Contextual Relevance Feedback in Web Information Retrieval

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

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

ABSTRACT

In this paper, we present an alternative approach to the problem of contextual relevance feedback in web-based information retrieval. Our approach utilises a rich contextual model that exploits a user's implicit and explicit data. Each user's implicit data are gathered from their Internet search histories on their local machine. The user's explicit data are captured from a lexical database, a shared contextual knowledge base and domain-specific concepts using data mining techniques and a relevant feedback approach. This data is later used by our approach to modify queries to more accurately reflect the user's interests as well as to continually build the user's contextual profile and a shared contextual knowledge base. Finally, the approach retrieves personalised or contextual search results from the search engine using the modified/expanded query. Preliminary experiments indicate that our approach has the potential to not only aid in the contextual relevance feedback but also contribute towards the long term goal of intelligent relevance feedback in web-based information retrieval.

Keywords: Contextual information retrieval, Contextual user profile, Relevance feedback, Query formulation.

1. INTRODUCTION

The growing amount of online digital information (*e.g. web documents*) on the Internet has created a need for contextual user profiles and relevance feedback (RF) [1] to be used in order to better meet users' information needs. The contextual user profile approach leverages a user's behaviour – *such as browsing, reading, and typing* – and their preferences – *such as explicit rankings, inputs, and instructions* – and then evaluates web page relevance in terms of its content and the user's context. The relevance feedback approach provides a means for automatically reforming a query to more accurately reflect the user's interests [2]. Both approaches have been studied for some time with varying degrees of success. Despite their long history in information retrieval (IR) research, these approaches have not been successfully implemented in web-based information retrieval [3]. This is partly due to the difficulty of capturing and representing knowledge about users, context, and tasks in a general web search environment [4]. In addition, this is partly due to the fact that users do not understand the mechanisms of the relevance feedback algorithms, creating user uncertainty concerning their purpose and impact. Also, providing relevance judgments requires additional effort on the part of the users [5]. The combination of these factors – the lack of a personalisation mechanism and the lack of understanding of relevance feedback algorithms – presents great research challenges related to contextual relevance feedback in web-based information retrieval (*e.g. search engines*).

In this paper, we present an alternative approach as a solution to the problem of contextual relevance feedback in web-based information retrieval. Our approach utilises a rich contextual model that exploits a user's implicit and explicit data to build a user's contextual profile. The approach builds a shared contextual knowledge base by consolidating various users' contextual profiles. It also employs a data mining technique to learn each user's specific information needs and employs a relevance feedback approach to support the iterative development of a search query by suggesting alternative terms/metakeywords/concepts for query formulation. The end result is a query enriched with additional search terms which can then be submitted to a search engine such as Google.

The remainder of this paper is organised in the following way. Section 2 presents related research. Section 3 describes the overview of the contextual relevance feedback architecture and presents our preliminary empirical work to date. Section 4 offers some concluding remarks as well as future research directions.

2. RELATED RESEARCH

The contextual relevance feedback approach has well recognised research challenges in web-based information retrieval. Various contextual relevance feedback approaches have been studied for some time with varying degrees of success. Here we review some of the related contextual relevance feedback research work.

A recent approach is that of Fonseca et al. [6] that uses extracted concepts from a special type of query relation graph to expand the original query. The extracted concepts are then shown to the user who selects the concept that is interpreted to be most related to the query. This concept is used to expand the original query and the expanded query is then processed in place of the original.

Shen et al.'s UCAIR [7] system uses a client-side web search agent that can perform eager implicit feedback, e.g., query expansion based on previous queries and immediate result re-ranking based on click-through information to provide a personalised search.

Unlike traditional relevance feedback methods, Sieg et al.'s ARCH [8] system uses the domain knowledge inherent in Web-based classification hierarchies such as Yahoo, combined with a user's profile information, to add just those terms likely to improve the match with the user's intent.

Rad et al.'s WAWA [9] system constructs a Web agent by accepting the user preferences in the form of instructions. These user-provided instructions are compiled into neural networks that are responsible for the adaptive capabilities of an intelligent agent. The system expands the initial query and uses machine-learning methods to retrieve and/or extract textual information from the Web.

