# Web Search Personalization: A Fuzzy Adaptive Approach

Mohammad. S Norouzzadeh, Ayoub Bagheri, Mohammad. H Saraee
Data Mining Lab, Department of Electrical and Computer Engineering
Isfahan University of Technology，Isfahan, Iran
m.norouzzadeh@ec.iut.ac.ir, a.bagheri@ec.iut.ac.ir, saraee@cc.iut.ac.ir

***Abstract*—Today the growing rate of web data has become so large and this is the reason for turning search engines into the major decision support systems for the Internet. In this paper, a novel and simple approach is proposed to improve Web search. The approach is a client-side method towards personalization of web search which adapts the results based on user interests. In addition we define a fuzzy variable to clarify the relevance of each document. In real world, notion of relevance is a fuzzy concept and certainly the relevancy ratios of some relevant documents are not equal. By applying our approach to the Google's datasets the result shows that this approach can improve performance in search.**

***Keywords- Web Search; Search Personalization; Fuzzy Theory; Adaptive Search***

## I. INTRODUCTION

The Internet Global Network currently consists of over than 63 trillion web pages across more than 80 million web sites and also continues to expand at an exponential rate. Under these circumstances searching a typical phrase on search engines is resulted over hundreds of pages and no one can check all of the results. Therefore search engines must continue to improve and refine their technology in order to remain useful.

On the other hand there are over than 1.4 trillion regular users with different interests and various degrees of computer literacy, language skills and a wide range of preferences. Retrieving the exact information for a particular user is one of the most challenging information retrieval tasks.

This process can be very difficult since a typical Internet user is unlikely to provide a well-defined search query to a search engine.

For example many search queries contain only two or three general and vague search keyword, and the inevitable outcome is that the search engine returns many irrelevant search results. Jansen et al [1] showed that 62% of queries submitted to the Excite web search engine contained only one or two terms. Such short queries can left out very useful search terms, which may decrease effectiveness of the search.

Traditional link-analysis search strategies such as PAGERANK [2-4] and HITS [5] are not capable to infer user's interests from these general queries and intelligent methods for inferring user's interests may be very costly for search engines [6]. These search engines that use typical information retrieval methods cannot provide precise results to the users. Regardless of who submitted the query, a typical search engine returns the same result when the same query is submitted by different users. This may not be suitable for users with different information needs. For example, if a query "ant colony" is issued to Google, about 631,000 results are returned. Since "ant colony" may refer to "ant colony optimization" or "ant swarm", two users with different interests may want the search results ranked differently: a biologist may expect biological relevant pages ranked highly; however, these pages may be unnecessary to be displayed for a computer scientist.

Web search personalization has been emerged as a potential solution to the mentioned problem. Web search personalization systems use gathered information from user such as profiles, cookies and … to conduct and revise search in order to maximize user satisfactions [7]. Personalization techniques have four common approaches.

i. Filtering methods are trying to filter typical search results based on gathered information such as [8].

ii. Augmenting methods may add some results to typical search results based on context such as [9].

iii. Re-ranking methods are trying to re-rank typical search results based on user's information such as [10].

iv. Blending methods combine former methods for personalization such as [11].

Since web expands on daily bases, search personalization becomes a necessity rather than a choice. Some search engines have offered the personalized search service including Google's Personalized Search which allows users to specify the Web page categories of interest [12].

In this paper we introduce a novel and robust method for web search personalization. In our approach we use fuzzy theory in order to model the natural uncertainty of user queries, user preferences and notion of relevance. Followed by reasons to prove our method is useful.

The remainder of the paper is structured as follows. Section II reviews related work, Section III presents the proposed approach for personalization, and finally we

**978-1-4244-4520-2/09/$25.00 ©2009 IEEE**

evaluate the method and conclude in Section IV and Section V respectively.

