C. Srinvas / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 6, November- December 2012, pp.234-241

Explicit User Profiles for Semantic Web Search Using XML

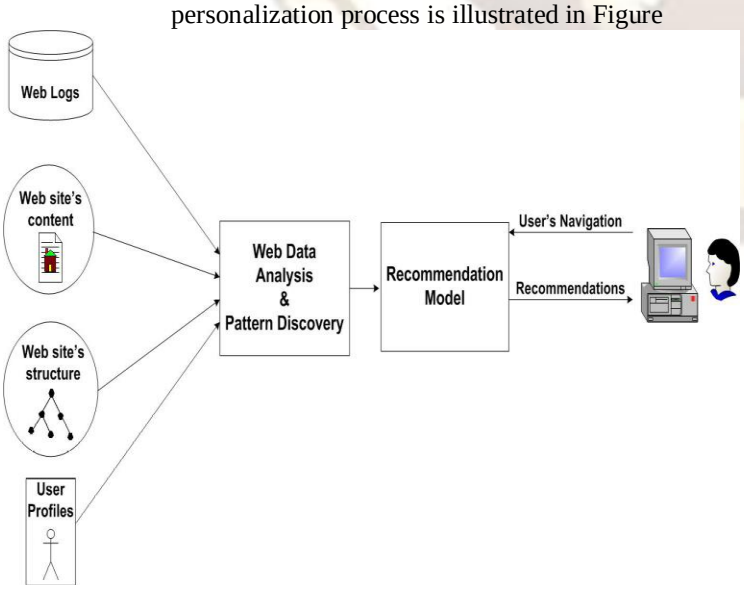
C. Srinvas
Associate Professor & HOD, Dept of Information Technology,
Sree Visvesvaraya Institute of Technology & Science (SVITS), Mahabubnagar, Andhra Pradesh, India

Abstract
The web is diffusing day to day as the number of users are increasing. Nowadays, end users always interested to extract meaningful information from the surplus of accessible Web resources on time. This leads to many problems such as information overload, irrelevant information supply. Filtering irrelevant information is a key issue there are many solutions, among them *personalization* of information services on the *Semantic Web* offers a promising solution to alleviate these problems, and to customize the web environment according to user's information needs, interests and preferences. Explicit user profiles are determined and maintained with the help of XML. This helps the search engine presenting the most relevant and expected results to the user.

Key wards: User profiles, Semantic web, Personalization

Introduction

Web personalization can be defined as any action that customizes the information or services provided by a web site to an individual user, or a set of users, based on knowledge acquired by their navigational behavior, recorded in the web site's logs, in other words, its usage. This information is often combined with the content and the structure of the web site, as well as the interests/preferences of the user, if they are available. The web personalization process is illustrated in Figure



above mentioned sources of information as input to pattern discovery techniques; the system tailors the provided content to the needs of each visitor of the web site. The personalization process can result in the dynamic generation of recommendations, the creation of index pages, the highlighting of existing hyperlinks, the publishing of targeted advertisements or emails, etc

Recommender Systems – A guidance based system tries to automatically recommend hyperlinks that are deemed to be relevant to the user's interests, in order to facilitate access to the needed information on a large website. It is usually implemented on the Web server, and relies on data that reflects the user's interest implicitly (browsing history as recorded in Web server logs) or explicitly (user profile as entered through a registration form or questionnaire). This approach will form the focus of our overview of Web personalization.

Different Ways to Compute Recommendations

Content-based filtering systems are solely based on individual users' preferences. The system tracks each user's behaviour and recommends them items that are similar to items the user liked in the past.

Collaborative filtering systems invite users to rate objects or divulge their preferences and interests and then return information that is predicted to be of interest for them. This is based on the assumption that users with similar behaviour (for example users that rate similar objects) have analogous interests.

In rule-based filtering the users are asked to answer to a set of questions. These questions are derived from a decision tree, so as the user proceeds on answering them, what she/he finally receives as a result (for example a list of products) is tailored to their needs. Content-based, rule-based and collaborative filtering may also be used in combination, for deducing more accurate conclusions.