Klink et al.'s RUBIC [10] system makes use of user preferences and search keywords to identify phrases using the relevance feedback approach. The system expands each phrase with a concept using existing stored concepts and results to produce a new query which is presented to the user for confirmation. After confirmation, such reformulated queries are submitted to search engines and the user is then presented with a hit list of relevant documents.

Zhang et al.'s WAIR [11] system learns the user's interests by observing their behaviour during interaction with the system. The system is then trained on the explicit feedback from the user. After this learning phase, the system estimates the relevance feedback implicitly based on the observations of the user actions. This information is used to modify the user profile. A retrieval agent constructs a query using the user profile and gets relevant URLs from existing Web-index services, e.g. AltaVista, Excite, and Lycos. The system then presents the highest-ranked documents to the user.

Budzik et al.'s Watson [12] system observes user interaction with everyday standard software tools – such as browsers and word processors – and generates queries on behalf of users as well as providing an interface by which the user can pose queries explicitly to WWW search engines for context-relevant information.

Chen et al.'s WebMate [13] learns and keeps track of user interests incrementally and with continuous update, it automatically provides documents that match the user interests. The system takes multiple pages provided by the user as relevance guidance and it extracts and combines relevant keywords from these pages and uses them for keyword refinement. It also provides relevance feedback during search to improve relevant search results.

Fensel et al.'s OntoBroker [14] system is a semantic indexing and instance querying technology for the WWW based on the use of ontologies. Visualisation (*or relevance feedback*) is employed to help users select classes and attributes for building queries. The hyperbolic technique allows a quick overview, which aids navigation of classes far away from the current focus, as well as allowing a closer examination of classes and their vicinity.

Krulwich et al.'s InfoFinder [15] system learns profiles of user interests from sample documents that users submit while browsing, without surveying users as to their interest in a set of sample documents. The system learns general profiles from the documents by heuristically extracting phrases that are likely to represent the document's topic. The InfoFinder's learning algorithm generates a search tree and translates this into a Boolean search string for submission to a generic search engine.

All these state-of-art contextual relevance feedback approaches expand the original search query by adding additional/extracted information using various techniques such as contextual user profiles *(i.e. user's behaviour and their preferences)*, ontology [16] or concept-based enhancement *(e.g. domain specific ontology)*, domain knowledge *(e.g. lexical databases, Web based classification hierarchies, etc.)*, or query logs *(e.g. search query logs, user logs, etc.)*.

Despite the success achieved by these approaches, to the best of our knowledge none of them offer a similar service to our approach. In contrast, our approach is distinct to previous approaches and consists of four steps:

Step (1) – it gathers the user's implicit data, *such as previously issued search queries, previously visited URLs and Meta keywords from those visited URLs*. This information is extracted from the user's Internet search histories on their local machine.

Step (2) – it captures the user's explicit data, *such as alternative term/phrases, Meta keywords, ontology and concepts*. This data is sourced from a lexical database, a shared contextual knowledge base and domain-specific ontology/concepts.

Step (3) – it constructs the user's contextual profile and a shared contextual knowledge base using data from step 1 and step 2.

Step (4) – finally, it modifies the user's initial query to more accurately reflect the user's interests using steps 1, 2 and 3.

To summarise, our approach captures the user's adaptive search context/intent by monitoring and capturing their implicit and explicit activities. In addition, the approach eases the user's relevance judgment task by presenting potentially relevant metakeywords/concepts in the course of query formulation. Hence, our approach should assist experienced and inexperienced users to find relevant information from the Internet and aims to make two main contributions. First, it will experimentally demonstrate the construction and use of an evolving contextual user profile and the shared contextual knowledge base to define the user's search context, which can be refined over the time. Second, it will experimentally demonstrate the formulation of a dynamic search query by employing a data mining technique to learn the user's specific information needs while employing the relevance feedback approach to support iterative development of the search query by suggesting alternative terms/concepts for query formulation. The following section describes the approach in more detail.