## II. RELATED WORKS

The idea of a personalized web search has been studied by various researchers with different ideas and many approaches have been proposed. Profiling and collaborative filtering are two types based on user's participation. In profiling, search conducted by the information collected from user which referred as 'profile'. In profiling strategies search results are personalized based on day relevance to user profile. Profiling strategy has three steps and each of these steps can be implement by different techniques. Steps of the profiling strategies are i) Input Data Selection, ii) Profile Construction, and iii) Profile Representation. In addition input data source includes three types which can be utilized as input data to build user profiles. These three types are i) server side data, ii) client side data and iii) proxies' data. The server side data are collected at the search server(s). Various types of server side data such as: Last visit and No. of visits [13] have been suggested for personalization. Data sources such as Bookmarks [14], Client Data [13], previous search queries [15-16] are client side data which have been suggested for personalization. In addition proxies' data such as collection of network level information [13] can be used as well.

In Profile Construction step, methods such as Text clustering, Classification [17-18], Discovery of Association Rules [19] and Temporal Pattern Discovery [19] have been utilized. Constructed Profile can be represented as Categories [19], List of URLs or Bag of words [13].

As mentioned above collected profiles can be used in three ways: query enhancement, result filtering and Separate Ranking Factor. In the first method the query is compared against user preferences. If the similarity between the query and user preferences is above a threshold the query is augmented with metadata and submitted to the search engine to obtain more precise results. In second method results are pruned based on user profile. Finally in last method the results that are returned by a search engine are re-ranked based on the user's profile. Here, we focus on personalizing search hits through result processing.

There are many methods that can be classified as profiling or collaborative filtering technique. Some of these methods generate biased PAGERANK vector and use them on PAGERANK algorithms. PAGERANK is the most efficient link analysis algorithm, used after user query for ranking the results returned by standard retrieval methods. The ranking is performed by evaluating the importance of a page in terms of its links from other important pages. In order to enhance the acquired results, many variations of this algorithm have been proposed. These approaches, use the so called "personalization vector" of PAGERANK in order to bias the results toward the user needs. Few examples of these methods are Topic-Sensitive PAGERANK [20], Modular PAGERANK [21] and BLOCKRANK [22].

Another type of methods in search personalization is collaborative filtering or social filtering. Collaborative filtering is any algorithm that filters information or patterns for a user based on a collection of users, agents, viewpoints, etc. The main idea behind collaborative filtering is making automatic predictions (filtering) about the interests of a user by collecting taste information from similar users (collaborating). Collaborative filtering is Similar to giving out recommendations to a friend. Collaborative filtering technique widely used in recommender systems.

The first system to use collaborative filtering was the Information Tapestry project at Xerox PARC [23]. This system allowed users to find documents based on previous comments by other users.

The basic mechanism behind collaborative filtering systems is as follow:

- A large group of people's preferences are registered;
- Using a similarity metric, a subgroup of people is selected whose preferences are similar to the preferences of the person who seeks advice;
- An (possibly weighted) average of the preferences for that subgroup is calculated;
- The resulting preference function is used to recommend or filter options on which the user has expressed no personal opinion as yet.

Collaborative filtering methods can be combined with other methods such as PAGERANK personalization methods [20].

In addition there exist some newer and different approaches such as search and ranking based on the PI, personalized search as an application of LSA and Exploiting Personal Data are reported in [24-26].

## III. PROPOSED APPROACH

In this paper we present a novel and robust approach for web search personalization. The proposed approach is a personalized adaptive search that uses fuzzy theory to improve results of query [9]. In our methodology adaptive means adjusting search results by user's feedback. This strategy has some advantages. One of these advantages is that there is no need to store any information or profile from users. Therefore in adaptive search all information that need would be collected at query time. Another benefit of this strategy is that users not have to fill profile forms and log in during search. This benefit makes search engines more user-friendly. The third advantage of the adaptive strategy for search is storing profile temporarily which no needs to any effort from part of user for changing preferences.

On the other hand, the adaptive web search personalization has some disadvantages. Redundant computation for each session is one of the disadvantages of this approach. Also in adaptive personalized search some valuable information maybe lost, this is because of users not filling any profile forms.