For example, lazy modeling is used in collaborative filtering which simply stores all users' information and then relies on K-Nearest-Neighbors (KNN) to provide recommendations from the previous history of similar users. Frequent itemsets, a partitioning of user sessions into groups of similar sessions, called session clusters or user profiles can

also form a user model obtained using data mining. Association rules can be discovered offline, and then used to provide recommendations based on web navigation patterns.

Among the most popular methods, the ones based on collaborative filtering and the ones based on fixed support association rule discovery may be the most difficult and expensive to use. This is because, for the case of high-dimensional and extremely sparse Web data, it is difficult to set suitable support and confidence thresholds to yield reliable and complete web usage patterns. Similarly, collaborative models may struggle with sparse data, and do not scale well to a very large number of users.

Several methods for extracting keywords that characterize web content have been proposed. The similarity between documents is usually based on exact matching between these terms. The need for a more abstract representation that will enable a uniform and more flexible document matching process imposes the use of semantic web structures, such as ontologies. By mapping the keywords to the concepts of an ontology, or topic hierarchy, the problem of binary matching can be surpassed through the use of the hierarchical relationships and/or the semantic similarities among the ontology terms, and therefore, the documents.

User profiling: In the Web domain, user profiling is the process of gathering information specific to each visitor, either explicitly or implicitly and representation of the user within the system. It is the first challenge to a personalized search system. A user profile includes demographic information about the user, their interests and even their behavior when browsing a Web site. This information is exploited in order to customize the content and structure of a Web site to the visitor's specific and individual needs.

Semantic Web: "The Semantic Web is an extension of the current web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation."

The users' navigation in a web site is typically content-driven. The users usually search for information or services concerning a particular topic. Therefore, the underlying content semantics should be a dominant factor in the process of web personalization. SEWeP (standing for Semantic Enhancement for Web Personalization), a web personalization framework that integrates content semantics with the users' navigational patterns, using ontologies to represent both the content and the usage of the web site. The whole process bridges the gap between Semantic Web and Web Personalization areas, to create a Semantic Web Personalization system.

The three stages in User Modeling

Stage-1: Collecting the Input Data

The Two Paradigms: - Explicit vs Implicit essentially, we have to have a model for collecting feedback from the user on various items. In Explicit User Modeling, the user is explicitly asked to rate his likeness for various items. The system records the ratings given by the users and analyze then to deduce future likeness for new items. In Implicit User Modeling, the system automatically gathers information about an user's interests and needs.

Time relevance of data collected: - Often the user is searching to quench an immediate information thirst, which will typically last for a short time span, as soon as this information need is done with, the user will generally be not interested in such information any more. To maximize the benefit for the user in such cases through implicit user modeling, propose an eager implicit feedback. That is, as soon as any new evidence has been collected, the systems belief on the information need about the user is updated, and the system then responds with updated model. Such a model however, increases the real time computational requirements and may not be scalable.

Surfing History Data for user profiling may be gathered from the surfing history and surfing behavior of the user. This includes time of visits, last visits, frequency of visits, number of outlinks followed, scrolling patterns, dwelling time, mouse clicks, mouse focuses etc. 'Curious browsers' have been designed to collect such data automatically and implicitly wile the user surfs the net. Human-Computer Interaction systems may be employed to gather more of such data, like eye movements and eye-focus. Snippets gathered from page title, text contents also give valuable clues. Surfing history for an user may also include query history and output URLs selected by the user in the past.

User Bookmarks The bookmarks of a user are a form of explicit feedback by the user and are a very accurate source for gathering feedback.

Client Data The IP address of a user gives us the first clue towards personalization. As soon as we have geographic classification of the user, we can present him with a personalized page that might contain weather information, local news etc. Such is geographic personalization. Even access methods (browser and OS used) could be used to weakly infer some personalization features.

External Client Side Data The data stored in a client's desktop could be used to infer personalization goals. This data may comprise of documents and emails in the harddrives, most frequented softwares etc.

Stage-2: Inferring Profiles

Preprocessing the collected data The collected data has to be cleaned first. The system has to be able to identify the user and track sessions. Log-in forms and cookies are simple enough to identify user.

Profiling Finally we attempt to build the profile. We can employ Machine Learning algorithms, statistical analysis etc. Domain knowledge and ontologies can be incorporated in user inferring profiles. Also, it is necessary to continuously update the user profiling. Some common models are clustering text (clustering is assigning items to groups), classification/decision trees, discovery of association rules, temporal pattern discovery, probabilistic models.