3. OVERVIEW OF THE CONTEXTUAL RELEVANCE FEEDBACK ARCHITECTURE

Figure 1 illustrates the overall architecture of the contextual relevance feedback approach. The two main components of the architecture are Behaviour Collector (BC) and Preference Collector (PC). The approach is prototyped using JAVA technology and deployed as a web-based application in Apache Tomcat servlet container. The approach is integrated with WordNet, Jena, Weka [17] and Google technology. The architecture is general and modular so that new ontology and search engines can be easily incorporated. The following paragraphs discuss each component in more detail.

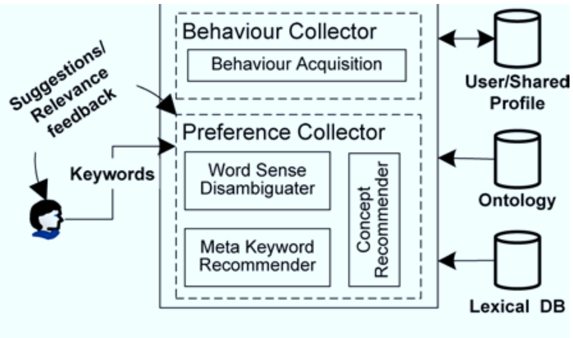

Fig. 1. The Contextual Relevance Feedback Architecture.

3.1 Collector (BC) Component

Figure 2 provides a summarised depiction of the functionality of the BC component, centred on a Behaviour Acquisition (BA) process. The BA process builds a user's contextual profile by extracting the user's information seeking behaviour from their Internet search history logs. The logs typically record information detailing previously submitted search queries and visited URLs. Various existing approaches have discussed the extraction of user information from Internet search history logs, including adaptive web search [18], personalizing search [19], mining navigation [20] and so on.

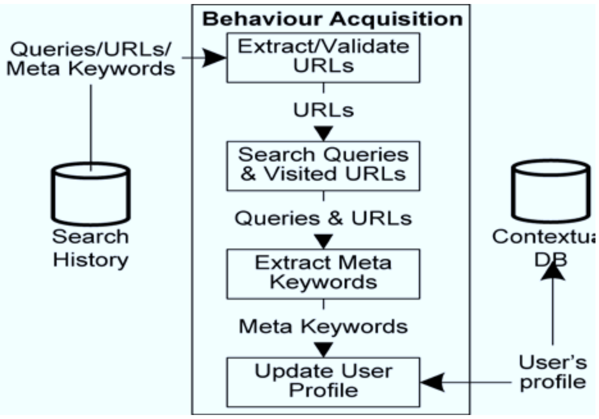

Fig. 2. The Behaviour Collector Functionality.

The BA process extracts and validates all visited URLs from the user's Internet search history logs. At the same time, it extracts the submitted search queries *(e.g. q0, q1, ..., qn)* and subsequently visited URLs *(e.g. u0, u1, ..., un)* from the validated URLs. In the next step, it extracts the Meta keywords *(e.g. m0, m1, ..., mn)* from each subsequently visited URL. Finally, it stores all this information incrementally as an initial contextual user profile as shown in Table 1.

Table 1. Example of User's Profile.

Query	Visited URLs	Meta Keywords
q1	*u1*	*m1, m2*
q1	*u2*	*m3, m4, m5*
q 2	*u3*	*m6*
q 2	*u4*	*m7, m8*
q 2	*u5*	*m9, m10, m11*

For example, for a query *q1*, the visited URLs are *u1* and *u2*, and the extracted Meta keywords for *u1* are (*m1* and *m2*) and for *u2* are (*m3, m4* and *m5*). The stored contextual user profile could be used to present long term and short term preferences. This contextual user profile information is later used by the PC component as described in Section 3.2.

3.2 Preference Collector (PC) Component

Figure 3 provides a summarised depiction of the functionality the PC component, consisting of Word Sense Disambiguater (WSD), Meta Keyword Recommender (MKR) and Concept Recommender (CR) processes.