As mentioned before, we need user's feedbacks to conduct an adaptive search [27-28]. User's feedbacks can be considered in two types including explicit or implicit. In explicit feedback, Users must indicate relevance explicitly using a binary or graded relevance system. Binary relevance feedback indicates that a document is either relevant or irrelevant for a given query. Graded relevance feedback indicates the relevance of a document to a query on a scale using numbers, letters, or descriptions (such as "not relevant", "somewhat relevant", "relevant", or "very relevant"). In implicit feedback, feedback is inferred from user behavior [29], such as determining which documents they do and do not select for viewing, the duration of time spent viewing a document [19], or page browsing or scrolling actions [19]. The proposed approach uses implicit feedbacks model to capture user's interests.

For personalization of search results Rocchio introduces a relevance feedback framework [17]. In Rocchio formulation each query has a query vector which is initially constructed from the query terms. This query vector is enhanced with user's feedback. That is,

$$\overline{q1} = \alpha\overline{q0} + \beta \sum_{\text{known relevant}} \frac{D_j}{|D_j|} - \gamma \sum_{\text{known non-relevant}} \frac{D_j}{|D_j|} \tag{1}$$

In this equation $\overrightarrow{q0}$ is the initial query vector and $\overrightarrow{q1}$ is the reformulated query vector. $D_j$ represents document's abstract vectors, and $|D_j|$ is the length of vectors. Where *α*, β and γ are set experimentally and control the query vector. The range of these parameters is restricted from 0 to 1.

In real world, notion of relevance is a fuzzy concept. It means if we have some relevant documents, their relevancies are not equal. Similarly the relevance degrees of non relevant documents are different.

To make easy and quick personalization of search results we introduce a formula based on the Rocchio formulation, which we adjusted it to consider membership grade of relevance for each document. The Rocchio formulation edits the query vector to retrieve relevant documents [30], but it cannot prevent some irrelevant results for the search. Just like Rocchio we associate with each query a query vector which is initially constructed from the query expressions. Subsequently this query is adjusted to distinguish relevant documents from irrelevant ones. Thus we introduce our method in the following manner:

$$\overline{q1} = \alpha\overline{q0} + \beta \sum_{\text{all documents}} \eta_j \frac{D_j}{|D_j|} \tag{2}$$

In this equation $\overrightarrow{q0}$ and $\overrightarrow{q1}$ are the initial query vector and the reformulated query vector respectively. Similarly to previous equation $D_j$ and $|D_j|$ represent document's abstract vector and the length of vectors correspondingly. Where *α* and *β* are set empirically and control the query vector. The range of these parameters is restricted from 0 to 1. The parameter *h* is the fuzzy variable which is differentiates each documents from another and also its range is restricted from 0 to 1.

In this work we utilize a fuzzy variable for relevance of each document from search results. Each document can be labeled "not relevant", "not many relevant", "somewhat relevant", "very relevant" or etc. To measure the ratio of document relevance – value of fuzzy variable – we used the abstract of each document which the search engines are returned, and the behavior of users such as noting which documents they do and do not select for viewing. The method we utilize for this purpose arises from fuzzy theory.

Based on fuzzy theory [31] each element in fuzzy set has a membership grade, which is defined by a membership function $\mu_A(x)$. In personalization of search, fuzzy set determines the relevance of the document. To calculate the relevance of each document we use abstract's terms and consider all abstracts relatively. Certainly each abstract has some common terms with the query vector, so we count the number of these terms. The ratio of document relevance – value of fuzzy variable – is calculated by the number of common terms with the query vector. For this purpose we use *Cosine similarity* formula as follows,

$$\mathrm{Sim}(q,p) = \frac{\sum_{k\in q\cap p} f_{kq} f_{kp}}{\sqrt{\left(\sum_{k\in p} f_{kp}^2\right)\left(\sum_{k\in q} f_{kq}^2\right)}} \tag{3}$$

Where *q* is the query vector, *p* is the abstract of page, and $f_{kd}$ is the frequency of term *k* in *d*. The *Cosine similarity* is a standard measure of similarity to estimate relevance in information retrieval and in most search engines.

