International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
Volume 3 Issue 1 –NOVEMBER 2013.

Personalized Web Search Based On User Histories in Feature Extraction

V.Vijayadeepa [#1], Dr.D.K.Ghosh [*2]

[#] *Head of the Department, Department of Computer Science, Muthayammal College of Arts and Science, Namakka,l TN, India.*

[1] deepkavishka@gmail.com

[*] *Professor, VSB Engineering College, Karur, TN, India.*

Abstract- **In our study, we implemented a wrapper for Google to examine different sources of information on which to base the user profiles: queries and snippets of examined search results. These user profiles were created by classifying the information into concepts from the Open Directory Project concept hierarchy and then used to re-rank the search results. We generate tautological positive and negative rules, based on which we calculate the interest probability to identify the user interest. Rules are generated using open directory project taxonomy and the probability is calculated using the previous history of search queries. We track the implicit behaviors like save, copy, bookmark and also the time spent on viewing the page. The implicit behaviors are used to re-rank the results.**

Index Terms- **User profiles, personalized search, conceptual search.**

I. INTRODUCTION

Personalization has been a very active research field in the last several years and user profile construction is an important component of any personalization system. Explicit customization has been widely used to personalize the look and content of many web sites, personalized search approaches focus on implicitly building and exploiting user profiles. Companies that provide marketing data report that search engines are utilized more and more as referrals to web sites, compared to direct navigation and web links [25] (i.e., StatMarket about WebSideStory product). As search engines perform a larger role in commercial applications, the desire to increase their effectiveness grows. However, search engines are affected by problems such as ambiguity and results ordered by web site popularity rather than user interests. Natural language queries are inherently ambiguous. For example, consider a user issuing the query “canon book”. Due to ambiguity in the query terms, we will obtain results that are either religious or photography related. According to an analysis of 2 months of their log file data conducted by OneStat.com [21], the most common query length submitted to a search engine (32.6 %) is two words and 77.2% of all queries are three words long or less. These short queries are often ambiguous, providing little information to a search engine on which to base its selection of the most relevant Web pages among millions. A user profile that represents the interests of a specific user can be used to supplement queries, narrowing down the number of topics considered when retrieving the results. For the user in our example, if we knew that they had a strong interest in photography but little or none in religion, the photography-related results could be presented to the user preferentially.

Our approach is based on building user profiles based on the user’s interactions with a particular search engine. For this purpose, we implemented GoogleWrapper: a wrapper around the Google search engine, that logs the queries, search results, and clicks on a per user basis. This information was then used to create user profiles and these profiles were used in a controlled study to determine their effectiveness for providing personalized search results.

The study was conducted through three phases:

1. Collecting information from users. All searches, for which at least one of the results was clicked were logged per user.

2. Creation of user profiles. Two different sources of information were identified for this purpose: all queries submitted for which at least one of the results was visited and all snippets visited. Two profiles were created out of both queries and snippets.

3. Evaluation: the profiles created were used to calculate a new rank of results browsed by users. The average of this rank was compared with Google’s rank.

Many approaches create user profiles by capturing browsing histories through proxy servers or desktop activities through the installation of bots on a personal computer. These require participation of the user to install the proxy server or the bot. In this study, we explore the use of a less-invasive means of gathering user information for personalized search. Our goal is to show that user profiles can be implicitly created out of short phrases such as queries and snippets collected by the search engine itself. We demonstrate that profiles created from this information can be used to identify, and promote, relevant results for individual users.

II. BACKGROUND

2.1 Ontologies and Semantic Web

According to Gruber [11], an ontology is a “specification of a conceptualization”. Ontologies can be defined in different ways but they all represent taxonomy of concepts along with the relations between them. In the context of the World Wide Web, ontologies are important because they formally define terms shared between any type of agents without ambiguity, allowing information to be processed automatically and accurately. OntoSeek [12] is an example of system based on ontologies. Utilizing information sources such as product catalogs and yellow pages it applies conceptual graphs to represent both queries and resources.

