Providing Privacy in Personalized Web Search

Sarika N. Zaware, Deepali Shah, Taahir Hamgi, Payal Khinvasara, Rasika Tribhuwan

Department of Computer Engineering, AISSMS IOIT, Pune

Abstract - Personalized web search is a promising way to improve search quality by customizing search results for people with individual information goals. The main aim is to provide results to users according to his interest.Unnecessary exposure will be avoided and relevant results will be provided to the user. For the the first time when the user fires the query , the results provided by the search engine and the proposed system is same if the interest specified by the user at the time of registration is not related to the query. But when a similar query is fired again, more relevant results will be displayed to user using vector space algorithm and ODP, that means the links that user visits earlier will be at the top of the webpage and rest will be listed below .Thus Personalized web search has been effectively carried out. It is good to use Vector Space Model and open directory project (ODP) to obtain personalized results. This paper describes the design and implementation of this system by using user's information and search behaviour.

Keywords: — Privacy providing, personalized search, Vector Space Model, ODP, user profile.

I. INTRODUCTION

As the amount of information on the web continuously increases, it has become difficult for web search engines to find information that satisfies users individual needs. Personalized search is a way to provide personalized web search results to the people in less time. One distinct approach is Re-ranking query results returned by search engines locally using personal information and another is by user's browsing history such as list of visited URLs, the number of visits to each, list of past search queries and pages clicked for these search queries .This browsing history is collected and stored and is visible to users in online as well as offline mode so that user can have immediately come to know his most favourite links. Even privacy is maintained by login mechanism as each user can view its own history and not of others, only after login to the application. This mechanism prevents from revealing likes, hobbies, and interest and search behaviour of other users. Thus personalized web search aims at providing privacy to the user along with prioritized results.

II. MOTIVATION

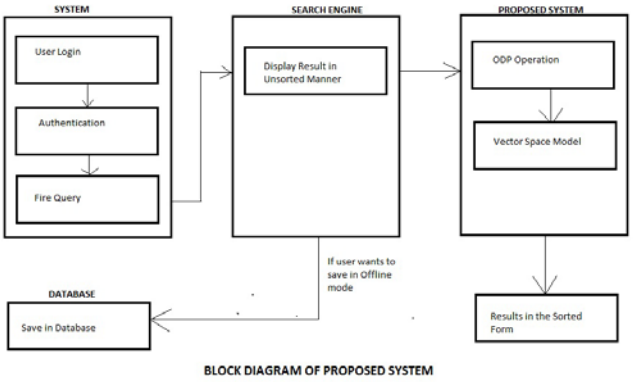
As more and more users are on the web, it is increasingly difficult to let the search engine know what we want. Re-ranking ambiguous queries has long been an important part in the research of Information Retrieval, but still remains to be a challenging task. Personalized search has done significant efforts to take over this challenge in the web search services, based on the fact that a user's general preference may help the search engine disambiguate the true intention of a query. However, studies have shown that users are reluctant to provide any explicit input on their personal preference. Even M. Richardson and P. Domingos did research related to re-ranking the pages according to queries fired by user. In this paper, we extend this research work to depict how a search engine can learn a user's preference automatically based on her past click history and how it can use the user preference to personalize search results. Our experiments show that users interest can be known accurately even from small click-history data and personalized search based on user preference will yield significant improvements over the years.

III. RELATED WORK

According to the research papers, there has been considerable research in personalized web search. Various algorithms have been used to re-rank the results according to user interest. Scalability and performance issues of personalized web search have been studied in these papers. A framework is being provided in these papers so that the bias vector space can be used to improve performance. Other than the scalability studies, [12] based on query terms (but not by individual users). In [15] with the help of user profiles explicitly specified by the users, personalized page ranks are computed. Researchers have also proposed ways to personalize web search based on ideas other than PageRank [16, 17, 18]. [17] Explores ways to consider the topic category of a page during ranking using user-specified topics of interest. [18] Does an analysis on the correlation between users, their queries and search results clicked to model user preferences. We believe this method is difficult to scale to general search engines due to difficulty in analysis, In [19], for example, multiple TF-IDF vectors computed, each value represents the user's interests in one area. In [20] pages visited by the user is categorized by their similarities and then compared to a set of pre-categorized pages, and user preferences are shown by the topic categories of pages in her browsing history. In [21] the user's preferences are learned from both pages she visited and those visited by users similar to her (collaborative filtering). Our work is slightly different from these studies, in that personalized results are obtained to

user's based on its interest using open directory project and vector space algorithm.

IV. PROPOSED METHODOLOGY



BLOCK DIAGRAM OF PROPOSED SYSTEM

The Proposed system mainly consists of eight modules having the following functionalities which help us in getting crisper and prioritized results based on user interests and behaviour.

Module 1 - User Login
1. Accept user id and password.
2. If new user then registers along with its interest else login directly.

Module 2 – Authentication
Verifying login id and password.

Module 3 - Fire Query
Accepting query from user & passing to the search engine.

Module 4 – Search Engine
Displaying results according to the fired query.

