 2003). All of the above-mentioned approaches utilise either user behaviour – such as browsing, reading, and typing – or user preferences – such as explicit ranking, explicit inputs, and explicit instructions – to construct a contextual profile, but not both. These approaches also do not use any form of shared intelligent knowledge base to formalise search queries and to provide relevance feedback or suggestions to the user. In addition, none of these approaches discuss how to use these captured contextual profiles in search engine server environments.

3. EXPECTED CONTRIBUTIONS

The expected contribution of this research project is an alternative framework for CIR from the WWW. The framework, under development at the Auckland University of Technology, has as its primary goals: 1) to develop/utilise technology which constructs for each user a contextual profile, by combining the user's behaviour, the user's preferences and a shared knowledge base; 2) to develop/utilise technology which collects millions of users' contextual profiles from millions of machines; 3) to develop/utilise technology which defines and constructs shared knowledge that can be used to refine search queries and to provide user feedback/suggestions; and 4) to integrate the outcomes of 1-3 in a single framework.

The proposed framework architecture is depicted in Figure 1. The architecture consists of two main models: Profile Collector and Context Manager. The Profile Collector resides on the user's desktop computer and consists of two specialised autonomous agents: Adaptive Agent and Preference Agent. They act as front-end brokers and gather contextual information from the user. The Context Manager resides on the search engine server and consists of four specialised autonomous agents: Context Crawler Agent, Context Knowledge Agent, Query Process Agent and Integration Agent. All these agents perform well-defined functions such as interacting with millions of machines to gather users' contextual profiles, processing the contextual profiles, maintaining the shared knowledge base, formulating contextual queries and filtering and presenting relevant results.

The framework centres on the construction of user contextual profiles by combining user behaviour, user preferences and shared knowledge base information. The shared knowledge base can be used to provide user feedback/suggestions and to refine search queries. The framework requires the collection of millions of users' contextual profiles from millions of machines. All these components are then integrated in a single comprehensive CIR framework. These features contribute to making this framework open, robust and scalable. However, capturing users' contextual profiles and sharing these contextual profiles to construct a shared knowledge base are typically person-dependent. Different users prefer different modes of information capture on their desktops and they may be concerned about the different social implications - such as privacy, spam, hacking and so on – for their shared contextual profiles. These issues warrant separate and extensive consideration. If the proposed framework is fully developed and deployed in a real search environment, the system should make these features explicit to users so they can have control over their preferred modes of information capture and the sharing of their contextual profiles.

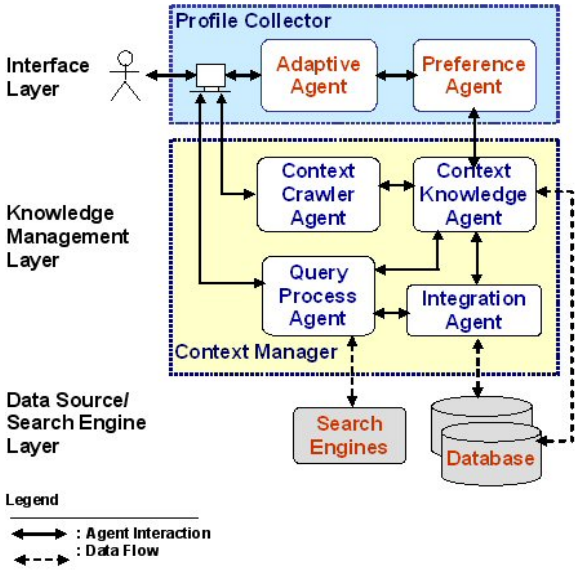

Figure 1. The proposed framework is divided into three layers (interface *layer*, *knowledge management layer* and *search engine layer)* and implementation of the framework requires basic research and development of tools in each layer.

A simple usage scenario of the proposed framework (as *a system)* is given here. A user spent time at his/her desktop computer planning a holiday to New Zealand (*visiting*

travel web pages, booking airline tickets and making hotel reservations online, using Microsoft Word to store travel information, and sending emails to friends about the trip.). The system continuously monitors the user's desktop activities and captures the user's contextual profile. When the user enters a query such as "Surfing", the system turns the "Surfing" keyword into shared concepts using the user's contextual profile and the existing knowledge base (using *various public ontology domains).* The system then understands the meaning of "Surfing" in the current context, i.e. "surfing waves" not surfing the Internet. In addition, the system is also aware of the surfing location and surfing dates (from *the hotel address and booking date).* With this information, the system generates contextual queries (*such as "Surfing in New Zealand", "Surf Tours", "Surf Lessons", "Surf Camps", "Surf Shops", "Northland surfing", "Auckland surfing", "East coast surfing" and so on*) and submits queries to a search engine. The system then filters results from the search engine using shared concepts and returns relevant information to the user. The system also provides useful suggestions/feedback (such *as "Check weather", "Surfing guide", "See surfing pictures" and so on)* to the user to get his/her specific interests.

