IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661, ISBN: 2278-8727Volume 6, Issue 6 (Nov. - Dec. 2012), PP 36-43
www.iosrjournals.org

Query- And User-Dependent Approach for Ranking Query Results in Web Databases

Sruthi Ambati[1], Raghava Rao[2]
[1]*Department of CSE, DRK Institute of Science & Technology, Ranga Reddy, Andhra Pradesh, India HOD*
[2]*Department of CSE, DRK Institute of Science & Technology, Ranga Reddy, Andhra Pradesh, India*

Abstract: *Internet has paved the way for the emergence of web databases. Querying such databases for required information has become a common task. Ranking such query results is an open problem to be addressed. The existing solutions such as user profiles, query logs, and database values perform ranking in user independent and/or query independent fashion. This can't provide efficient ranking. This paper presents a new approach known as Query and User Dependent Ranking for giving ranking to query results of deep web. The proposed ranking framework is based on two fundamental aspects to the problem of ranking query results. They are query similarity and user similarity. These similarities are exploited to make efficient ranking of query results. A prototype application is built to test the efficiency of our model. The empirical results revealed that our approach is efficient and can be used in real world applications.*

Index Terms: *Deep web, ranking query results, user similarity, query similarity*

I. Introduction:

Due to emergence of Internet and its related technologies, people of all walks of life started storing data over web. This helps them to access the content from anywhere in the world. Thus web databases also emerged. These web databases are known as deep web [1], [2]. The web databases are from various domains such as vehicles, real estate, education, health care and so on. These web databases are searched by online users through a search mechanism provided. The queries can have criteria that correspond to the attributes of the database schema. When results returned are huge in number, user time gets wasted in brewing for required information. To overcome this problem the present web databases simplify the results by sorting them in a particular attribute. This may not be suitable to the requirements of many users who prefer ordering on multiple attributes. There are many existing web databases in the world that can be accessed through WWW. For instance Google's Google Base which is a web database that stored information about vehicles with all relevant attributes such as Make, Mileage, Price etc. In this table each record represents a vehicle in the real world that is ready for sale. Two scenarios which are common are considered for making a new ranking model in this paper. The first scenario is that various web users can have different ranking preferences towards the results of the same query. This is because each user requires different information that is part of the same query results. Thus ranking preferences of each user is different. The second scenario is that same user ma has different ranking preferences for the results of different queries. From these two scenarios it can be understood that various users make queries to web databases and the queries they make may be similar or different. Therefore it is ideal to have user dependent and also query dependent similarity for ranking query results. Many existing web databases follow simple sorting for ranking while the extension of SQL allows providing attribute weights [3], [4]. For most web databases this approach is not user friendly and users have to waste some of their time in browsing the query results. For this reason an automated ranking of query results is studied and some techniques are proposed in [5], [6], and [7]. However these approaches either user query independent or user independent way of ranking query results. Another approach used in [8] if to build extensive user profiles and in that case users are supposed to order the records. This is proposed for user-dependent ranking and do not differentiate the different between different queries and provide a single ranking order for any query. Even recommender systems made use of either user similarity or query similarity. Some of them are collaborative in nature [9], [10], [11] and some of them are content based filters [12], [13]. The work in this paper is inspired from those works. However, there is difference between them. Our work is based on user similarity and also queries similarity. It does mean that the proposed approach is user and query dependent ranking for web database.

In this paper the ranking model is based on two notions such as user similarity and query similarity. User similarity indicates that different users can have same preferences. Query similarity indicates that different users can have same queries. In order to achieve this ranking of users and queries are to be maintained. We have developed a workload file that contains the user and query ranking functions. When new record is entered into web database, obviously that is given by a user. There might be many users who issued that query previously and there might be same queries issued earlier. The workload file is in tabular format and it gets updated with ranking functions as per the proposed algorithm as and when new queries are made. The proposed model has

two models mixed. They are known as user dependent ranking model and query dependent ranking model. However, we prefer applying both of them for better results. The proposed ranking modal in this paper is a linear weighted sum function. It contains attribute weights and value weights. Attribute weights indicate the importance of attributes while the value weight indicates the importance of values of attributes. Relevance feedback techniques [14] are utilized for making the workload minimal. The main contributions of this paper are

- User and query dependent approach for ranking web databases.
- Ranking model based on user similarity and query similarity notions.
- Two synthetic databases such as college and hospital used for experiments. However, the model can be tested with web databases.
- We used a new workload concept for maintaining the updated ranking of users and queries.