If document is not relevant then $\mu_A(x) = 0$, If it is fully relevant then $\mu_A(x) = 1$, and if it is partially relevant then $0 < \mu_A(x) < 1$ as in Fig. 1.

In our method $\mu_A(x)$ must be replaced by the relevance function. The relevance function calculates by the abstract of each search results. Also in relevance function each of the documents weighted. Each document weighted with respect of the users open them for viewing or not. According to (2), regarding to the relevance degree of a result, the key terms of its abstract can be appended to query vector. Therefore from (3), all pages which selected for viewing (relevant), value of their fuzzy variables approach to one. The value of fuzzy variables for all pages that do not selected would be less than viewed pages.

For example if a user searches a query “ant colony”, search engine returns some pages as result. Assume that user is looking for “ant colony optimization” technique; therefore she selects those pages that contain relevant terms like “Optimization” among results. As noted above our method assigns more weights to those pages which contain interest key terms to the user and vice versa. The query vector will be augmented with key terms of relevant pages. Therefore the query vector for this example completed with “optimization” term.

The proposed method in this paper for personalization of web search has some noticeable advantages over previous approaches. As mentioned before, our approach is based on the client side computation. This makes the system quick and provides privacy for users. In addition using abstracts of the documents reduces the computation time than utilizing full documents. If the hardware and software tools would advance in future we can use full documents with the same speed. In addition, using explicit user’s feedback would be more helpful and progressive to the approach. Finally, the proposed method can be easily combined with other approaches to make better results for the users.

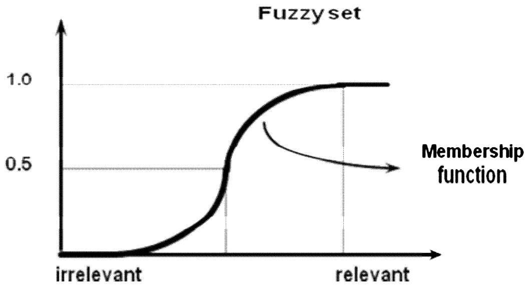

Figure 1. Typical fuzzy membership function of relevance

## IV. EVALUATION

To evaluate effectiveness of our proposed method with respect to our primary goal, learning the user’s topic preference vector from her click stream and using this vector to personalize search ranking, we may consider one of the following evaluation metrics.[30]

Let us suppose that we have a query $q$ and some description $r_q$ of a subset of the relevant pages. In our method $r_q$ is the abstract of relevant pages prepared by search engine. We call $C_q$ the set of all pages obtained from the search engine. We can then define an estimated precision,

$$P(C_q) = \frac{1}{|C_q|}\sum_{p \in C_q} Sim(r_q, p) \tag{4}$$

Where, $Sim(r_q, p)$ is the cosine similarity function defined in (3).

We can further define an estimated recall as:

$$R(C_q) = P(C_q)|C_q| = \sum_{p \in C_q} Sim(r_q, p) \tag{5}$$

That intuitively approximates the recall measurement, which is obtained by multiplying precision by the size of all retrieved set.

As mentioned in previous section in our method fuzzy weights of relevant documents will be increased and according to (4) and (5), we can intuitively see that if some terms have been added to query vector the size of result set, $C_q$, will be decreased or in worst case will not change. On the other hand based on the (3), since we follow user interests, results that are not so relevant will be removed and Cosine similarities improve. Therefore it is clear that, these two measurements would be improved for pages which have been selected by user.

For example if a user searches the phrase “ant colony” in some search engine and then selects results that contain “ant colony optimization”, According to (2) weight of “optimization” term in query would be increased. Thus “optimization” term will be added to query and results are confined and *Cosine similarities* improve, this means that $|c_q|$ getting smaller and summation of similarities higher. Therefore, based on these improvements and (4), the estimated precision will be increased.

Since we know there is trade-off between estimated precision and recall, we cannot improve both factors simultaneously for our approach. In above example we saw the improvement for two factors, but always it is not true. For example in some cases one or more terms in query vector may be removed and result set will be expanded. In such cases using estimated recall could be better measure for examining performance.

