A System for Adaptive Information Retrieval

Ioannis Psarras and Joemon Jose

Department of Computing Science, University of Glasgow, Glasgow, G12 8QQ

Abstract. In this paper, we describe the design and development of personal information assistant (PIA), a system aiming to meet individual needs of the searchers. The system's goal is to provide more up-to-date and relevant information to users with respect to their needs and interests. The main component of the system is a profile learner for capturing temporal user needs, based on implicit feedback gathering techniques. It monitors the system usage, the documents viewed and other user actions in order to infer users' changing needs.

1 Introduction

Web search engines, designed for discovering documents online, are very popular and generally perceived to do a good job in finding relevant information on the web. However, recent studies, such as [4, 5], have highlighted that users interact only with a limited number of search results usually among the first page. [4, 5] also demonstrated that, searchers usually choose some relevant information within the first page of results having viewed very few documents. Uncertain about the availability of other relevant documents most users end their search sessions after one or two iterations. In fact, most of the time, they keep looking for information regarding the same topics, for example things that relate to their work. Often such information requirements change by sliding into new topics, based on the changes of user interests. Only way to satisfy such needs is to search on a continuous basis, that is keep looking for information regularly.

In this paper, we argue that a personal information assistant will improve search experience recommending additional documents, relevant to the interests of users. We have developed a system, called PIA (Personal Information Assistant), which makes it easier for people to locate information regarding their needs. Our system adapts to the changing needs of users, manages their multiple search interests, and pro-actively fetches and presents relevant documents on a regular basis. The aim was to build a system capable of modeling people's evolving needs, in an effective way, and use the information provided in order to create a personalized information source for users. The main feature of the system's design, that supports this, is the profiling learning algorithm responsible to discover users' interests. Another key aspect of PIA's design is the extractor algorithm that facilitates implicit information gathering from the sources the user showed some interest. More details, about profile creation and the various algorithms used, are provided in later sections.

V. Wade, H. Ashman, and B. Smyth (Eds.): AH 2006, LNCS 4018, pp. 313–317, 2006.

2 Motivation

With the growth of the World Wide Web the need for tools to address problems with information overload, [6], has become more apparent. However, in many situations the information seeking experience is less than satisfactory: often searchers have difficulty finding relevant information. The main reason for this is the lack of effective search interaction and retrieval tools. The existing tools are often ineffective for all but the most simple search tasks [2]. There are three main areas of user interaction with a search engine: selection of initial query words, the assessment of retrieved pages and query modification [3]. To build an effective search tool one has to address the problems of query formulation and support the formation of information needs that are prone to develop or change during a search.

Past solutions, like [1], used mostly explicit feedback gathering to model the user's searching behavior. WebMate [1] is a search agent that supports both Internet searching and browsing. Using multiple vectors to keep track of user's interests, relevant documents can be suggested to the user. The system automatically attempts to learn the user's categories of interest by requiring the explicit marking of pages during normal browsing. However, this form of relevance feedback increases the user's responsibility which can cause inconvenience or introduce confusion. Other such systems are aimed to pro-actively find and filter relevant information that matches our interests. New interests are stored in a simple profile, containing terms related to different interests and hence resulted in poor performance.

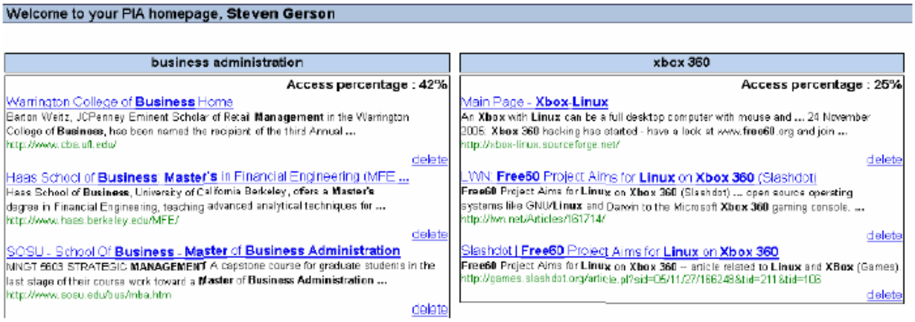
We need a pro-active search assistant that addresses these issues, identifies the multiple facets of user needs and can fetch relevant information. In order to reduce the cognitive load in the feedback issues a combination of implicit and explicit feedback gathering can be much more powerful, since they only require minimal user interaction.