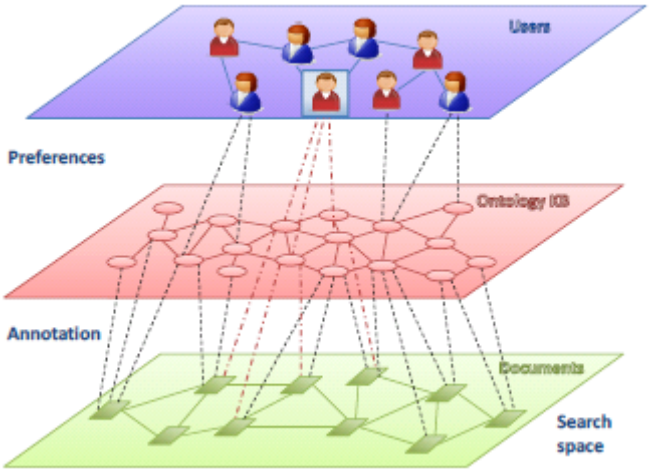


Figure. Link between user preferences and search space

Approaches for inferring profiles may be generalized into two categories. Offline approaches pre-process history information mining relationships between queries and documents visited by users. Online approaches use these data as soon as they are available. Online approaches are more dynamic and updated. But offline approaches can employ more complex algorithms because there is no pressure to meet urgent time restrictions.

Stage-3: Profile Representation

Categories The user profile may be conceptualized as a mapping of queries to categories. The internet directories may be used for a hierarchical categorization of the web. The ODP is the most popular directory, along with a search engine. As we shall see, a lot of systems have been experimented with the ODP as a base reference.

Bag Of Words, List Of URLs In bag-of-words model, the user profile is simply an unordered collection of words. Such models can be built by applying naive Bayes classifiers. However, such a model is a very weak representation of the user. So is the case with list of URLs model. **Fitting in The**

User Modeling Component

In personalized search systems the user modeling component can affect the search in three distinct phases:

Part of retrieval process: The ranking is a unified process wherein user profiles are employed to score Web contents. This is computationally most demanding.

re-ranking: user profiles take part in a second step, after evaluating the corpus ranked via non-personalized scores. This is a client side approach.

Query modification: User profiles affect the submitted representation of the information needs, e.g., query, modifying or augmenting it.

Personalization as a separate ranking factor

The idea is to use the 'distance' from the user profile to the output URL as a separate ranking factor. Thus, a second stage reranking is done on the top n results returned by the search engine. This reranking may be done with the profile as a bag of words as shown. They also show an innovative separation of the "permanent profile" from the current profile. The current profile more effectively reflects the immediate goals of an user session.

Reranking may be done by exploiting the ODP as shown by. The profiles are topics from ODP, and reranking is done as per the conceptual similarity amongst the topics in profile and topics associated with each output URL. The misearch system was developed by Speretta and Gauch. User profiles are represented as weighted concept hierarchies. The ODP (Open Directory Project) is chosen as the base hierarchy. GOOGLE was chosen as the search engine. There is a wrapper built to anonymously monitor activities. Two types of data are collected for each user: the submitted queries for which at least one result was visited, and the snippets, i.e., titles and textual summaries, of the results selected by the user. A classifier trained on the ODPs hierarchy, chooses the concepts most related to the collected information, assigning higher weights to them. After a query is submitted to the wrapper, the snippets are classified into the same reference concept hierarchy.

Outline the Solution

In this project, we studied the problem of personalized search. While there is some existing related work, it is far from optimal. We attempt to address the issue How to personalize more elegant manner.