II. Related Work:

Usage of web databases has brought the ranking the query results concept. There is no such requirement in case of relational databases. However it is there in case of IR for some time. Ranking gained popularity with emergence of deep web. Ranking has become an essential task as the results of query results in large number of records that waste user's time as he has to browse the results for actual information required. Recommender systems have been using ranking for making best recommendations to end users of online applications. With respect to user and query similarity this paper resembles to the work done in [9], [11] and [10]. It also has some relevance with content filtering mechanisms explored in [13] and [15]. There is main different between ranking a database and making recommendations. This way our work differs from existing work. The existing web databases make use of simple ordering for ranking while our proposed framework focuses on user similarity and query similarity based. Moreover the existing techniques for ranking do not user both similarities. They are either user independent or query independent or independent of both of them. In the case of recommender systems [16], and [17] each attribute holds presence or absence of the user input tag. In case of user and query similarity matrix each cell contains ordered set of functions represented by ranking function. These way recommender systems differ from web databases. However, both of them make use of ranking phenomenon. The notion of similarity also makes our work different from existing ones as specified earlier approaches don't use query and user similarity together. This paper makes use of query and user similarity at the same time and the resultant ranking function is updated in to a workload file which is nothing but a relational table in this paper. Incase of content filtering concepts, the similarity is found by domain expert or user profiles are used [15]. The same can be achieved by using user profiles [15]. Direct user profiles usage in our paper is not possible as we need to find both user and query similarities. Thus the work load file concept in this paper is having prominent importance. This is because the user profile gives importance to user information rather than the queries. We also assume that different users have different query preferences and same user may have different query preferences. The notion of user similarity is same as the concept used in collaborative filtering; however the technique used is different. Based on ratings users are compared in collaborative filtering; our work extends user personalization besides considering query similarity notion. Thus our work stands different from many existing works. As web databases and query results of them has received attention from academic circles, the user and query dependent ranking has not been addressed. Chaudhuri et al. [5] addresses only query – dependent ranking using vector and IR models. In [6] user – dependent ranking for web databases is explored. In either case both user and query dependent ranking is not used. Thus this paper is first in its solution for both user and query dependent ranking of query results of web databases. In [18] the approach for user similarity requires the user to specify ordering without making queries. These approaches do not recognize the fact that users are having different ranking preferences.

The closest of our approach are [3], [4] and [19]. However, these techniques are not suitable for efficient ranking of query results of web databases as they do not consider both user and query similarity notions. The cosine similarity metric proposed in [20] and the IR method proposed in [21] and relevance feedback approaches in [14], [22], [23] and [24] are not suitable for direct use for web databases. For this reason we implement a ranking model that is based on the user and query similarity. It does mean that the proposed model is query and user dependent ranking model.

III. Proposed Ranking Algorithm

```
INPUT: Ui, Qj, Workload W (M queries, N users)
OUTPUT: Ranking Function Fxy to be used for Ui, Qj
STEP ONE:
For P ¼ 1 to M do
%% Using Equation 2 %%
Calculate Query Condition Similarity (Qj, Qp)
End for
%% Based on descending order of similarity (Qj, Qp)
Sort(Qi, Q2,.....QM)
Select QKset i.e., top-K queries from the above sorted set
STEP TWO:
Forr ¼ ! to N do
%% Using Equation 7 %%
Calculate User Similarity(Ui, Ur) over QKset
End for
%% Based on descending order of similarity with Ui %%
Sort(U1, U2,.....UN) to yield Uset
STEP THREE:
For Each Qs 2 QKset do
For Each Ut 2 Uset do
Rank (Ut:Qsp ¼
```

Listing 1 – Ranking algorithm [25]

This algorithm is based on the algorithm given in [36] and implemented in the prototype application which shows both user and query dependent rankings for query results of web databases.

IV. Sample Workload File

The sample workload file is given in fig. 1 which shows queries, users and the ranking functions calculated as per the algorithm given in listing 1.

Table 1 – Sample Workload

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
U1	??	F12	--	--	F15	--	F17	
U2	F21	F22	--	F24	--	F26	F27	--
U3	F31	F32	F33	F34	--	--	F37	--

V. Experimental Evaluation

The experiments are made using a prototype application with two synthetic web databases such as college and hospital. The experimental results are evaluated by visualizing the results in the form of graphs. Fig. 1 and 2 shows the ranking quality of query similarity models for both databases with 10% work load.