3 PIA - Main Components

Personal Information Assistant was developed as an adjunct to the current web search engines. The system was developed using JAVA Enterprise Edition (J2EE) and is based on a three-tier architecture. User queries and other interaction data is captured and processed at the server. The queries are forwarded to Google, and the results are parsed and presented to the browser. At this stage, the user's profile gets updated to exploit the information gathered from the previously issued search. At some future point, the assistant will analyze the information stored and attempt to retrieve additional relevant documents regarding the user's evolving needs.

In order to help the user in judging relevant information PIA uses a summarization system. We have implemented a version of the system described in [7]. It generates summaries of result pages based on the queries and known as query biased summaries. As demonstrated in [7], such summaries will facilitate more interaction with the system.

The main user interface features a personalized homepage for each user and a profile editor. Each user's home page is similar to a portal, where people can view the documents recommended by the system with respect to their interests. Interests are displayed on a priority basis aiming to improve the retrieval performance of the system, since high-priority facets are likely to be more attractive for the user. The other components of the system are the term extractor algorithm and the profile generation scheme, described in subsequent sections.



Welcome to your PIA homepage, **Steven Gerson**

business administration

Access percentage : 42%

Warrington College of **Business** Home
Barton Weitz, JCPenney Eminent Scholar of Retail **Management** in the Warrington College of **Business**, has been named the recipient of the third Annual ...
http://www.cba.ufl.edu/
delete

Haas School of **Business**: **Master's** in Financial Engineering (MFE ...
Haas School of **Business**, University of California Berkeley, offers a **Master's** degree in Financial Engineering, teaching advanced analytical techniques for ...
http://www.haas.berkeley.edu/MFE/
delete

SOSU - School Of **Business** - **Master** of **Business Administration**
MNGT 5603 STRATEGIC **MANAGEMENT** A capstone course for graduate students in the last stage of their course work toward a **Master** of **Business Administration** ...
http://www.sosu.edu/bus/mba.htm
delete

xbox 360

Access percentage : 25%

Main Page - **Xbox-Linux**
An **Xbox** with **Linux** can be a full desktop computer with mouse and ... 24 November 2005: **Xbox 360** hacking has started - have a look at www.**free60**.org and join ...
http://xbox-linux.sourceforge.net/
delete

LWN: **Free60** Project Aims for **Linux** on **Xbox 360** (Slashdot)
Free60 Project Aims for **Linux** on **Xbox 360** (Slashdot) ... open source operating systems like GNU/**Linux** and Darwin to the Microsoft **Xbox 360** gaming console. ...
http://lwn.net/Articles/161714/
delete

Slashdot | **Free60** Project Aims for **Linux** on **Xbox 360**
Free60 Project Aims for **Linux** on **Xbox 360** -- article related to **Linux** and **XBox** (Games).
http://games.slashdot.org/article.pl?sid=05/11/27/166248&tid=211&tid=106
delete

Fig. 1. The personalized home page, displaying additional documents discovered by the system

3.1 Profile Representation and Management

A profile consists of a set of interests that relate to the user requirements. PIA recognizes that user interests are multiple and hence their profile contain multiple facets of user needs. Fundamentally, an interest constitutes a weighted keyword vector, distinguished by a representative name. Such interests can be temporal, which will be eventually discarded, or long-term needs. As discussed in [4], people usually interact very little with information retrieval tools, such as search engines, so forcing them to add their own profiles would definitely decrease the functionality and usability of our system. Therefore, PIA features techniques to make it possible to modify a user's profile implicitly. Using a term extraction algorithm and a profile learning scheme, interests can be discovered and populated without any user interaction. Explicit profile creation is possible, as well as modification of the system's suggested interests, available through the profile editor interface, but it only constitutes an optional feature.

3.2 The Extractor Algorithm

The extraction algorithm strives to extract a set of representative words, from user search iterations, with respect to their information needs. Apart from the search terms, it takes into consideration the search engine snippets and summaries of recently viewed documents, since these directly reflect to the user's information need. Query terms directly express the user's search requirements and are applied an extra weight compared to the other words in the set. For experimentation purposes, three retrieval models have been made available: A boolean model, a frequency model and the well-known TF-IDF model.