The expression “Semantic Web” was introduced by ETAI (Electronic Transactions on Artificial Intelligence) in 2000 to describe the extension of the Web to deal with the meaning of available content rather than just its syntactic form. Many XML based projects such as Resource Descriptor Framework (RDF), Notation 3 (N3), and OWL started from there and each aims to define syntax capable of describing and/or manipulating ontologies. One of the main bottlenecks in the evolution of the Web along these lines is the amount of manual effort usually required to create, maintain, and use ontologies. Our approach shares many of the same goals as the Semantic Web, however we focus on automatic techniques wherever possible.

2.2 Personalization

Personalization is the process of presenting the right information to the right user at the right moment. In order to learn about a user, systems must collect information about them, analyze the information, and store the results of the analysis in a user profile. Information can be collected from users in two ways: explicitly, for example asking for feedback such as preferences or ratings; and implicitly, for example observing user behaviors such as the time spent reading an online document. Explicit construction of user profiles has several drawbacks. The user provide inconsistent or incorrect information, the profile built is static whereas the user’s interests may change over time, and the construction of the profile places a burden on the user that they may not wish to accept. Thus, many research efforts are underway to implicitly create accurate user profiles [6][7][22]. User browsing histories are the most frequently used source of information about user interests. Trajkova and Gauch [26] use this information to create user profiles represented as weighted concept hierarchies. The user profiles are created by classifying the collected Web pages with respect to a reference ontology. Kim and Chan [15] also build user profiles from the same source, however they use clustering to create a user interest hierarchy. The collected Web pages are then assigned to the appropriate cluster. The fact that a user has visited a page is an indication of user interest in that page’s content. Extending this idea, Chan describes a metric to estimate the level of user interest; for example the percentage of links visited on a page or URL presented in bookmarks.

To achieve effective personalization, profiles should distinguish between long-term and short-term interests and include a model of the user’s context, i.e., the task in which the user is currently engaged and the environment in which they are situated [19]. Several systems have attempted to provide personalized search that are tailored based upon user profiles that capture one or more of these aspects.

In the OBIWAN project [8], search results from a conventional search engine are classified with respect to a reference ontology based upon the snippets summarizing the retrieved documents. Documents are re-ranked based upon how well their concepts match those that appear highly weighted in the user profile. PERSIVAL [18] is a system that provides personalized search on specific medical libraries. Rather than building a user profile, PERSIVAL allows users to augment queries by providing contextual information such as a patient record. PERSIVAL then extracts concepts from the context and uses them to expand the query. The patient record is also used to filter the search results, removing information that is not related to the specific case described in the context. They have extended their personalized search to also be applied to multimedia information.

Competitive Intelligence Spider and Meta Spider [5] are part of a client-based application that collects and organizes Web documents on the user’s machine. Spiders may gather information directly from Web sites or through search engines. Collected documents are then analyzed and noun phrases are extracted to create a personal dictionary for the user to guide future searches. The noun phrases are also used to organize the documents and a graphical map of the results is generated. Users can personalize the search explicitly by selecting specific Web sites, the number of Web pages to collect, and the noun phrases used in the final map of results.

The Personal Search Assistant [14] is an application that a background process that collects information on behalf of a user by submitting queries to various search engines. Results are stored on the local machine and are analyzed so that they can be organized conceptually. The user manually creates a conceptual database that is input to a personal agent responsible for building a user profile. The profile is used to filter the results of later searches.

