Journal of Computer Science 8 (7): 1029-1035, 2012
ISSN 1549-3636

A Novel Page Ranking Algorithm for a Personalized Web Search

[1]Jayanthi, J. and [2]K.S. Jayakumar
[1]Department of Computer Science and Engineering,
Sona College of Technology, Salem-5, Tamil Nadu, India
[2]Department of Mechanical Engineering,
SSN College of Engineering, Chennai, Tamil Nadu, India

Abstract: Problem statement: Information on the web is growing exponentially. Today, traditional search engines provide results mainly based on the user's query. Though the context of the query varies, the returned result seems to be same for all users. Accordingly users are expected to search for the relevant results, which is an added overhead to the users. **Approach:** We propose a Personalized Preference Network based Web Search Ranking (PPN based WSR) framework that uses Personalized Page Ranking (PPR) algorithm for re-ranking the search results. **Results:** Our methodology aims to compute the User Interest Score (UIS) over the search results. **Conclusion:** The proposed method can yield preferred results since it considers both the User Interest Score and Term Frequency and Inverse Document Frequency (TF-IDF) for re-ranking.

Key words: Personalized Page Ranking (PPR), User Interest Score (UIS), Term Frequency and Inverse Document Frequency (TF-IDF), User Interest Hierarchy (UIH)

INTRODUCTION

The impressive growth in the amount of information on the internet has attracted a huge variety of users towards it. Search engines present a well organized way to search the relevant information from the web. However, the search results acquired might not always be helpful to the users, as search engine fails to recognize the user intention behind the query.

A particular query could mean different things in varying context and the anticipated context can be interpreted by the user alone. For illustration, the specified query "skate", a user might be searching about the glide on ice or for a kind of fish. Traditional search engines provide similar set of results without considering the intention behind the query. Thus, in spite of recent development on web search technologies there are still many conditions in which search engine users are not satisfied with the search results. Therefore, the requirement arises to have personalized web search system which gives an output appropriate to the users as highly ranked pages. A personalized web search has various levels of efficiency for different users, queries and search contexts. A personalized web search has various levels of efficiency for different users, queries and search contexts.

Related work: Search Engine return results based on simple keyword matches without any concern for the information needs of the user. Ramadhan *et al*. (2006) proposed a heuristic based solution to differentiate the significance of various backlinks by assigning a different weight factor to them depending on their location in the directory tree of the Web space. This Rank computation completely relies on the link structure of a web page and hence it fails to consider the user's interest.

Web systems utilize the User Relevance Feedback (Algarni *et al*., 2010) to interpret the user's information needs. The vector space model computes the similarity between the query and the document and is based on the terminological overlap between them. Relevance Feedback requires the user to classify the documents into relevant and irrelevant groups. Rocchio algorithm is used to expand the queries from the feedback thus obtained. Users are generally reluctant to provide information on whether they are interested in a particular document or not, so relevance feedback is not satisfying mechanism to fulfill the user needs.

Web personalization could be achieved by organizing the user profile as User Interest Hierarchy (UIH) (Kim and Chan, 2005). UIH tracks the user interest implicitly and DHC algorithm is used for the

Corresponding Author: Jayanthi, J., Department of Computer Science and Engineering, Sona College of Technology Salem-5, Tamil Nadu, India

same in order to classify the results. Different characteristics of a term are derived and accordingly the terms are scored. This approach does not present any consideration for merging the current term which is similar to the existing term in the hierarchy. UIH could be refined by specifying two new characteristics namely term and node specificity (Hu and Chan, 2008). Using these features the top results can be re-ranked. But the same approach fails to handle some new queries that are provided by users.

News search is personalized (Dali *et al*., 2010) in some news portals by using demographic information. The results are re-ranked based on the information that is fetched during registration of the users. Zhuang and Cucerzan (2006), Q-Rank is used to refine the ranking of the search results by constructing the query context from search query logs. Definitions of the query context are extracted from the query logs in order to extract the context of the new query. Using the extracted context the results are re-ranked. Page rank vectors (Aktas *et al*., 2004) are personalized by weighting the links based on the match between hyperlinks and user profiles. User specified interests are organized as binary vectors where each feature corresponds to a set of one or more DNS tree nodes. Topic-Sensitive Page Rank (Haveliwala, 2002) scores are computed using the topic in the context in which the query appeared. Multiple importance scores for each page with respect to various topics are captured and at query time these importance scores are combined to form the composite PR scores using that the results are ranked.