General Approach:

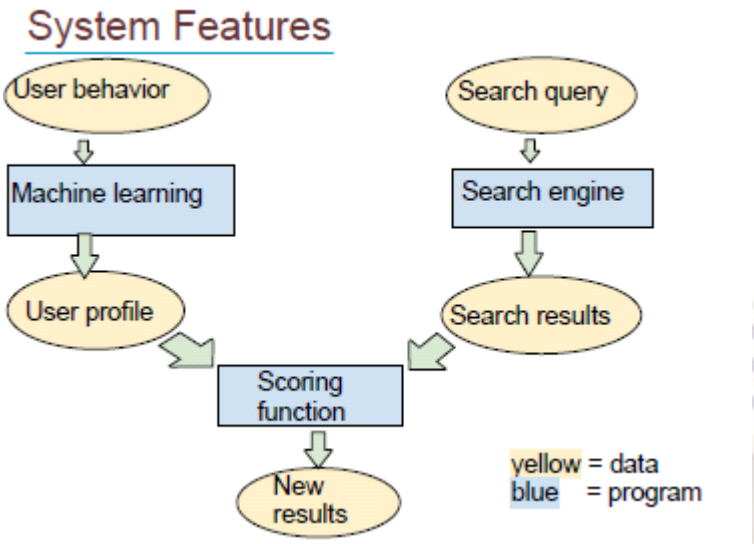


Figure. General approach for this project

The major challenges for personalized search are two fold. The first is, using ontology to identify topics that might be of interest to a specific web user and building appropriate user profile. The second is how to utilize the user profile to improve search accuracy. Another important challenge is evaluation of experiments. There are no standard and bench mark datasets available on which experiments can be performed. This makes comparison with earlier work in the literature and replicating their results difficult. There are also no standard metrics available to effectively evaluate personalized search algorithms. Commonly used metrics in Information Retrieval systems are usually used.

An important requirement for building personalized web search is to build user profiles that represent the users' interests. There are two representations commonly used for user profiles. One is using frequently occurring words in user documents. This creates large profiles where profile terms have low precision and have insufficient context to determine the user interests. The other is using a pre-existing ontology such as DMOZ.

Our User Modeling Approach

Reference ontology and Interest score annotation for the concept. Our approach models the user's profiles by reusing the knowledge available in the domain ontology that is why we named them ontology-user-profiles. Specifically, we propose a semantic model for each user that gives information about: Concepts in the The user context is represented using an ontological user profile, which is an annotated instance of reference ontology. The purpose of using Ontology is to identify topics that might be of interest to a specific Web user. Therefore, we define our ontology as a hierarchy of topics, where the topics are utilized for the classification and categorization of Web pages. The hierarchical relationship among the concepts is taken into consideration for building the ontological user profile as we update the annotations for existing concepts using spreading activation method. It is used to maintain the interest scores based on the user's ongoing behavior.

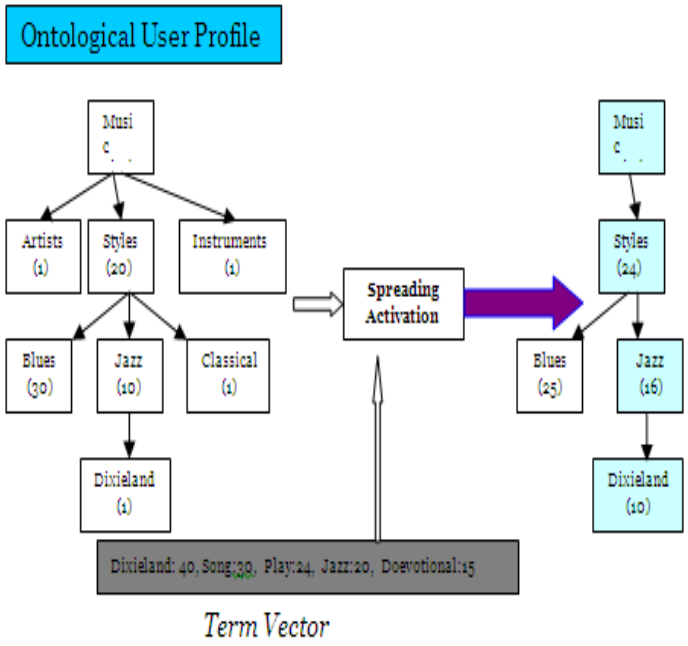
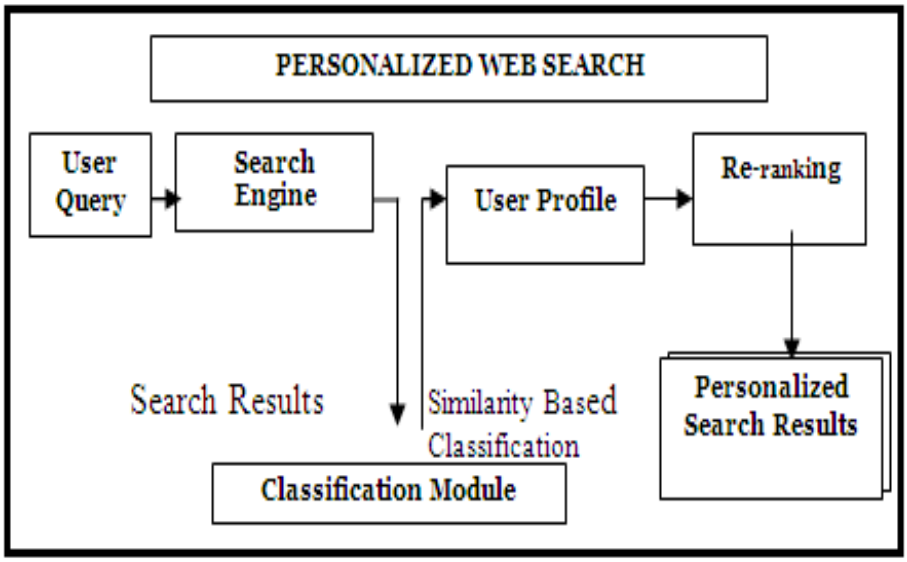