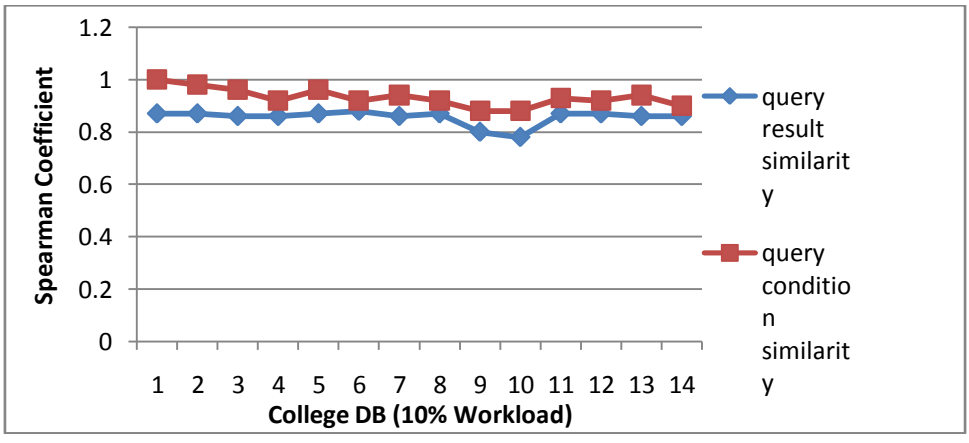

Fig. 2 – Ranking quality of query similarity (College DB)

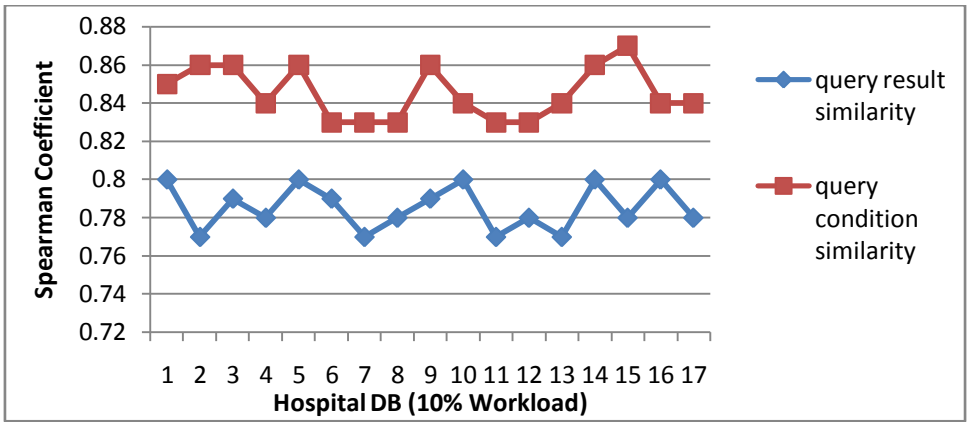

Fig. 3 – Ranking quality of query similarity (Hospital DB)

As can be seen in fig. 2 and 3, query condition similarity average is found across all queries. The X axis shows queries while the Y axis shows spearman coefficient. As it is evident in the graphs, the query condition model outperforms query result model. The loss of quality is due the restricted workload that is 10%. When workload increases, the quality also increases.

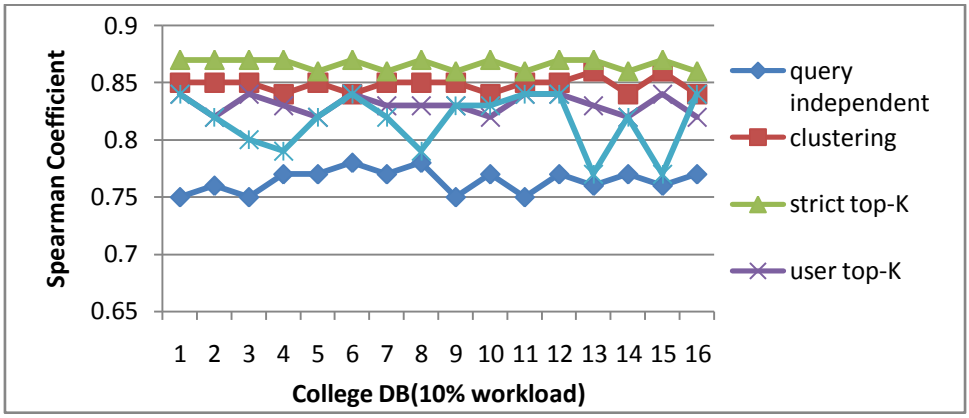

Fig. 4 – Ranking quality of user similarity model (College DB)