Historical query logs are learned and from which the results are optimized so that user intended pages are ranked higher. Queries from the logs are clustered using the similarity function (Shanna *et al*., 2010) and the sequential patterns from the selected web pages are captured and based on the patterns the results are re-ranked. Similarly the frequent phrases from the past queries are obtained using frequency meaning based algorithm (Barouni-Ebrahimi *et al*., 2008) and accordingly the appropriate results are re-ranked. User behaviors are modeled (Agichtein *et al*., 2006) and by learning those models the preferred results for the users are predicted. User behavior beyond click through are modeled so that the re-ranking thus obtained is far better than the one that is obtained by considering only click through methods. The user profile (Bhowmick *et al*., 2010; Brin and Page, 1998) is constructed based on many data sources and framework uses three types of monitors. Various types of ontology and their relationship is discussed.

Kavita and Gawali (2010) and Ratnakumar (2008), various web mining techniques are widely used for search result personalization. A weighted URL ranking algorithm is used to rank the web search results based on the features extracted from hyperlinks, anchor terms and user interested domains. The retrieved results from the search engines are weighed according to the occurrence of tokens and are again weighed in accordance with the user interested domain and the same are retained for re-ordering the results according to the match with the query weight. For personalization (Teevan *et al*., 2005) some client side algorithms are developed. The different algorithms (Kumar and Singh, 2010) used for link analysis like Page Rank (PR), Weighted Page Rank (WPR) and Hyperlink-Induced Topic Search (HITS) algorithms are discussed and compared.

A classic algorithm such as Hub Finder algorithm (Paul-Alexandru *et al*., 2004) is used to find the related pages and the result is used to provide a platform for personalized ranking. This algorithm uses the user's bookmarks as input and the hubs with higher page rank are filtered for further processing. Thus the technique contributes for personalized ranking. Harb *et al*. (2009), a personal search engine is designed which provides relevant results according to user's interests. Three factors contributing to accurate retrieval of results are important of document category, user interest and the degree of relevance of the document.

Based on the click history (Qui and Cho, 2006) the user model is developed where the representation of user preference is given based on the topic and page.

MATERIALS AND METHODS

Proposed work: We propose a method to re-rank the search results by considering the user interest over the search results that are returned by the traditional search engines. The architecture of the proposed system is illustrated in Fig. 1.

The proposed preference network based page ranking algorithm includes the following functionalities to extract the relevant result for personalized search:

- A set of documents that matches the user query is fetched from the search engine (top K documents)
- The terms in the initial set of documents are weighed using TF-IDF measure and by using the same the user preferred network of concepts is framed
- The network is tracked for UIS and the proposed feature weights are computed
- The result set is ranked based on computed UIS and TF-IDF value