Figure. Portion of Ontological user profile where Interest

Scores are updated on Spreading Activation

The first challenge is building the user profile. Building the user profile itself is a huge research area called construction of User Profiles. In this project, user context refers to the context of past searches made by the user. Many kinds of context information can be potentially exploited. For example, explicit context, implicit context, short term context, long term context etc. Explicit context consists of information given by a user explicitly like one or more documents explicitly marked by a user to be relevant to him. Implicit context refers to any context information naturally available while a user interacts with a retrieval system. For example, if a user clicks a result among the list of results given for a query, the clicked result was probably found to be interesting to the user. While explicit context information is more reliable than implicit context, it is often not available because it requires extra effort from the user. Due to this, implicit context information became an interesting alternative.



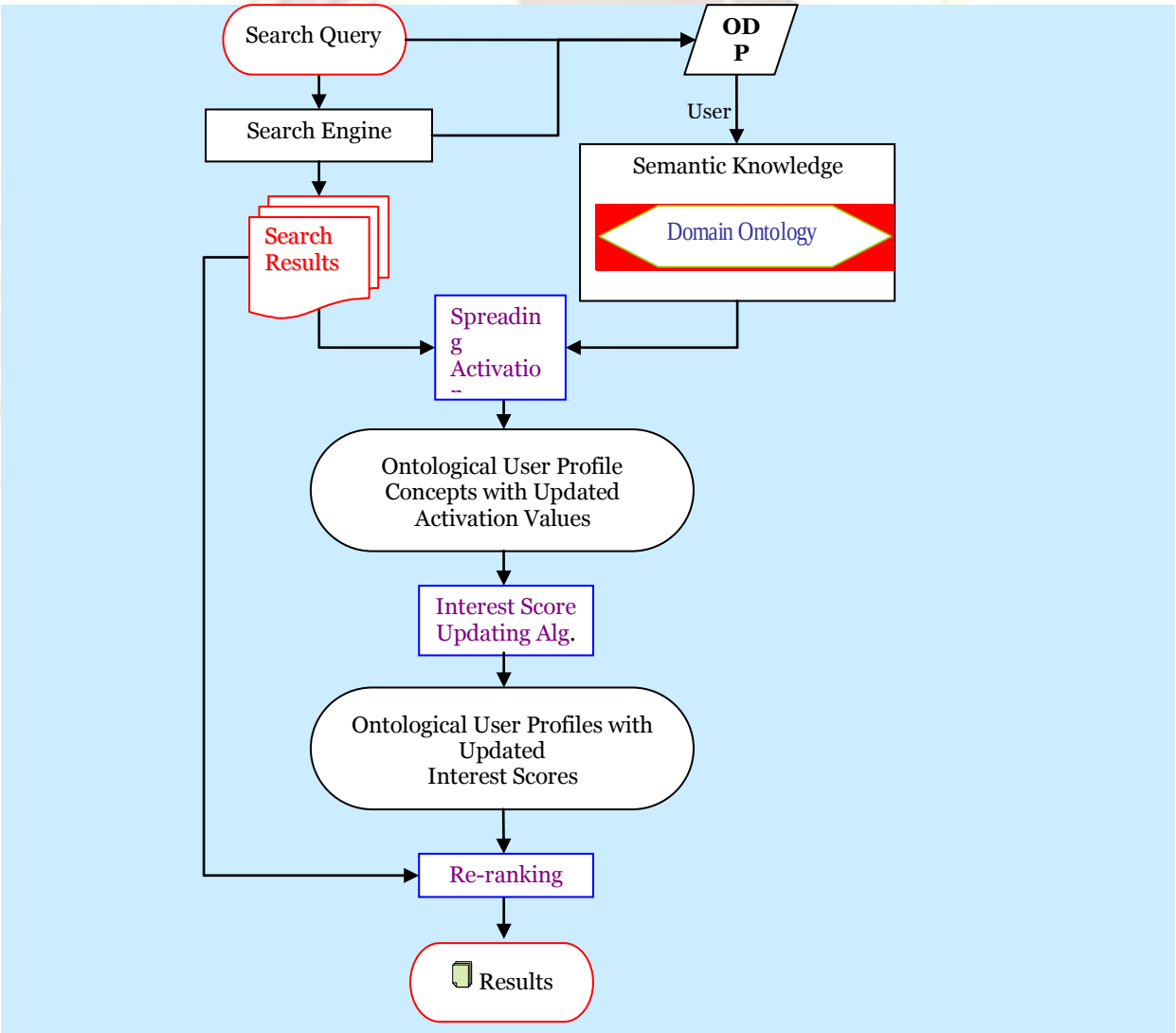


Figure. Diagrammatic overview of this paper.

The second major challenge identified is: re-ranking to improve search results.

The main goal of improving search results is to show more relevant results to the user on the top few results because a user mostly sees only the top few (typically 10) results.

Our approach of re-ranking is as follows: We first retrieve results for a query from a major

search engine and consider the top few results and then compute a score for each document from the top few results based on the user profile of the corresponding user.

Re-rank algorithm is utilized to re-rank the search results based on the interest scores and the semantic evidence in the user profile. A term vector r is computed for each document r € R, where R is the set of search results for a given query. The term weights are obtained using the tf.idf formula described earlier. To calculate the rank score for each document, first the similarity of the document and the query is computed using a cosine similarity measure. Then, we compute the similarity of the. document with each concept in the user profile to identify the best matching concept.