III. METHODS

Feature Ex traction

The process starts with feature extraction ,at this stage all the search queries submitted by the user from the browsing history is extracted. From the search queries stemming process is applied. Pure terms are extracted as only nouns from the search queries.

```
Algorithm: Feature Extraction
Given: Set of User Queries Qi
Procedure:
step1: Remove stop words from Qi.
step2: Tag the word using pos tagger.
step3: Identify nouns.
step4: return index terms IW(Iw1,Iw2,…).
step6: End
```

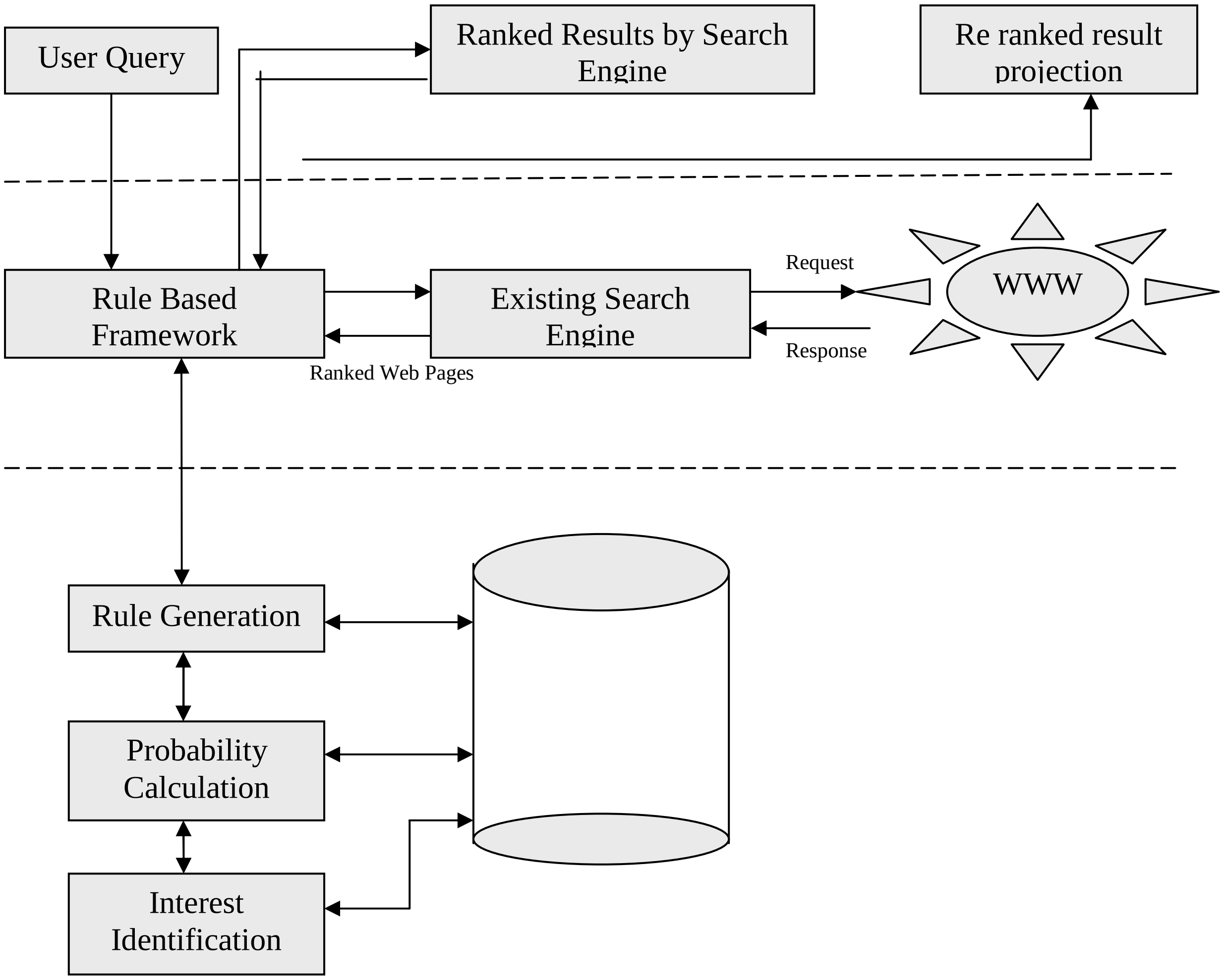

Figure 1: System Architecture

Rule Generation
Ranked Web Pages
Rule Generation
Probability Calculation
Interest Identification
User Query
Ranked Results by Search Engine
Re ranked result projection
Rule Based Framework
Existing Search Engine
WWW
Request
Response

At this stage of the process the odp taxonomy is read into the system. The read text is splitted into set of terms. From the set of terms the root concept is identified. With the identified root concept , positive and negative rules are generated against index terms. Generated rules are indexed for later usage.