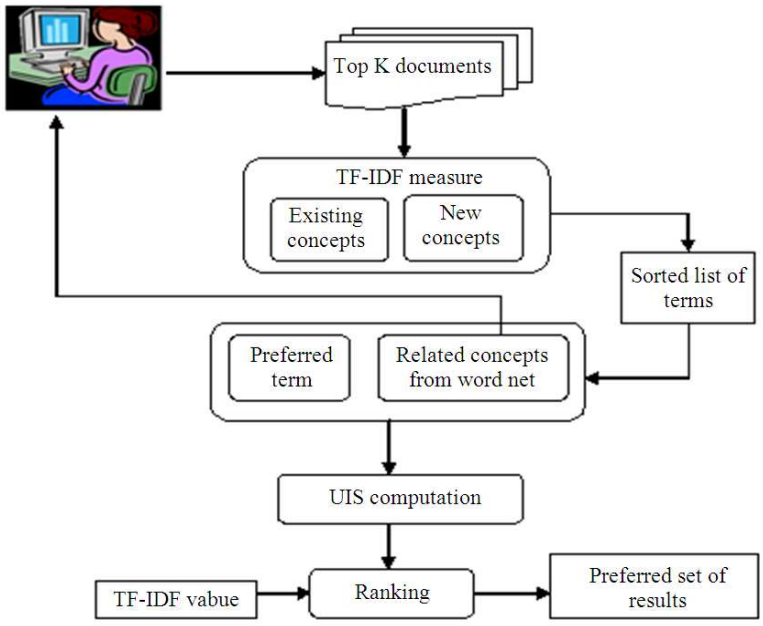

Fig. 1: System architecture

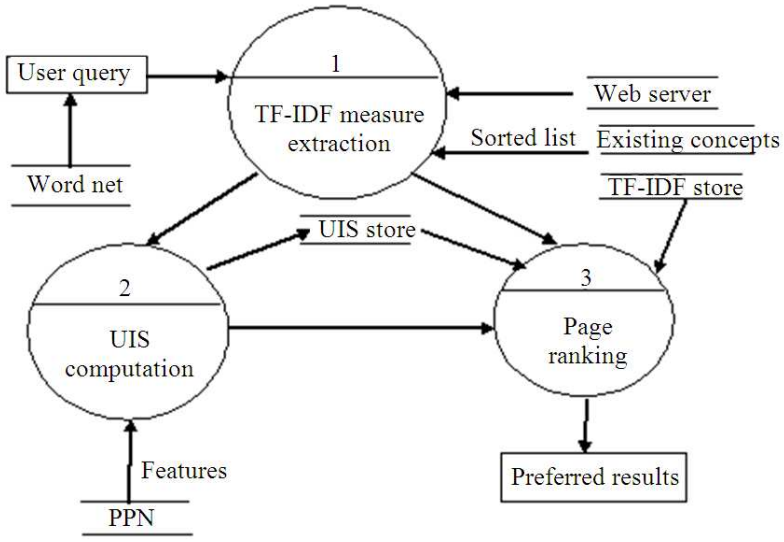

Fig. 2: Process flow outline

Method: The proposed system proceeds through the below processes namely:

- TF-IDF Measure Extraction
- UIS computation
- Page Ranking

The proposed Personalized Preference Network based Web Search Ranking Framework process the query for authenticated users and provides the personalized or preferred results by weighting the relevant results in accordance with user's interest. When the user issues the query the search engine retrieves the set of results. From the results retrieved top K documents are selected and it serves as the initial input to the PPN based WSR framework. The proposed framework is realized through three different processes and the data flow could be interpreted using Fig. 2.

TF-IDF measure extraction: The top K documents from the web server are analyzed for each term TF-IDF measure is computed and the same could be retained in the TF-IDF store. Terms are sorted based on the TF-IDF value measured and from this the top N terms with higher weights are used for further processing. From the above term sheet, the identical terms in all documents are collected and their weights are added up and from the outcome the higher weighted terms are again selected for building the personalized preference network.

Term frequency and Inverse document frequency can be obtained as below Eq. 1:

$$tf_i = \frac{n_i}{\sum kn_k} \quad (1)$$

Where:
n_i = No of occurrence of a term i
nk = Total no of terms in a document Eq. 2:

$$idf_i = \log \frac{N}{df_i} \quad (2)$$

Where:
N = Total number of documents that are relevant
df_i = Number of documents that contain the term i at least once Eq. 3:

$$\text{TF-IDF weight} = tf_i * idf_i \quad (3)$$

Thus the term frequency and inverse document frequency are computed.

UIS computation: User Interest Score is computed by considering the various features through which the individual's interest can be tracked. Features are extracted from the PPN and the same are weighted to obtain the UIS.

Features to be considered are:

- Frequency of usage
- Link Access pattern
- Time spent over a concept
- Usage count

User's weight over a concept could be rendered using the top three features and the last feature renders the concept's weight.

The proposed mathematical model computes the UIS and the definitions incorporated are as follows:

User Set U = $\{U_i\}$ where i={1,2,3,……,n}

Concept Set C= $\{C_j\}$ where j={1,2,3,……,m}:

C_{ij} = $\{ F_{ij}, A_{ij}, T_{ij}, UC_{ji} \}$
F_{ij} = {V}
A_{ij} = {N}
T_{ij} = {P}
C_{ij} = represents the j^{th} concept for i^{th} user
F_{ij} = represents the frequency of usage of j^{th} concept by i^{th} user
A_{ij} = represents the access pattern of j^{th} concept by i^{th} user
T_{ij} = represents the time spent over the j^{th} concept by i^{th} user
UC_{ji} = represents the usage count of j^{th} concept by all users

Frequency of usage calculates how frequently an individual views a particular concept. Frequently used concept with respect to particular user over a fixed span is computed and it gains the maximum weight among other concepts Eq. 4:

$$F_{ij} = \frac{v_R(C_j)}{\sum_s v(C)} \quad (4)$$

where, $V_R(C)$ corresponds to repeated visits and $\sum V(C)$ corresponds to total number of visits of all concepts over a session.