After performing thorough tests with all schemes and measuring their performance in a variety of circumstances, we deduced that the frequency model gives the best results for out application. Taking into account the frequency model's formula, where the weight of each term equals to the number of times it occurs in the collection, it is easy to observe that the extracted set of terms is less likely to be random.

3.3 The Profiling Generation Algorithm

The profile learning model is based on the assumption that users will always visit documents related to their search requirements. Therefore, after the user performs a search, web pages that have been viewed are considered to be more relevant than the rest in the result collection. The profiler extracts the most representative words for a query, by continuously monitoring user interaction and exploiting this information to discover representative terms. Before providing a more detailed description of how profiles are created implicitly, some knowledge is needed regarding the extractor algorithm.

We used clustering techniques to detect various facets of users' interests. Having extracted a set of terms from the visited documents in the result set, a single-pass clustering algorithm is applied, using using cosine coefficient as the similarity matching function.

As discussed earlier on, the aim was to profile users' requirements with the least possible effort from them. One of the main issue, that arises, is labeling interests. Asking users to fill in interests' names explicitly is not an option, because we want to go beyond this explicit feedback gathering model. Cluster labeling is one of the most major information retrieval issues. Due to time constraints, simple, but efficient, cluster and interest labeling algorithms were used for this purpose. The most frequent terms, appearing in an interest or a cluster, are most likely to describe it correctly. So, by labeling the interest using the n most popular terms seems to work effectively enough and hence used.

3.4 Finding Additional Documents

The reason for implementing all these algorithms is to present to the user adequate additional relevant documents related to his interests. When the user re-visits the portal, he will get a listing of new documents associated with his interests in a personalized home page, illustrated in figure 1 above. The system takes advantage of the profiler algorithm and formulates a new query by extracting n most frequent words from an interest. During the implementation and evaluation of our system, it was observed that 4-6 query terms are adequate to retrieve additional relevant documents from the web. Finally, it issues an online search using the query formulated and adds m documents retrieved to the related interest.

This process occurs whenever a user logs in to the system, but at most once a day to avoid updating the suggested documents too often. In fact, the whole document updating procedure is completely transparent to the user with no slow-down at all. The system suggests them some documents that might be of interest and they decide whether they want to delete them or not.

4 Conclusion

We have designed, deployed and evaluated a system aiming to supply users with up-to-date information regarding their personal needs. By using an implicit information gathering model we eliminate the necessity of forcing users to create their profiles explicitly. By formulating queries based on the users' interests and automatically seek more information on the web, the assistant recommends additional documents that might be of interest to the users. We also present techniques to keep up with users' evolving needs effectively such as the term extractor scheme and the profile management algorithm.

Acknowledgments

The work reported in this paper is partly funded by the Engineering and Physical Research Council (EPSRC), UK grant number EP/C004108/1.

References

1. Chen, L., & Syraca, K., Webmate: A Personal Agent for Browsing and Searching, Proceedings of the 2nd International Conference on Autonomous Agents, 132-139
2. Dennis, S., McArthur, R. and Bruza, P. (1998). Searching the WWW made easy? The Cognitive Load imposed by Query Refinement Mechanisms. Proceedings of the 3rd Australian Document Computing Symposium.
3. Ingwersen, P. and Willett, P. (1995). An introduction to algorithmic and cognitive approaches for information retrieval. Libri. 45(3/4), 160-177.
4. Jansen, B.J. and Pooch, U. (2000). A Review of Web Searching Studies and a Framework for Future Research. Journal of the American Society for Information Science and Technology. 52(3), 235-246
5. Jansen, B.J., Spink A. and Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of users on the Web. Information Processing and Management. 36(2), 207-227.
6. Nelson, M.R. (1994). We Have the Information You Want, But Getting It Will Cost You: Being Held Hostage by Information Overload. ACM Crossroads. 1(1).
7. White, R. W., Jose, J. M. and Ruthven, I. (2003). A task-oriented study on the influencing effects of query-biased summarisation in Web searching. Information Processing & Management, 39(5), 707- 733.