Once the best matching concept is identified, a rank score is assigned to the document by multiplying the interest score for the concept, the similarity of the document to the query and the similarity of the specific concept to the query. If the interest score for the best matching concept is greater than one, it is further boosted by a tuning parameter Once all documents have been processed, Then, we re arrange the documents in descending order of the score with respect to this new rank score.

For re-ranking the results, we followed a common set up. User query is posed to a web search engine (Google) and top few results matching the query are retrieved. These top results are then reranked using the user profile1.

We performed experiments on data extracted from query logs collected over 3 months by a popular search engine. We performed our experiments on 17 users from the query log. The first two months of data is used to learn a profile and the third month data is used for evaluation. Our experiments showed that our approaches outperformed the baseline with a wide margin.
Our study of the same data has lead to some interesting observations closely related to the problem addressed in this project. This has also motivated us to see if clickthrough data can be created in dynamic way by simulating behaviour of users searching a search engine. We developed a basic system which performs the same.

Results and Discussion

This Project focused on building the User profile based on the reranking and assembly of the search results. User's search query is first passed to a search engine such as Google and the resultant URL's are stored as an initial reference results on the Users local machine. The search query along with the URL's from the result is then passed as part of a request to another engine. This engine parses the ODP standard database. The engine takes the URL from the Google results, parses through the ODP standard database xml to identify the relevant categories and the corresponding URL matches. If there is a match of the URL with the ODP content, the corresponding ODP elements and XML structure are then reranked and stored as part of the Output which is the requisite User profile that we build. The data is stored as part of a Collections object within the Engine.

Consistently over several rounds of search query and the corresponding ODP searches resulted in output user profiles of similar nature. The ODP database is not a comprehensive one which would have enabled us to run through multiple levels and search queries for better analysis and evaluation. The sample database was a rudimentary one having a small content for us to the search and parse.