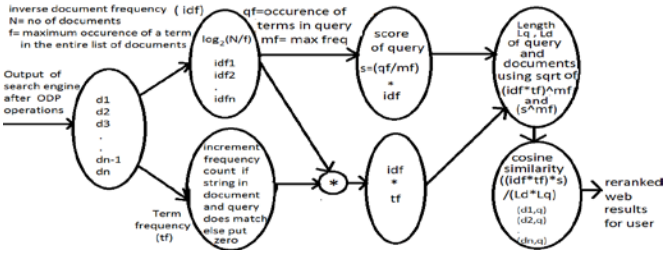
Module 5 – ODP
Open directory project for creating dynamic user interest tree(hierarchical model).

Module 6 – Vector Space Model
For re-ranking results obtained from search engine based on user interest and behaviour.

Module 7 – Results in Sorted Form
To display relevant results with higher priority.

Module 8 – Offline Mode
Storing user specified web links in our database for offline use

Following is the illustration of the Vector Space Model used in our system.



This is a simplified example of the vector space retrieval model.

Query fired through the application to the search engine is:-

Q: "what are apple company products?"

These are the five documents obtained as a result of firing the query to the search engine

D1: "apple is a fruit"

D2: "apple fruit is red"

D3: "it is good for health"

D4: "apple is a multinational company"

D5: "iphone is apple company product"

The grammatical connectives are truncated in query as well as the documents as shown below:

Q: "apple company products"

D1: "apple fruit"

D2: "apple fruit red"

D3: "good health"

D4: "apple multinational company"

D5: "iphone apple company product"

The truncated documents will be re-ranked using the following steps in vector space algorithm:-

Step1: we calculate the idf(inverse document frequency) values for the terms in the document. From these idf values we come to know the number of times a term occurs in entire list of documents.

Step 2: we calculate the tf (term frequency)scores for all the terms in the document . From these tf values the occurrence of a term in each document is known.

Step 3:now we multiply the tf scores by the idf values of each term, obtaining the following matrix of documents-by-terms:

Step 4:we calculate the score for the truncated query in order to find how many times the terms in the query match with those in the document .

Step 5: we calculate the length of each document and of the query:

Step 6: then the cosine similarity values are computed to rank the documents based on their maximum match with the terms in the query .

Then according to highest value of the similarity values, the documents are re-ranked in their descending order

Thus the final order in which the documents are presented as result of the query will be: d5, d4, d1, d2, d3.

D5: “iphone is apple company product”

D4: “apple is a multinational company”

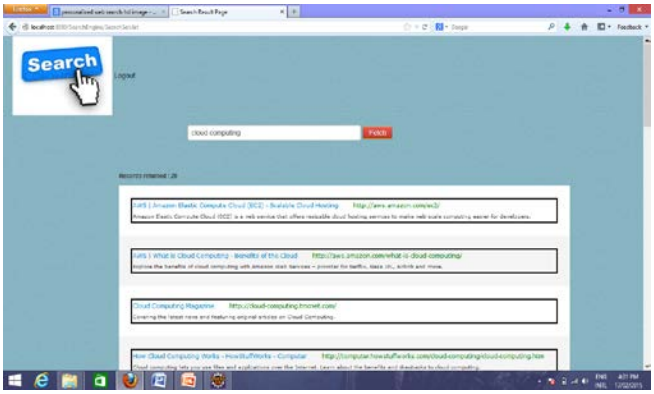
D1: “apple is a fruit”

D2: “apple fruit is red”

D3: “it is good for health”

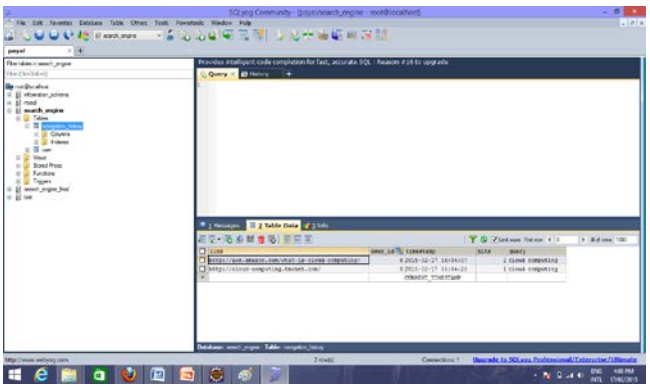
V. Simulation/Experimental Result

When the query “Cloud Computing” is fired for the first time. The results displayed in our system are in the same order as that of which are provided by our search engine i.e Google.