Link Access Pattern illustrates the navigation pattern of a single user in association with a specified query. Depth of access for a particular concept with respect to particular user over a fixed span is computed Eq. 5:

$$A_{ij} = \frac{N_v(C_j)}{\sum N(C_j)} \quad (5)$$

where, $N_V(C_j)$ corresponds to the number of nodes visited and $\sum N(C_j)$ corresponds to the total number of nodes.

Time spent over a concept depicts how long a particular concept is viewed by the individual under study. It is obtained by computing the percentage of scroll Eq. 6:

$$T_{ij} = \frac{P_s(C_j)}{\sum P(C_j)} \quad (6)$$

where, $P_S(C_j)$ corresponds to the number of pages scrolled and $\sum P(C_j)$ corresponds to the total number of pages.

Usage count depicts how wide a concept is viewed by various users. This in turn extracts the concept popularity Eq. 7:

$$UC_{ij} = \sum U_i (C_j) \quad (7)$$

where, $\sum U_i (C_j)$ corresponds to the number of users of a concept C_j.

Using the above proposed computation, the higher weighted concept from each user's perspective is obtained. From the higher weighted concept, the weights of the remaining concepts are also calculated relative. Relative weight is interpreted as below Eq. 8:

$$Wt[Feature(C)] = \frac{Max[Wt(Feature) \times Feature(C)]}{Max(Feature)} \quad (8)$$

Once the features are weighed, the user's interest score of all concepts can be derived using the proposed scoring function Eq. 9:

$$UIS = \sum_{i=1}^{n} \sum_{j=1}^{m} [F_{ij} + A_{ij} + T_{ij} + UC_{ij}]$$
$$= \sum_{i=1}^{n} \{ \sum_{j=1}^{m} \left[\left(\frac{V_j}{\sum_s v(C)} \right) + \left(\frac{N_v(C_j)}{\sum N(C_j)} \right) + \left(\frac{P_s(C_j)}{\sum P(C_j)} \right) + \sum U_i (C_j) \right] \} \quad (9)$$

The above suggested formula calculates the UIS for the maximum weighted concept. Likewise, the same could be derived from all the remaining concepts that are relatively weighed.

Page ranking: The rank of the relevant results is computed in accordance with the user interest. The ranking of a result considers both TF-IDF measure and user interest score. Personalized page rank is computed as Eq. 10:

$$PPR = 0.55 * (UIS) + 0.45 * (TF\text{-}IDF) \quad (10)$$

While computing the rank, the weight of the UIS and TF-IDF are varied according to the nature of the query and the user.

RESULTS AND DISCUSSION

In result analysis, specified query is considered and accordingly the preferred network with respect to single user could be computed as below:

- User issues the query "Web Mining" and the results are retrieved by the traditional search engine
- Initially, the user selected documents say {d1, d2, d3, d6, d7, d8}, from the retrieved results are retained for analysis
- From the retained document set, keywords are extracted to construct the preferred network

Using the preferred network in Fig. 3, the page rank of the results could be computed as illustrated in Table 1.