To evaluate the performance of our method practically, we have used two datasets each containing 98 documents. These documents have obtained form Google search engine. To do this test let us suppose a user wants to get some documents about “Ant Colony Optimization”. But when she searches documents in Google search engine, she only uses two terms including “Ant” and “Colony”. Therefore with respect to the user’s goal, she will select documents which contain the term “Ant Colony Optimization” through the results of search engine. The proposed method gets the first query term and user feedback and then produce the refined query term for better resulting in search.

The Google search engine for the query term “Ant Colony” returns 10 results in the first page. Three of these ten documents are relative to the “Ant Colony Optimization” and certainly are those which the user selects for viewing. Based on the proposed method in (2), the first query term refines with abstract and title of the results. In

(2), the fuzzy parameter $h$ obtains from (3). The $a$ and $b$ parameters have valued 0.5 and 0.5 by trial and error respectively. Output of this equation is a vector that contains words and their frequency. This vector filters by a threshold that shows the frequency of a word. In this test we used the value 1 for the threshold. The new obtained query term contains "Optimization" term in addition to terms in the first query. Therefore results get update by using the new query.

The table 1 shows the Precision and Recall for the initial and the refined query terms. These values are obtained based on the first 98 results in Google search engine for each query from (4) and (5) respectively. The results have gotten from the first 10 pages in the output of Google search engine.

Table 1: Comparing Precision and Recall

| | **Initial query "Ant Colony"** | **Refined query "Ant Colony Optimization"** |
|---|---|---|
| **Precision** | 44.92 | 58.17 |
| **Recall** | 44.02 | 57.01 |

Based on table 1, the proposed approach improves the performance of Google's method and modifies the results with respect to the user demand and her interests. Hence we can conclude the proposed method presents good performance and can overcomes the search methods in famous and popular search engines.

## V. CONCLUSION AND FUTURE WORKS

In this paper we proposed a fuzzy adaptive approach for web search personalization. In our method we adjust search results by user's feedback. There is no need to store any information from users and they do not have to fill profile forms and log in when they want to search. The proposed approach improves the performance of Google's method, modifies the results with respect to the user demand and presents good performance. There are a number of interesting directions to investigate for future works. First we can use some extra information about user such as location information and profile forms [10, 32]. Next, the proposed method can be combined with other methods such as collaborative filtering and Diversification [33]. Also we can examine other fuzzy membership functions for improving our approach.

## REFRENCES

[1] Jansen, B.J., Spink, A. and Saracevic, T. Real life, real users, and real needs: A study and analysis of users on the web, Information Processing & Management 36. 2. (2000) 207-227.

[2] Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks 30 (1998) 107–117.

[3] Pavel Berkhin,P: Survey: A Survey on PageRank Computing. Internet Mathematics 2(1): (2005)

[4] Amy N. Langville and Carl D. Meyer. Deeper Inside PageRank. Internet Mathematics, Vol. 1(3): , (2005) 335-380.

[5] Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46 (1999) 604–632.

[6] B. Mobasher, S. Anand, A. Kobsa, and D. Jannach (Eds.), AAAI Press Technical Report WS-07-08, PP. 17-26, July 2007.

[7] Feng Qiu, Junghoo Cho: Automatic identification of user interest for personalized search. WWW 2006: 727-736.

[8] F. Liu et Al.: Pers. Search for Improving Retrieval Effectiveness. TKDE 2004, IEEE Transactions on Knowledge and Data Engineering Volume 16 , Issue 1 (January 2004) (2004) 28 - 40 ISSN:1041-4347.

[9] Shady Elbassuoni, Julia Luxenburger, Gerhard Weikum "Adaptive Personalization of Web Search" Proceedings of the 1st Workshop on Web Information Seeking and Interaction, Amsterdam, The Netherlands, SIGIR 2007.