The proposed architecture is general and modular so that new categorisations, ontologies (Gruber, 1993) and search engines can easily be incorporated. Figure 2 shows how the proposed framework could be integrated with an existing search engine. We believe that the framework will be a significant contribution in CIR research as well as enhancing Information Retrieval (IR) in general.

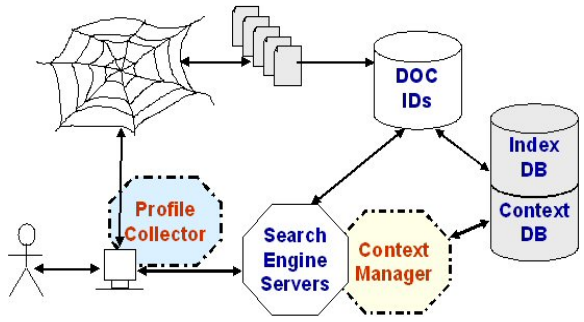

Figure 2. The proposed framework can be integrated with existing search engine technology or deployed as an entity in its own right.

4. RESEARCH METHOD

Our current work is focused on the investigation, design, development and testing of an alternative (new) CIR framework that seeks to create innovations, define new ideas (or practices) and technical capabilities. As such, the System Development Research Methodology (SDRM) (Jay F. Nunamaker, Chen, & Purdin, 1991) and the Design-Science (DS) research guidelines (Hevner, March, Park, & Ram, 2004) are the most appropriate research methodology and research guidelines for this research project. The SDRM methodology consists of several iterative phases, where each phase consists of various research activities. Similarly, the DS research guidelines consist of seven well-defined research guidelines, which essentially complement the iterative phases of the SDRM methodology.

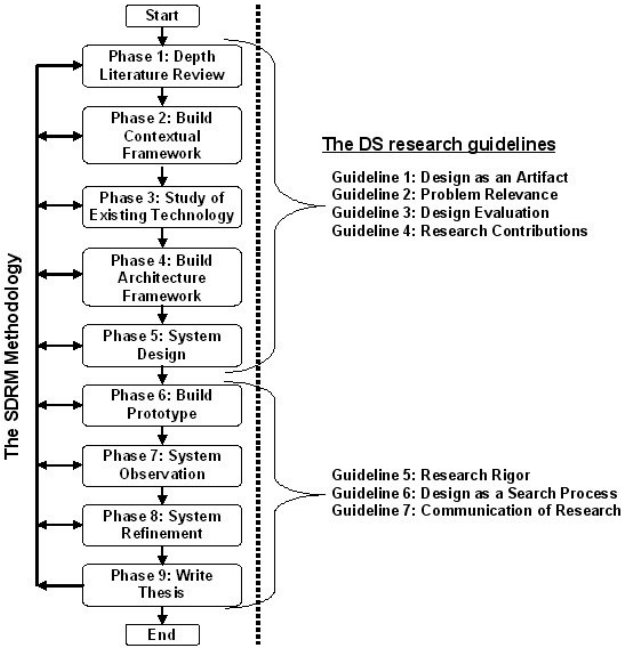

Figure 3. The SDRM Methodology and the DS research guidelines for CIR.

Figure 3 shows the diagrammatic form of the SDRM methodology and the DS research guidelines for CIR. Both the research methodology and the research guidelines must be addressed in some manners to complete this research project. In addition, depending on the research phases or the research guidelines, the research project uses both qualitative (such as observation of user's behaviour) and quantitative empirical research methods (such as testing experiments and simulations) as required.

The research project starts with the *Depth Literature Review* phase, which carries out the in-depth literatures review of existing CIR related literatures. This phase identifies/learns the existing research status, their challenges, future directions and research questions. The *Build Contextual Framework* phase builds a CIR conceptual framework that provides a very high-level system overview of processes and basic components. The *Study of Existing Technology (e.g. tools and techniques)* phase explores and evaluates the existing CIR related technology and identifies the potential usage of these technologies in the proposed conceptual framework. The *Build Architecture Framework* phase builds the proposed CIR architecture that provides the system components details, their relationships and their basic functionalities. The *System Design* phase provides the detailed design of the whole proposed system (such as detail functionality of each component, their independency and database design and so on).