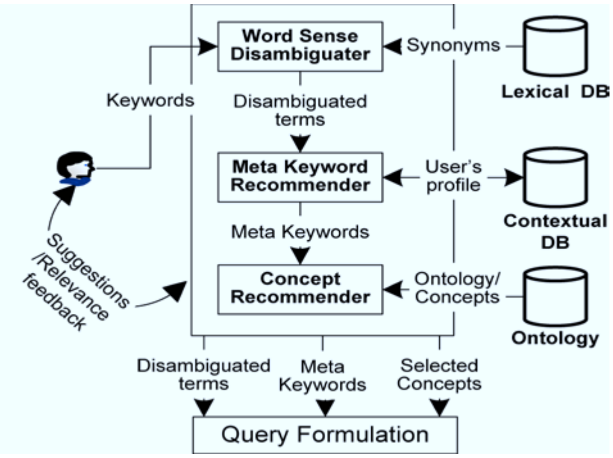

Fig 3. The Preference Collector Functionality.

The main objective of the PC component is to a capture user's preferences and at the same time to expand a simple keyword query into a more effective query in order to improve the results of that query. The MKR and CR processes employ the nearest neighbour data mining technique to learn each user's specific information needs. The nearest neighbour algorithm has previously been applied successfully to classify/recommend relevant information to users including adaptive nearest neighbour [21], nearest neighbor search, [22], and query chains [23]. In addition, the MKR and CR processes employ the relevance feedback approach to support the iterative development of a search query by suggesting alternative meta-keywords/concepts for query formulation. Similarly, the relevance feedback approach has been successfully applied to a wide variety of areas including interactive text-based image retrieval [24], content-based music [25], and misuse detection in information retrieval systems [26].

The PC component starts with the WSD process by accepting the user's input *(i.e. one/more keywords)* as a search query *(e.g. "Surfing")*. The process uses WordNet to disambiguate the entered keyword(s) *(i.e. user's input)* by presenting the user with potentially relevant terms/phrases. WordNet has been used as a word sense disambiguation tool in queries [27], geographical information retrieval systems [28], text-to-concept mappings [29] and so on. The user selects the disambiguated terms/phrases *(e.g.d0, d1, ..., dn)* that best describe the subject of their query.

Next, the MKR process takes the selected terms/phrases and computes the nearest neighbours in the user's contextual profile and the shared contextual knowledge base to extract a pool of Meta keywords (*i.e. extracted through the BA process*) relevant to the user's query. The user then selects the Meta keywords *(e.g. m0, m1, ..., mn)* from the pool that best describe their search intent. Similarly, the CR process takes the disambiguated terms/phrases and the selected Meta keywords and computes the nearest neighbours in the user's contextual profile and the shared contextual knowledge base to map to the associated domain specific public domain concept hierarchy *(e.g. computer science ontology)*. Systems such as OntoBroker [14], RUBIC [10], and WebSifter II [30] have used publicly available ontologies to extract additional query terms/concepts. By default, the CR process presents the user with the selected ontology and domain specific concepts using his/her initial search terms, the disambiguated search terms/phrases and the selected Meta keywords. However, the user has the option to alternatively select a more relevant ontology and a list of relevant concepts using the shared contextual knowledge base. The user may select classes *(e.g. c0, c1, ..., cn)* that describe their information needs. Finally, the PC component stores the user's preferences data as a shared contextual user profile for future use as shown in Table 2.

Table 2. Example of a Shared Contextual User's Profile.

Query	Disambiguated Terms	Meta Keywords	Concepts
q1	*d1, d2*	*m1, m2*	*c1, c2*
q 2	*d3*	*m3, m4, m5*	*c3*
q 3	*d4*	*m6*	*c4, c5, c5*

For example, for a query *q1*, the selected disambiguated terms are *d1* and *d2*, selected Meta keywords are $m1$ and $m2$ while the selected concepts are $c1$ and $c2$.

The shared contextual knowledge base could be used to suggest or recommend Meta keywords, ontology and concepts to other users with similar contextual profiles. Finally, the Query Formulation (QF) component, which is an independent component, uses a Boolean query expansion technique to formulate search queries using all of the above information (or parts, thereof) for submission to Google's search engine. The simple formula for the Boolean query expansion is as follows:

With all information;

$$qm = q0\ \text{AND}\ (d1\ \text{OR}\ ..dn)\ \text{AND}\ (m1\ \text{OR}\ldots mn)\ \text{AND}\ on\ \text{AND}\ (c1\ \text{OR}\ldots cn)(1)$$

With disambiguated terms/phases and selected concept information;

$$qm=q0\text{AND}(d1\ \text{OR}\ ...dn)\ \text{AND}\ on\ \text{AND}\ (C1\ \text{OR}\ ...Cn)\ (1)$$

In the above formulae (1 & 2), qm = modified query; qo = original query; $d1$ to dn = disambiguated term(s); $m1$ to mn = selected Meta Keyword(s); on = selected domain name, $c1$ to cn = selected concept(s).