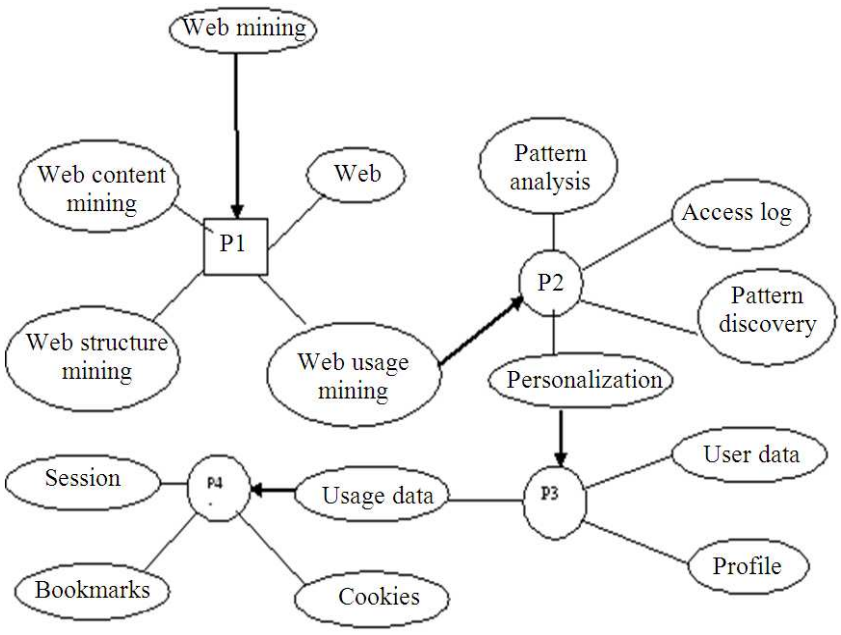

Fig. 3: Tracking the user interest through preference network

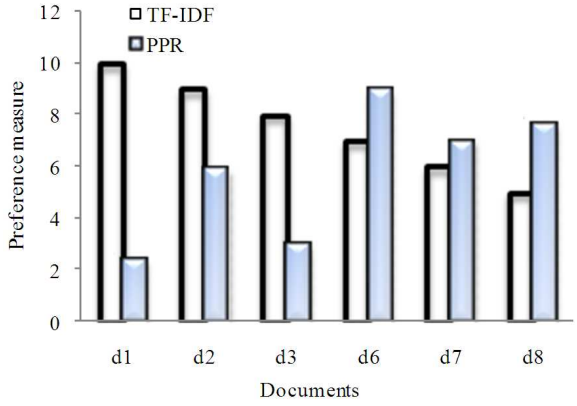

Fig. 4: TF-IDF and PPR based preference measure of the documents

Table 1: PPR computation

Preferred term	TF-IDF	UIS	PPR
Web	0.530	0.12	0.30
Web usage mining	0. 950	0.70	0.81
Web structure mining	0. 600	0.46	0.52
Web content mining	0. 112	0.05	0.07
Personalization	0.900	0.83	0.86
Pattern analysis	0. 606	0.44	0.51

Table 2: Query-term preference list

	Keyword Indexing	
Query		
Web Mining	Existing	Proposed
	Web	Personalization
	Web Mining	Usage data
	Web content mining	User data
	Web structure mining	Profile
	Internet	Access log
	Web usage mining	Pattern analysis
	Data mining	Web usage mining

Existing page rank of search results for the specified query “Web Mining” could retrieve the pages mainly based on the occurrence of the query term in the retrieved web pages.

Query-term preference list of the existing and the proposed system is illustrated in Table 2. It shows the way in which the proposed work re-ranks the search results based on the user preference. User preferred terms are the major contributing factor towards search result personalization. According to the terms list extracted, the web pages containing these preferred terms will gain higher rank than those that contain simply the query term.

The user clicked document set {d1, d2, d3, d6, d7, d8} among the retrieved results for the same query under study was considered for preference computing. The preferred measure for each document in the prescribed set are computed from both existing and proposed perspectives and the same is shown in the Fig. 4.

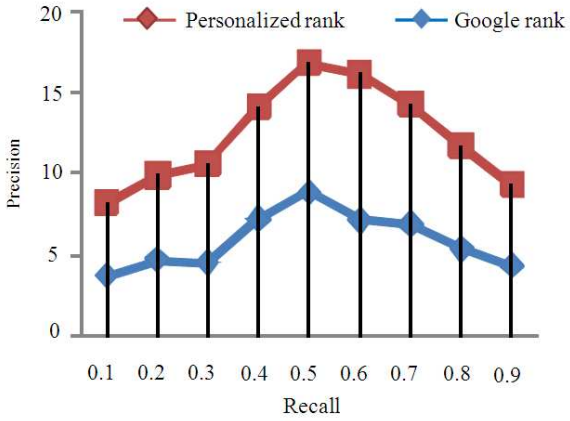

Fig. 5: Precision-recall plots of two different ranking schemes

Precision (the ratio between the number of relevant results retrieved for the number of retrieved documents) recall (the percentage of relevant documents retrieved) measure corresponding to Google ranking and the proposed ranking method are compared and the same is shown in Fig. 5.

CONCLUSION

We introduced a strategy for personalizing the Page Rank based on the user's interest score computed from the preferred network based profile. User interested categories are tracked without user intervention. Based on the UIS, the corresponding results will be mapped and produced at the user end. The user can easily identify the relevant pages among the search results. Our method relies on the quality of the extracted preferred term list and the results prove that the proposed scheme can obtain more personalized results. We are analyzing on the profile convergence features which may further improve the ranking of the search results.