User Profiles that are built form a basis for future search optimization based on the interest levels of the user. The interface that we have built could be bettered through usage of JSP and Servlets. We built the whole interface as a Java command line based interface where higher user interaction and manual work is necessary. This is a limitation in terms of capturing the user clicks and getting to the desired URL and search category.

Parsing technology and the engine have been built using the SAXParser that is available as part of the JDK standard library set. Also, a standard search engine such as google has been used. We have also used the much acclaimed Google Ajax search API for Java to actually interface with Google site and obtain the results. One limitation is the usage of the URL Connection object which actually can have some security implication.

Definite enhancement is envisioned in terms of providing a better user interface, generation of the user profile with the application of the Spreading activation algorithm.

Conclusion and Future work

This Project develops a novel technique for web search personalization, where short-term context is taken into consideration, not only as another source of preference, but as a complement for long-standing user profiles, in order to aid in the selection of the context-relevant preferences that can produce more reliable and "in context" results.

We have presented a framework for contextual information access using ontologies and demonstrated that the semantic knowledge embedded in an ontology combined with long-term user profiles can be used to effectively tailor search results based on users' interests and preferences. The models and methods proposed in the project build upon a user profile representation

based on ontological concepts, which are richer and more precise than classic keyword or taxonomy based

In our future work, we plan to continue evaluating the stability and convergence properties of the ontological profiles as interest scores are updated over consecutive interactions with the system. Since we focus on implicit methods for constructing the user profiles, the profiles need to adapt over time. Our future work will involve designing experiments that will allow us to monitor user profiles over time to ensure the incremental updates to the interest scores accurately reflect changes in user interests.

LIST OF REFERENCES

[1] M. Aktas, M. Nacar, and F. Menczer. Using hyperlink features to personalize web search. In *Advances in Web Mining and Web Usage Analysis, Proceedings of the 6th International Workshop on Knowledge Discovery from the Web, WebKDD 2004*, Seattle, WA, August 2004.

[2] H. Alani, K. O'Hara, and N. Shadbolt. Ontocopi: Methods and tools for identifying communities of practice. In *Proceedings of the IFIP 17th World Computer Congress - TCI 2 Stream on Intelligent Information Processing*, pages 225-236, Deventer, The Netherlands, The Netherlands, 2002.

[3] J. Allan, et al. Challenges in information retrieval and language modeling. *A CM SIGIR Forum*, 37(l):31-47, 2003.

[4] O.Boydell and B. Smyth. Capturing community search expertise for personalized web search using snippet-indexes. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM 2006, pages 277-286, Arlington, VA, November 2006.

[5] H. Chang, D. Cohn, and A. McCallum. Learning to create customized authority lists. In Proceedings of the 7th International Conference on Machine Learning, ICML 2000, pages 127-134, San Francisco, CA, July 2000.

[6] P. Chirita, C. Firan, and W. Nejdl. Summarizing local context to personalize global web search. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM 2006, pages 287-296, Arlington, VA, November 2006.

[7] P. A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschutter. Using odp metadata to personalize search. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2005, pages 178-185, Salvador, Brazil, August 2005.

[8] P. A. Chirita, D. Olmedilla, and W. Nejdl. Pros: A personalized ranking platform for web search. In Proceedings of the 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH 2004, Eindhoven, The Netherlands, August 2004.

[9] S. Dumais, T. Joachims, K. Bharat, and A. Weigend. Implicit measures of user interests and preferences. A CM SIGIR Forum, 37(2), 2003.

[10] S. Gauch, J. Chaffee, and A. Pretschner. Ontology-based personalized search and browsing. Web Intelligence and Agent Systems, 1(3-4), 2003.

[11] T. R,. Gruber. Towards principles for the design of ontologies used for knowledge sharing. In Formal Ontology in Conceptual Analysis and Knowledge Representation, Deventer, The Netherlands, 1993.

[12] H. Haav and T. Lubi. A survey of concept-based information retrieval tools on the web. In 5th East-European Conference, ADBIS 2001, pages 29-41, Vilnius, Lithuania, September 2001.

[13] T. H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th International World Wide Web Conference, WWW 2002*, Honolulu, Hawaii, May 2002.

[14] G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web, WWW 2003*, pages 271-279, Budapest, Hungary, May 2003.