Using these simple formulae, the QF generates an enhanced (*or expanded*) query and submits it to the search engine. The enhanced query is said to represent the user's search intent more accurately and potentially improves recall and precision.

We have tested our approach with six users in preliminary experiments with simulated data (*i.e. contextual user profile and shared contextual knowledge base*). All subjects were educated to graduate level and used the Internet on a regular basis. Users performed a series of search tasks and were asked to compare the results achieved using our approach and to those achieved using their normal search engine. Four out of six users agreed that our approach improved the effectiveness of the search query and improved precision and recall in search results. In our approach, the precision is improved since ambiguous query terms are disambiguated using the lexical database. Similarly, the recall is improved since additional Meta Keywords and concept-based query terms are added to the original search query that would not be retrieved by using only the original query.

However, in our work there remain many research issues and technical details that need to be investigated. First of all, it is known that users are often reluctant to make the extra effort to provide explicit relevant feedback [31]. As a result, building the user's contextual profile and the shared contextual knowledge base is a challenge as the system requires a large number of such profiles to train the nearest neighbour classifier. That said, this may be a transitional issue that will resolve itself as users receive more value from and place more trust in the system and/or in the Internet generally. Second, data mining may also present ethical challenges as information on individual users' browsing behaviour is scrutinised. Security and privacy are major issues that warrant separate and extensive consideration. Third, the scalability of our approach has not been investigated to any extent at this preliminary stage. Fourth, the precision of query formulation is directly proportionate to the availability and number of shared contextual profiles. Once addressed, we hope that the system will be a significant contribution to contextual relevance feedback research as well as enhancing information retrieval in general.

4. CONCLUSION

This paper has presented ongoing research on the implementation of the contextual relevance feedback approach in web-based information retrieval. The approach builds a contextual user profile employing the user's implicit data (*i.e. from Internet browsing history*) and explicit data (*i.e. from a lexical database, a shared contextual knowledge base and domain-specific ontology/concepts*) to provide relevant information to users that potentially satisfies their information needs.

Preliminary experiments have shown that the system generally 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 Meta Keyword and concept-based query terms are added to the 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 queries but also contribute towards the long term goal of intelligent search query formulation in web-based information retrieval.

Our future work will involve adding to and testing the functionality of the system, analysing the search results and enhancing the effective query expansion using Boolean methods.

5. REFERENCES

1. Rocchio, J., Relevance feedback in information retrieval, in The SMART Retrieval System: Experiments in Automatic Document Processing, G. Salton, Editor. 1971, Prentice Hall.p. 313-323.

2. Allan, J. Incremental relevance feedback for information filtering. In the 19th annual international ACM SIGIR conference on Research and development in information retrieval. 1996. Zurich, Switzerland: ACM Press.

3. Croft, W.B., S. Cronen-Townsend, and V. Lavrenko. Relevance Feedback and Personalization: A Language Modeling Perspective. In DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries. 2001. Dublin City University, Ireland.

4. Allan, J., et al., Challenges in Information Retrieval and Language Modeling. ACM SIGIR Forum, 2003. Volume 37: p. 31 - 47.

5. Vinay, V., et al. Comparing relevance feedback algorithms for web search. In 14th international conference on World Wide Web. 2005. Chiba, Japan: ACM Press.

6. Fonseca, B.M., et al. Concept-based interactive query expansion. In 14th ACM international conference on Information and knowledge management. 2005. Bremen, Germany: ACM Press.

7. Shen, X., B. Tan, and C. Zhai. Implicit User Modeling for Personalized Search. In 14th ACM international conference on Information and knowledge management. 2005. Bremen, Germany: ACM Press.

8. Sieg, A., et al. Using Concept Hierarchies to Enhance User Queries In Web-Based Information Retrieval. In The IASTED International Conference on Artificial Intelligence and Applications. 2004. Innsbruck, Austria.