REFERENCES

Agichtein, E., E. Brill, S. Dumais and R. Ragno, 2006. Learning user interaction models for predicting web search result preferences. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (RDIR’ 06), ACM, Seattle, WA, USA., pp: 3-10. DOI: 10.1145/1148170.1148175

Aktas, M.S., M.A. Nacar and F. Menczer, 2004. Using hyperlink features to personalize web search. Adv. Web Mining Web Usage Anal., 3932: 104-115. DOI: 10.1007/11899402_7

Algarni, A., Y. Li and X. Tao, 2010. Mining specific and general features in both positive and negative relevance feedback. Proceedings of the 19th Text REtrieval Conference: Relevance Feedback Track, Nov. 16-19, Gaithersburg, MD, USA.

Barouni-Ebrahimi, M., E. Bagheri and A.A. Ghorbani, 2008. A frequency mining-based algorithm for re-ranking web search engine retrievals. Adv. Artif. Intell., 5032: 60-65. DOI: 10.1007/978-3-540-68825-9_6

Bhowmick, P.K., S. Sarkar and A. Basu, 2010. Ontology based user modeling for personalized information access. Int. J. Comput. Sci. Appli., 7: 1-22.

Brin, S. and L. Page, 1998. The anatomy of a large-scale hypertextual web search engine. Proceedings of the 7th International World Wide Web Conference, Apr. 14-18, Brisbane, Australia, pp: 101-117.

Dali, L., B. Fortuna and J. Rupnik, 2010. Learning to rank for personalized news article retrieval. J. Machine Learning Res.-Proc. Track, 11: 152-159.

Harb, H.M., A.R. Khalifa and H.M. Ishkewy, 2009. Personal search engine based on user interests and modified page rank. Proceedings of the International Conference on Computer Engineering and Systems, Dec. 14-16, IEEE Xplore Press, Cairo, pp: 411-417. DOI: 10.1109/ICCES.2009.5383228

Haveliwala, T.H., 2002. Topic-sensitive PageRank. Proceeding of the 11th International Conference on World Wide Web, May 7-11, ACM Press, Honolulu, HI, USA., pp: 517-526. DOI: 10.1145/511446.511513

Hu, J. and P.K. Chan, 2008. Personalized web search by using learned user profiles in re-ranking. Florida Institute of Technology, USA.

Kavita, D.S. and S.Z. Gawali, 2010. Web search result personalization using web mining. Int. J. Comput. Appli., 2: 29-32.

Kim, H.R. and P.K. Chan, 2005. Personalized ranking of search results with learned user interest hierarchies from bookmarks. Proceedings of the 11th SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 21-21, Chicago, Illinois, USA.

Kumar, P.R. and A.K. Singh, 2010. Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval. Am. J. Applied Sci., 7: 840-845. DOI: 10.3844/ajassp. 2010.840.845

Paul-Alexandru, C., D. Olmedilla and W. Nejdl, 2004. PROS: A personalized ranking platform for web search. Adaptive Hypermedia Adaptive Web-Based Syst., 3137: 431-461. DOI: 10.1007/978-3-540-27780-4_7

Qui, F. and J. Cho, 2006. Automatic identification of user interest for personalized search. Proceedings of the 15th international conference on World Wide Web, May 22-26, ACM Press, Edinburgh, Scotland UK., pp: 727-736. DOI: 10.1145/1135777.1135883

Ramadhan, H.A., K. Shihab and J.H. Ali, 2006. Improving the ranking capability of the hyperlink based search engines using heuristic approach. J. Comput. Sci., 2: 638-645. DOI: 10.3844/jcssp. 2006.638.645

Ratnakumar, A.J., 2008. An implementation of web personalization using web mining techniques. J. Theoretical Applied Inform. Technol.

Shanna, A.K., N. Aggarwal, N. Duhan and R. Gupta, 2010. Web search result optimization by mining the search engine query logs. Proceeding of the International Conference on Methods and Models in Computer Science, Dec. 13-14, IEEE Xplore Press, New Delhi, pp: 39-45. DOI: 10.1109/ICM2CS.2010.5706716

Teevan, J., S.T. Dumais and E. Horvitz, 2005. Personalizing search via automated analysis of interests and activities. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 15-19, ACM Press, Salvador, Brazil, pp: 449-456. DOI: 10.1145/1076034.1076111

Zhuang, Z. and S. Cucerzan, 2006. Re-ranking search results using query logs. Proceedings of the 15th ACM international Conference on Information and Knowledge Management, Nov. 5-11, ACM Press, Arlington, VA, USA., pp: 59593-433. DOI: 10.1145/1183614.1183767