[15] R. Kraft, F. Maghoul, and C. C. Chang. Y!q: contextual search at the point of inspiration. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management, CIKM 2005*, pages 816-823, Bremen, Germany, November 2005.

[16] S. Lawrence. Context in web search. IEEE Data Engineering Bulletin, 23(3):25-32, 2000.

[17] F. Liu, C. Yu, and W. Meng. Personalized web search for improving retrieval effectiveness. IEEE Transactions on Knowledge and Data Engineering, 16(1):28-40, 2004.

[18] A. Micarelli and F. Sciarrone. Anatomy and empirical evaluation of an adaptive web-based information filtering system. User Modeling and User-Adapted Interaction, 14(2-3):159-200, 2004.
[19] S. Middleton, N. Shadbolt, and D. D. R.oure. Capturing interest through inference and visualization: Ontological user profiling in recommender systems. In Proceedings of the International Conference on Knowledge Capture, K-CAP 2003, pages 62-69, Sanibel Island, Florida, October 2003.
[20] M. Porter. An Algorithm for suffix stripping. Program. 14 (3); 130-137, 1980.
[21] F. Qiu and J. Cho. Automatic identification of user interest for personalized search. In Proceedings of the 15th International World Wide Web Conference, WWW 2006, pages 727-736, Edinburgh, Scotland, May 2006.
[22] D. Ravindran and S. Gauch. Exploting hierarchical relationships in conceptual search. In Proceedings of the 13th International Conference on Information and Knowledge Management, ACM CIKM 2004, Washington DC, November 2004.
[23] C. R.ocha, D. Schwabe, and M. P. de Aragao. A hybrid approach for searching in the semantic web. In Proceedings of the 13th international conference on World Wide Web, WWW 2004, pages 374-383, New York, NY, USA, 2004.
[24] G. Salton and C. Buckley. On the use of spreading activation methods in automatic information. In Proceedings of the llth annual international ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 1988, pages 147-160, Grenoble, France, 1988.
[25] G. Salton and M. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, New York, NY, 1983.
[26] B. Schilit and M. Theimer. Disseminating active map information to mobile hosts. IEEE Network, 8(5):22-32, 1994.
[27] X. Shen, B. Tan, and C. Zhai. Ucair: Capturing and exploiting context for personalized search. In Proceedings of the Information Retrieval in Context Workshop, SIGIR IRiX 2005, Salvador, Brazil, August 2005.
[28] A. Sieg, B. Mobasher, S. Lytinen, and R. Burke. Using concept hierarchies to enhance user queries in web-based information retrieval. In Proceedings of the International Conference on Artificial Intelligence and Applications, IASTED 2004, Innsbruck, Austria, February 2004.
[29] A. Singh and K. Nakata. Hierarchical classification of web search results using personalized ontologies. In Proceedings of the 3rd International Conference on Universal Access in Human-Computer Interaction, HCI International 2005, Las Vegas, NV, July 2005.
[30] M. Speretta and S. Gauch. Personalized search based on user search histories. In Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2005, pages 622-628, Compigne, France, September 2005.
[31] A. Spink, H. Ozmutlu, S. Ozmutlu, and B. Jansen. U.S. versus european web searching trends. ACM SIGIR, Forum, 15(2), 2002.
[32] F. Tanudjaja and L. Mui. Persona: A contextualized and personalized web search. In Proceedings of the 35th Annual Hawaii International Conference on System Sciences, HICSS 2002, page 67, Big Island, Hawaii, January 2002.
[33] J. Teevan, S. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR, 2005, pages 449-456, Salvador, Brazil, August 2005.
[34] J. Trajkova and S. Gauch. Improving ontology-based user profiles. In Proceedings of the Recherche d'Information Assiste par Ordinateur, RIAO 2004, pages 380-389, University of Avignon (Vaucluse), France, April 2004.

Authors

C. Srinivas, *M.Tech(CSE), MCA, B.Sc(CSE)* working as a Associate Professor & Head Department of Informstion Technology, Sree Visvesvaraya Institute of Technology & Science (SVITS), Chowderpally (V), Devarkadra (Mdl), Mahabubnagar, A.P-India. Research area includes Artificial Intelligence & Neural Networks, Network Security, Web Mining,