[10] Kazunari Sugiyama, Kenji Hatano, Masatoshi Yoshikawa, Adaptive Web Search Based on User Profile Constructed without Any Effort from Users, Proceedings of the 13th international conference on World Wide Web New York, NY, USA 675 - 684 (2004)

[11] Rohini U and Vamshi Ambati: A Collaborative Filtering based Re-ranking Strategy for search in Digital Libraries. In the proceedings of 8th ICADL - 2005, Bangkok.

[12] Google personalized search,
http://labs.google.com/personalized

[13] D. Pierrakos, G. Paliouras, C. Papatheodorou, C.D. Spyropoulos, "KOINOTITES: A Web Usage Mining Tool for Personalization", Proceedings of the Panhellenic Conference on Human Computer Interaction, Patras, December 2001.

[14] Ben Markines and Lubomira Stoilova and Filippo Menczer, Bookmark Hierarchies and Collaborative Recommendation Proceedings of the 3rd international workshop on Link discovery, Chicago, Illinois
66 - 73 (2005) ISBN:1-59593-215-1

[15] Fang Liu, Clement Yu, Weiyi Meng, Personalized Web Search for Improving Retrieval Effectiveness, IEEE Transactions on Knowledge and Data Engineering 16(2004) 28-40.

[16] Mehmet S. Aktas, Mehmet A. Nacar, Filippo Menczer: Using Hyperlink Features to Personalize Web Search. WebKDD 2004: 104-115.

[17] Peter Jackson, Isabelle Moulinier, Natural Language Processing for Online Applications, John Benjamins Publishing Company Amsterdam / Philadelphia, 2002.

[18] 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.

[19] Soft Computing for Knowledge Discovery and Data Mining Maimon, Oded; Rokach, Lior (Eds.) 2008, XIV, 434 p. 74 illus., Hardcover Springer Science and Business Media, Inc.

[20] Taher Haveliwala. "Topic-Sensitive PageRank," Proceedings of the Eleventh International World Wide Web Conference, May 2002

[21] G. Jeh and J. Widom, Scaling personalized web search, Proceedings of the 12th international conference on World Wide Web table of contents Budapest, Hungary
271 - 279 (2003) ISBN:1-58113-680-3

[22] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub, Exploiting the block structure of the web for computing PageRank, Stanford University Technical Report, 2003

[23] David Goldberg, David Nichols, Brain Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. Comm of the ACM, 35(12):61–70, 1992.

[24] Jian-tao Sun, Yuchang Lu, Cubesvd: A novel approach to personalized web search In Proc. of the 14 th International World Wide Web Conference 2005.

[25] J. Teevan, S. T. Dumais & E. Horvitz (2005). Personalizing search via automated analysis of interests and activities. SIGIR 2005.

[26] Dae-Young Choi, Enhancing the power of Web search engines by means of fuzzy query, Decision Support Systems Volume 35 , Issue 1 (April 2003) Web retrieval and mining
31 - 44 (2003)

[27] Gerard Salton and Chris Buckley, Improving Retrieval Performance by Relevance Feedback 355 - 364 (1997) Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

[28] Ryen W. White, Ian Ruthven, Joemon M. Jose The use of implicit evidence for relevance feedback in web retrieval, Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval 93 - 109 (2002)

[29] Maria Fasli, Udo Kruschwitz, using implicit relevance feedback in a web search assist, Proceedings of the First Asia-Pacific Conference on Web Intelligence: Research and Development
356 - 360 (2001) ISBN:3-540-42730-9

[30] Filippo Menczer, Complementing search engines with online web mining agents Decision Support Systems Volume 35 , Issue 2 (May 2003) Special issue: Web data mining 195 - 212 (2003)

[31] Roger jang, chuen sun, eiji mizutani, neuro-fuzzy and soft computing, ISBN 13: 9780132610667,prentice-hall, Inc. 1996.

[32] Dae-Young Choi, Personalized local internet in the location-based mobile web search Decision Support Systems Volume 43 , Issue 1 (February 2007) 31-45 (2007)

[33] Filip Radlinski, Susan Dumais, Improving Personalized Web Search using Result Diversification , Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval Seattle, Washington, USA 691 - 692 (2006)