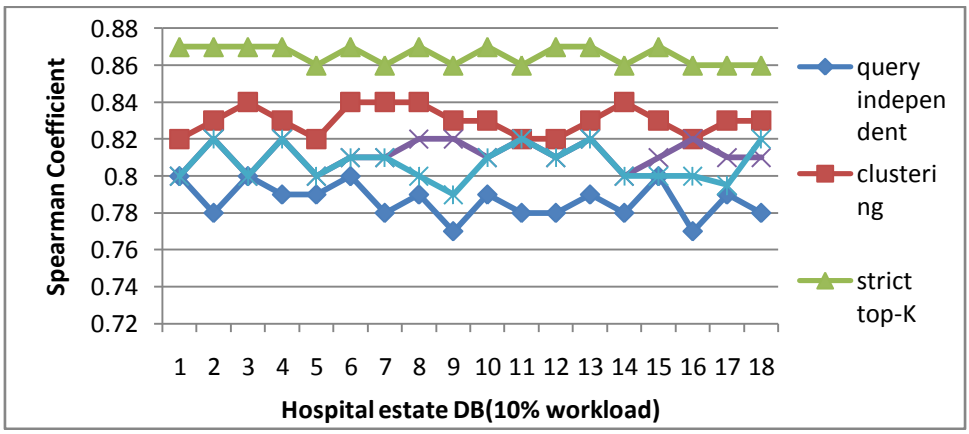

Fig. 5 – Ranking quality of user similarity model (Hospital D)

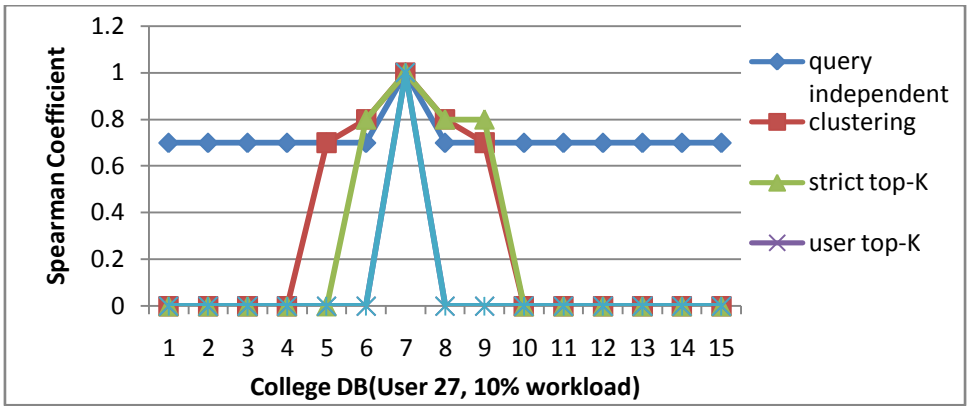

Fig. 6 – Ranking quality of user similarity model (College DB)

Fig. 4, 5, and 6 show the average ranking quality achieved fro both college and hospital database across all queries for all users. The results reveal that strict top-K model performs bittern than other models. However, the strict top-K has no ranking functions for many queries.

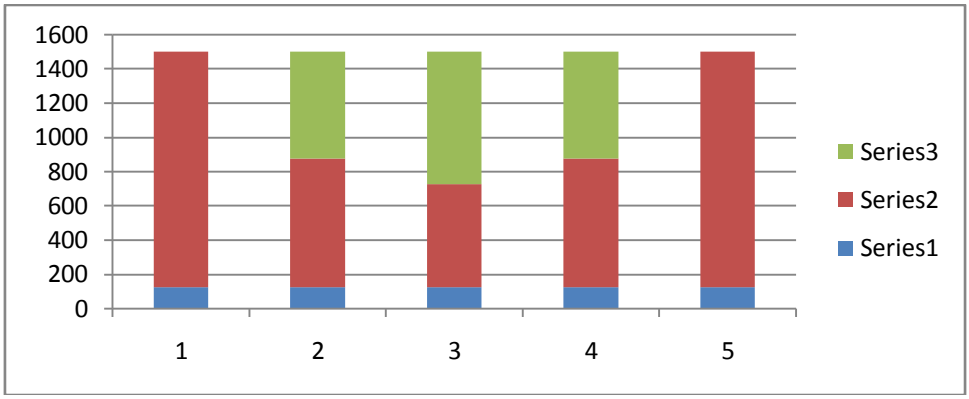

Fig. 7 – Ranking functions derived for user similarity (College DB)