IV. CONCLUSION

With the semi-structure of information on the Internet and the arbitrariness of releasing the enormous amount of web pages, turns finding desired information quickly and exactly to be a crucial task. Search engine is playing an increasing important role in information retrieval on the Internet. The search results given by search engines are generally sorted on descendent importance of its usage. Humans think in terms of concepts but the concept may be differing from one another. Hence the importance of a page is gained from users with different Concepts. Thus this contradictory importance does not be feasible in future. Hence user-centric personalization is the only solution to solve the problem. We further investigate this framework to increase the relevancy of the web links to the search query.

V. REFERENCES

[1] Daniel Fogares and Balazs Racz, Practical Algorithms and Lower Massive Bounds for Similarity Search in Massive Graphs, IEEE transactions on knowledge and Data Engineering,vol 19,No5,pages 585-598, May 2007.

[2] Ioannis Anagnostopulos,Ilias Maglogiannis, Adapting user’s Browsing and Web Evolution features for Search in medical portals, First IEEE International Workshop on semantic media adoption and personalization, pages 37-42, 2006.

[3] Liang Deng, Martin D. F. Wong, An Exact Algorithm for the Statistical Shortest Path Problem, ACM conference on Asia South Pacific design automation, pages 965-970, 2006.

[4] S.Sendhilkumar and T.V. Geetha, An Evaluation of Personalized Web Search for Individual User, International Conference on Artificial Intelligence and Pattern Recognition (AIPR07), FL, USA, pages 484 - 490, 2007.

[5] Orland Hoeber and Xue Dong Yang, ExploringWeb Search Results Using Coordinated Views, Fourth IEEE International Conference on Coordinated & Multiple Views in Exploratory Visualization, pages 3-13, 2006.

[6] O. Hoeber and X.D. Yang,The visual exploration of web search results using HotMap, International Conference on Information Visualization, pages157-165, 2006.

[7] Wenxue Tao, Wanli Zuo, QuerySensitive SelfAdaptable Web Page Ranking Algorithm, Second International Conference on Machine Learning and Cybernetics, Xi, pages 413-418, 2003.

[8] Rahul R. Joshi and Y. Alp Aslandogan, Conceptbased Web Search using Domain Prediction and Parallel Query Expansion, IEEE International Conference on Information Reuse and Integration, Waikoloa, Hawaii, pages16-18, 2006.

[9] A. Broder, “A Taxonomy of Web Search,” ACM SIGIR Forum, vol. 36, no. 2, pp. 3-10, 2002.

[10] U. Lee, Z. Liu, and J. Cho, “Automatic Identification of User Goals in Web Search,” Proc. 14th Int’l World Wide Web Conf. (WWW ’05), pp. 391-400, 2005.

[11] B.J. Jansen, A. Spink, J. Bateman, and T. Saracevic, “Real Life Information Retrieval: A Study of User Queries on the Web,” ACM SIGIR Forum, vol. 32, no. 1, pp. 5-17, 1998

[12] K. Sugiyama, K. Hatano, and M. Yoshikawa, “AdaptiveWebSearch Based on User Profile Constructed without Any Effort from Users,” Proc. 13th Int’l World Wide Web Conf. (WWW ’04), pp. 675-684, 2004.

[13] F. Liu, C. Yu, and W. Meng, “Personalized Web Search for Improving Retrieval Effectiveness,” IEEE Trans. Knowledge and Data Eng., vol. 16, no. 1, pp. 28-40, Jan. 2004.