The *Build Prototype* phase develops a working conceptual prototype of the proposed system. This phase tests the prototyped system's feasibility, reliability and performance. The *System Observation* phase observes and evaluates the performance of the prototyped system. The *System Refinement* phase identifies and overcomes any major perceived flaws or shortcomings with the proposed

framework. The phase also outlines direction for future work as it is recognised that this research area is potentially huge and not all desirable features may feasibly be implemented in the timescale. Finally, the research method ends with the *Write Thesis* phase that documents all the findings of above phases.

In addition, all these phases must ensure the DS research guidelines are addressed in some manner. For phases 1 – 4, the research project must produce a viable artefact, which must develop technology-based solutions to important and relevant business problems. The research project must rigorously demonstrate the utility, quality, and efficacy of a design artifact via well-executed evaluation methods. The research project must provide clear and verifiable contributions in the areas of the design artefact and design foundations. Similarly, for phases 5 - 9, the research project must rely upon the application of rigorous methods and search for the best, or optimal, design, as this is often intractable for realistic information systems problems. Finally, the research project must be presented effectively to both technology-oriented as well as management-oriented audiences.

5. CONCLUSION

This paper has presented a research framework for CIR from the WWW that will improve query results (or make search results more relevant). The proposed framework utilises various approaches/techniques to address some of the many acknowledged challenges that exist in the CIR domain.

The proposed framework architecture consists of two main models: Profile Collector and Context Manager. Both models consist of various specialised autonomous agents that perform well-defined functions. These agents support interactive monitoring and capturing of each user's behaviour and preferences, query specification and query processing, contextual profile gathering and categorisation, as well as relevance result filtering and presentation.

The proposed research project uses the SDRM methodology and the DS research guidelines to complete this research project, as it involves constructing an alternative (new) framework for CIR from the WWW that seeks to create innovations, define new ideas (or practices) and technical capabilities.

6. REFERENCES

Allan, J., Harper, D. J., Hiemstra, D., Hofmann, T., Hovy, E., Kraaij, W., et al. (2003). Challenges in inormation retrieval and language modeling. *ACM SIGIR Forum, 37*(2), 31 - 47.

Bauer, T. L., & Leake, D. B. (2003). Detecting context-differentiating terms using competitive learning. *ACM SIGIR Forum, 37*(2), 4 - 17.

Budzik, J., & Hammond, K. J. (2000). User interactions with everyday applications as context for just-in-time information access. Paper presented at the International Conference on Intelligent User Interfaces (ICIUI2000).

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

Fan, W., Gordon, M. D., & Pathak, P. (2004). Discovery of context-specific ranking functions for effective information retrieval using genetic programming. *Knowledge and Data Engineering, IEEE Transactions, 16*(4), 523 - 527.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition,* 5(2), 199-220.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in Information systems research. *MIS Quarterly, 28*(1), 75-105.

Jay F. Nunamaker, J., Chen, M., & Purdin, T. D. M. (1991). Systems development in information systems research. *Journal of Management Information Systems, 7*(3), 89-106.

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.

Lieberman, H. (1995). Letizia: An agent that assists Web browsing. Paper presented at the IJCAI-95.

McQuistan, S. (2000). Techniques for current answers: Part 1: Information overload and the internet. *The Journal of Audiovisual Media in Medicine, 23*(3), 124.

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

Pazzani, M., Muramatsu, J., & Billsus, D. (1996). Syskill & Webert: Identifying interesting web sites. P*roceedings of the National Conference on Artificial Intelligence (AAA1996)*, Portland.

Rad, T. E., & Shavlik, J. (2003). A system for building intelligent agents that Learn to retrieve and extract information. *User Modeling and User - Adapted Interaction, 13*(1-2), 35.

Sahami, M., Mittal, V., Baluja, S., & Rowley, H. (2004). The happy searcher: Challenges in Web information retrieval. Paper presented at the 8th Pacific Rim International Conference on Artificial Intelligence, Auckland, New Zealand.

Wen, J-R., Lao, N., & Ma, W-Y. (2004). *Probabilistic model for contextual retrieval.* Paper presented at the Annual ACM Conference on Research and Development in Information Retrieval, United Kingdom.