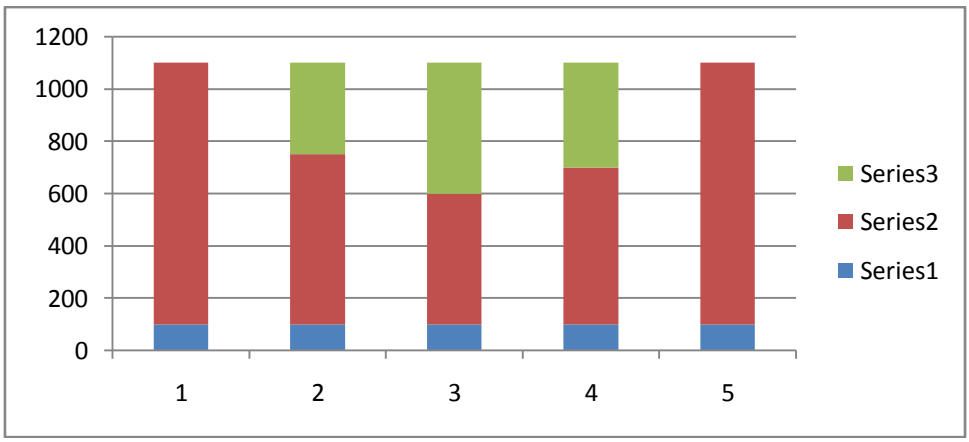

Fig. 8 – Ranking functions derived for user similarity (Hospital DB)

Fig. 7 and 8 confirm the fact that different models have different abilities for determining ranking functions across the workload. Nevertheless, the strict top-K is accurate and superior to all other models from the perspective of ranking function.