[14] M. Speretta and S. Gauch, “Personalized Search Based on User Search Histories,” Proc.IEEE/WIC/ ACM Int’l Conf. Web Intelligence (WI ’05), pp. 622-628, 2005.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank Citation Ranking: Bringing Order to the Web,” technical report, Computer Science Dept., Stanford Univ., 1998.

[16] T.H. Haveliwala, “Topic-Sensitive Pagerank,” Proc. 11th Int’l World Wide Web Conf. (WWW), 2002.

[17] Y. Zhou and W.B. Croft, “Query Performance Prediction in Web Search Environments,” Proc. 30th Ann. Int’l ACM SIGIR Conf.Research and Development in Information Retrieval (SIGIR ’07), pp. 543-550, 2007.

[18] P.A. Chirita, C. Firan, and W. Nejdl, “Summarizing Local Context to Personalize Global Web Search,” Proc. ACM Int’l Conf. Information and Knowledge Management (CIKM), 2006.

[19] J. Chaffee and S. Gauch, “Personal Ontologies for Web Navigation,” Proc. ACM Int’l Conf. Information and Knowledge Management (CIKM ’00), pp. 227-234, 2000.

[20] S. Gauch, J. Chaffee, and A. Pretschner, “Ontology-Based Personalized Search and Browsing,” Web Intelligence and Agent Systems, vol. 1, no. 3/4, pp. 219-234, 2003.

[21] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, “Rank Aggregation Methods for the Web,” Proc. 10th Int’l World Wide Web Conf., pp. 613-622, 2001.

[22] P.W. Foltz and S.T. Dumais, “Personalized Information Delivery: An Analysis of Information Filtering Methods,” Comm. ACM, vol. 35, no. 12, pp. 51-60, 1992.

[23] W.B. Frakes and R. Baeza-Yates, “ Information Retrieval: Data Structures and Algorithms”. Prentice Hall, 1992.

[24] N. Fuhr, “A Decision-Theoretic Approach to Database Selection in Networked IR,” ACM Trans. Information Systems (TOIS), vol. 17, no. 3, pp. 229-249, 1999.

[25] S. Gauch, G. Wang, and M. Gomez, “ProFusion: Intelligent Fusion from Multiple, Distributed Search Engines,” J. Universal Computer Science, vol. 2, no. 9, pp. 637-649, 1996.

[26] E.J. Glover, G.W. Flake, S. Lawrence, W.P. Birmingham, A. Kruger, C.L. Giles, and D.M. Pennock, “Improving Category Specific Web Search by Learning Query Modifications,” SAINT,pp. 23-34, 2001.

[27] G.H. Golub and C.F. Van Loan, “Matrix Computations”, third ed. 1996.

[28] L. Gravano and H. Garcia-Molina, “Generalizing GlOSS to Vector-Space Databases and Broker Hierarchies,” Proc. 21st Int’l Conf. Very Large Databases (VLDB), pp. 78-89, 1995. [29] D.A. Grossman and O. Frieder, “Information Retrieval: Algorithms and Heuristics”. 1998.

V.Vijayadeepa received her B.Sc degree in computer science discipline from university of Madras and M.Sc degree in the same discipline from Periyar University,Salem She has also completed her M.Phil in computer science discipline at Bharathidasan University. She is having 10 years of experience in collegiate teaching and She is the head of computer Science and applications department in Muthayammal college of Arts and Science affiliated by Periyar University. Her main research interests include personalized Web search, Web information retrieval, data mining, and information systems.

Dr D.K.Ghosh got his B.Sc degree from Calcutta University and M.Sc degree from IIT, Kharagpur.He has received his doctorate from IIT, Kharagpur. He is working as a professor in V.S.B Engineering College, Karur, India. He is very much interested in reading academic books and research papers all the time. Moreover, He is the research guide of Anna University, Chennai.