9. Rad, T.E. and J. Shavlik, A System for Building Intelligent Agents that Learn to Retrieve and Extract Information. User Modeling and User - Adapted Interaction, 2003. 13(1-2): p. 35.

10. Klink, S., et al. Collaborative Learning of Term-Based Concepts for Automatic Query Expansion. In 13th European Conference on Machine Learning. 2002. Helsinki, Finland.

11. Zhang, B.T. and Y.-W. Seo. Personalized Web document Filtering Using Reinforcement Learning. In Applied Artificial Intelligence. 2001.

12. Budzik, J. and K. Hammond. Watson: Anticipating and Contextualizing Information Needs. In American Society for Information Science. 1999.

13. Chen, L. and K. Sycara. WebMate: A Personal Agent for Browsing and Searching. In International Conference on Autonomous Agents. 1998. Minneapolis, Minnesota, United States.

14. Fensel, D., et al. Ontobroker: How to make the WWW Intelligent. In 11th Knowledge Acquisition Workshop. 1998. Banff, Alberta, Canada.

15. Krulwich, B. and C. Burkey, The InfoFinder agent: learning user interests through heuristic phrase extraction. IEEE Expert, 1997. 12(5): p. 22 - 27.

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

17. Witten, I.H. and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. 2005, San Francisco: Morgan Kaufmann. 525.

18. Sugiyama, K., K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In 13th international conference on World Wide Web. 2004. New York, USA: ACM Press.

19. Teevan, J., S.T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In 28th annual international ACM SIGIR conference on Research and development in information retrieval. 2005. Salvador, Brazil: ACM Press.

20. Fu, X., J. Budzik, and K.J. Hammond. Mining navigation history for recommendation. In 5th international conference on Intelligent user interfaces. 2000. New Orleans, Louisiana, United States: ACM Press.

21. Ku, W.S., et al. Adaptive nearest neighbor queries in travel time networks. In 13th annual ACM international workshop on Geographic information systems. 2005. Bremen, Germany: ACM Press.

22. Tešic, J. and B.S. Manjunath. Nearest Neighbor Search for Relevance Feedback. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2003. Madison, Wisconsin.

23. Radlinski, F. and T. Joachims. Query Chains: Learning to Rank from Implicit Feedback. In 11th ACM SIGKDD international conference on Knowledge discovery in data mining. 2005. Chicago, Illinois, USA: ACM Press.

24. Zhang, C., J.Y. Chai, and R. Jin. User term feedback in interactive text-based image retrieval. In 28th annual international ACM SIGIR conference on Research and development in information retrieval. 2005. Salvador, Brazil: ACM Press.

25. Hoashi, K., K. Matsumoto, and N. Inoue. Personalization of user profiles for content-based music retrieval based on relevance feedback. In 11th ACM international conference on Multimedia. 2003. Berkeley, CA, USA: ACM Press.

26. Ma, L. and N. Goharian. Query length impact on misuse detection in information retrieval systems. In 2005 ACM symposium on Applied computing. 2005. Santa Fe, New Mexico: ACM Press.

27. Liu, S., C. Yu, and W. Meng. Word sense disambiguation in queries. In 14th ACM international conference on Information and knowledge management. 2005. Bremen, Germany: ACM Press.

28. Buscaldi, D., P. Rosso, and E.S. Arnal. A WordNet-based Query Expansion method for. Geographical Information Retrieval. in Cross-Language Evaluation Forum 2005 WORKSHOP. 2005. Vienna, Austria.

29. Bonino, D., F. Corno, and F. Pescarmona. Automatic learning of text-to-concept mappings exploiting WordNetlike lexical networks. In the 2005 ACM symposium on Applied computing. 2005. Santa Fe, New Mexico: ACM Press.

30. Kerschberg, L., W. Kim, and A. Scime. WebSifter II: A Personalizable Meta-Search Agent Based on Weighted Semantic Taxonomy Tree. In International Conference on Internet Computing 2001(IC'2001). 2001.

31. Kelly, D. and J. Teevan, Implicit feedback for inferring user preference: A bibliography. SIGIR Forum, 2003. 32(2).