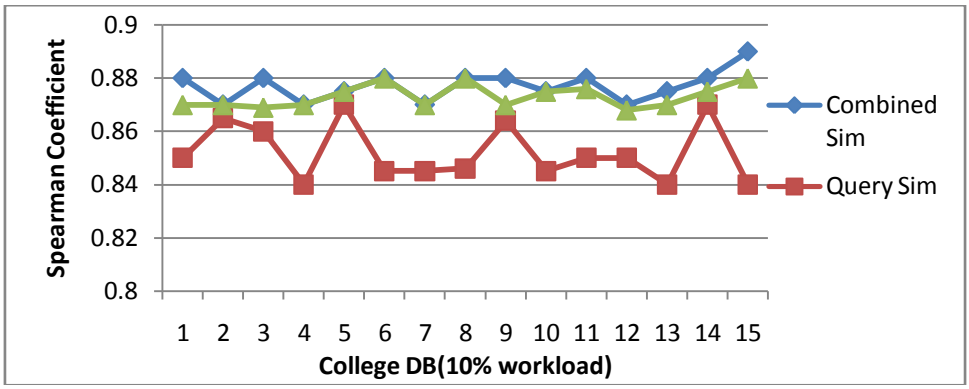

Fig. 9 – Ranking quality of combined similarity model

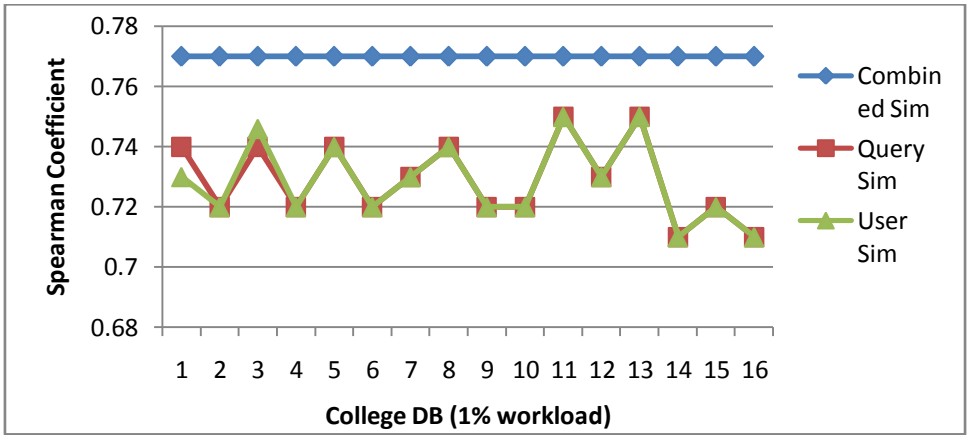

Fig. 10 – Ranking quality of combined similarity model

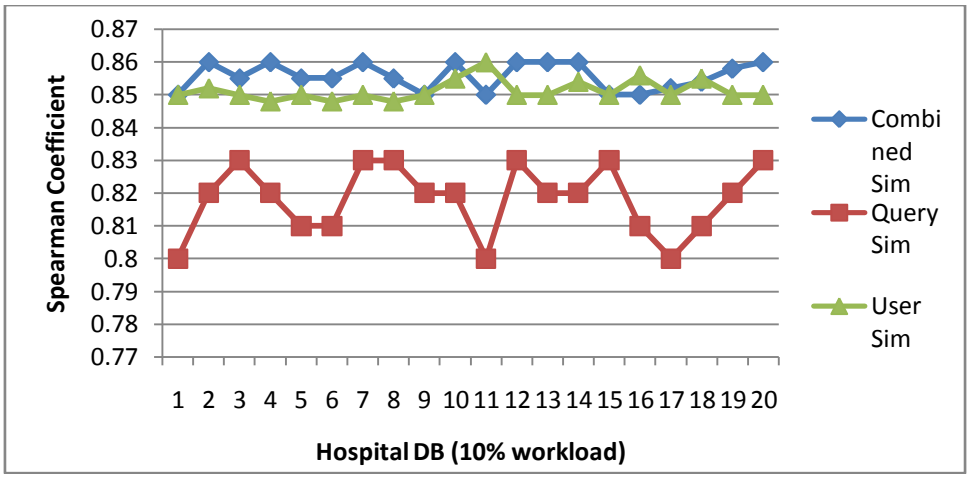

Fig. 11 – Ranking quality of combined similarity model

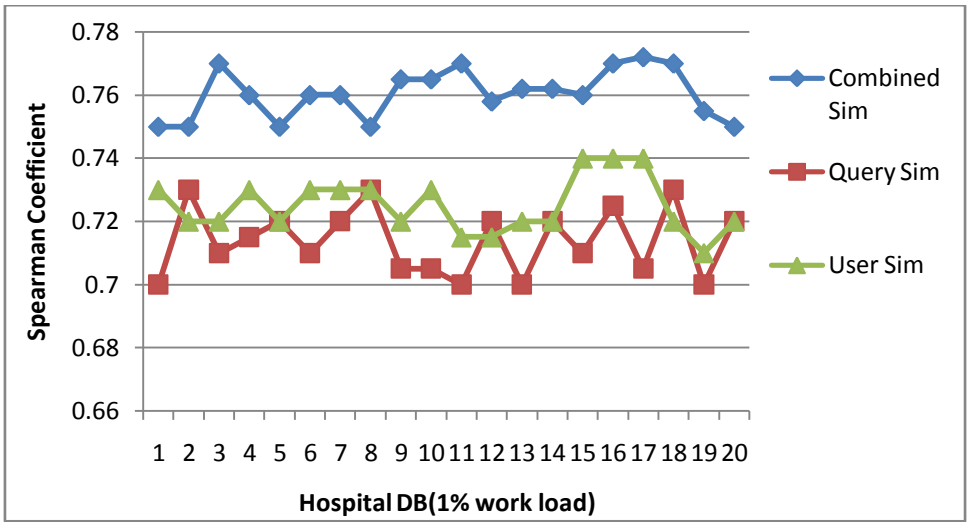

Fig. 12 – Ranking quality of combined similarity model

Fig. 9, 10, 11 and 12 show the quality of combined models for both databases with 1% and 10% workload. The important observation is that the composite model is performing better than other individual models. Another fact established here is that with more ranking functions in workload better similarity and quality of results is achieved.

VI. Conclusion

This paper proposed a new ranking model for ranking query results of web databases. We used two synthetic web databases for experiments. They are college database and hospital database. The model is based on both query similarity and user similarity. We have also built a prototype web based application that demonstrates the efficiency of the proposed ranking model. A workload file is maintained that that continually stores updated ranking functions for both user similarity and query similarity. When a new query is made, this workload file is used for giving ranking to the query results. Designing and maintaining a workload is challenging in the context of web databases. We have implemented an algorithm for computing user and query similarities and update workload consistently. The experimental results revealed that our new ranking model works well and it can be explored for real world web databases.