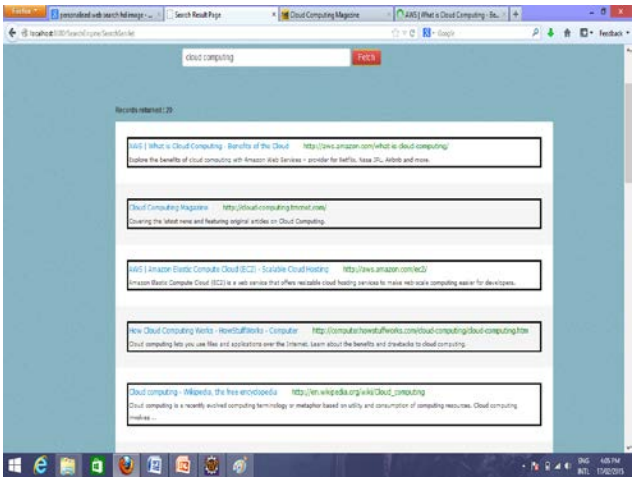
When the user clicks on several links, the pages visited by him are being prioritized depending upon the number of times he has visited those pages.

Below we can see our database where the links and the number of times they have been hit by the user is displayed in the backend.



As we can see the Amazon website has been viewed twice by the user, and the tmcnet website has been viewed only once. Thus, the amazon link will be prioritized and will be displayed at the top, and the tmcnet website will be displayed at the second position. Following will be the other results provided by the search engine.

We can see the re-ranked results below.



VI. Conclusion

The use of Internet in the recent years is growing rapidly which makes the need of a technique which can give accurate and relevant results to the user. Even though there are many search engines currently present, it has been observed that they fails to capture user’s preference and behaviour and hence the search results may or may not provide relevant results to the user. In this paper, hence we proposed a possible technique such as vector space algorithm and open directory project which can give users an experience of

personalized web search and ultimately users can get what they want in a crisp manner in shorter time and fewer clicks as well.

VI. FUTURE SCOPES

In future, image or logo can be added along with the query so that it will guide the user to get the desired information in a very specific context with comparably less effort.

REFERENCES

[1] B. J. Jansen, A. Spink, and T Saracevic. Real life, realusers, and real needs: A study and analysis of userqueries on the Web. Information Processing andManagement, 36(2):207 – 227, 2000.

[2] Robert Krovetz and W. Bruce Croft. Lexicalambiguity and information retrieval. InformationSystems, 10(2):115–141, 1992.

[3] Nielsen netratings search engine ratings report.http://searchenginewatch.com/reports/article.php/2156461,2003.

[4] J. Carroll and M. Rosson. The paradox of the activeuser. Interfacing Thought: Cognitive Aspects ofHuman-Computer Interaction, 1987.

[5] T. Haveliwala. Topic-sensitive pagerank. InProceedings of the Eleventh Int'l World Wide WebConf., 2002.

[6] S. Brin and L. Page. The anatomy of a large-scalehypertextual Web search engine. In Proc. of WWW'98, 1998.

[7] F. Qiu and J. Cho. Automatic identification of userpreferences for personalized search. Technical report,UCLA technical report, 2005.

[8] J. Cho and S. Roy. Impact of Web search engines onpage popularity. In Proc. of WWW '04, 2004.

[9] Dimitri P. Bertsekas and John N. Tsitsiklis.Introduction to Probability. Athena Scientific, 2002.

[10] M. Kendall and J. Gibbons. Rank CorrelationMethods. Edward Arnold, London, 1990.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd. Thepagerank citation ranking: Bringing order to the web.Technical report, Stanford Digital LibraryTechnologies Project, 1998.

[12] M. Richardson and P. Domingos. The IntelligentSurfer: Probabilistic Combination of Link and ContentInformation in PageRank. In Advances in NeuralInformation Processing Systems 14. MIT Press, 2002.

[13] G. Jeh and J. Widom. Scaling personalized websearch. In Proceedings of the Twelfth Int'l World WideWeb Conf., 2003.

[14] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub.Exploiting the block structure of the web forcomputing pagerank. Technical report, StanfordUniversity, 2003.

[15] M. Aktas, M. Nacar, and F. Menczer. Personalizingpagerank based on domain profiles. In Proc. ofWebKDD 2004: KDD Workshop on Web Mining andWeb Usage Analysis, 2004.

[16] F. Tanudjaja and L. Mui. Persona: A contextualizedand personalized web search. In Proc. of the 35thAnnual Hawaii International Conference on SystemSciences, 2002.

[17] P. Chirita, W. Nejdl, R. Paiu, and C. Kohlschuetter.Using odp metadata to personalize search. InProceedings of ACM SIGIR '05, 2005.

[18] J. Sun, H. Zeng, H. Liu, Y. Lu, and Z. Chen.Cubesvd: A novel approach to personalized websearch. In Proceedings of the Fourteenth Int'l WorldWide Web Conf., 2005.

[19] L. Chen and K. Sycara. Webmate: a personal agentfor browsing and searching. Proc. 2nd Intl. Conf. onAutonomous Agents and Multiagent Systems, pages132–139, 1998.

[20] A. Pretschner and S. Gauch. Ontology basedpersonalized search. In ICTAI, pages 391–398, 1999.

[21] K. Sugiyama, K. Hatano, , and M. Yoshikawa.Adaptive web search based on user profile constructedwithout any effort from users. In Proceedings of theThirteenth Int'l World Wide Web Conf., 2004.

[22] Supporting privacy protection in personalized web search, Lidan Shou, He Bai, Ke Chan, And Gang Chen, 2014.