References:

[1] M.K. Bergman, "The Deep Web: Surfacing Hidden Value," J. Electronic Publishing, vol. 7, no. 1, pp. 41-50, 2001.
[2] K.C.-C. Chang, B. He, C. Li, M. Patil, and Z. Zhang, "Structured Databases on the Web: Observations and Implications," SIGMOD Record, vol. 33, no. 3, pp. 61-70, 2004.
[3] C. Li, K.C.-C. Chang, I.F. Ilyas, and S. Song, "Ranksql: Query Algebra and Optimization for Relational Top-k Queries," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 131-142, 2005.
[4] A. Marian, N. Bruno, and L. Gravano, "Evaluating Top-k Queries over Web-Accessible Databases," ACM Trans. Database Systems, vol. 29, no. 2, pp. 319-362, 2004.
[5] S. Chaudhuri, G. Das, V. Hristidis, and G. Weikum, "Probabilistic Ranking of Database Query Results," Proc. 30th Int'l Conf. Very Large Data Bases (VLDB), pp. 888-899, 2004. 684 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 9, SEPTEMBER 2012 Fig. 10. Ranking quality of learning models.
[6] W. Su, J. Wang, Q. Huang, and F. Lochovsky, "Query Result Ranking over E-Commerce Web Databases," Proc. Conf. Information and Knowledge Management (CIKM), pp. 575-584, 2006.
[7] H. Yu, Y. Kim, and S. won Hwang, "Rv-svm: An Efficient Method for Learning Ranking Svm," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), pp. 426-438, 2009.
[8] G. Koutrika and Y.E. Ioannidis, "Constrained Optimalities in Query Personalization," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 73-84, 2005.
[9] J. Basilico and T. Hofmann, "A Joint Framework for Collaborative and Content Filtering," Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 550- 551, 2004.
[10] T. Hofmann, "Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis," Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 259-266, 2003.
[11] D. Billsus and M.J. Pazzani, "Learning Collaborative Information Filters," Proc. Int'l Conf. Machine Learning (ICML), pp. 46-54, 1998.
[12] M. Balabanovic and Y. Shoham, "Content-Based Collaborative Recommendation," Comm. ACM, vol. 40, no. 3, pp. 66-72, 1997.
[13] C. Basu, H. Hirsh, and W.W. Cohen, "Recommendation as Classification: Using Social and Content-Based Information in Recommendation," Proc. 15th Nat'l Conf. Artificial Intelligence (AAAI/IAAI), pp. 714-720, 1998.
[14] B. He, "Relevance Feedback," Encyclopedia of Database Systems, pp. 2378-2379, Springer, 2009.
[15] S. Gauch and M. Speretta, "User Profiles for Personalized Information Access," Adaptive Web, pp. 54-89, 2007.
[16] S. Amer-Yahia, A. Galland, J. Stoyanovich, and C. Yu, "From del.icio.us to x.qui.site: Recommendations in Social Tagging Sites," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 1323-1326, 2008.

[17] A. Penev and R.K. Wong, "Finding Similar Pages in a Social Tagging Repository," Proc. 17th Int'l Conf. World Wide Web (WWW), pp. 1091-1092, 2008.
[18] S.-W. Hwang, "Supporting Ranking For Data Retrieval," PhD thesis, Univ. of Illinois, Urbana Champaign, 2005.
[19] K. Werner, "Foundations of Preferences in Database Systems," Proc. 28th Int'l Conf. Very Large Data Bases (VLDB), pp. 311-322, 2002.
[20] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. ACM Press, 1999.
[21] N. Fuhr, "A Probabilistic Framework for Vague Queries and Imprecise Information in Databases," Proc. 16th Int'l Conf. Very Large Data Bases (VLDB), pp. 696-707, 1990.
[22] Y. Rui, T.S. Huang, and S. Mehrotra, "Content-Based Image Retrieval with Relevance Feedback in Mars," Proc. IEEE Int'l Conf. Image Processing, pp. 815-818, 1997.
[23] L. Wu et al., "Falcon: Feedback Adaptive Loop for Content-Based Retrieval," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 297- 306, 2000.
[24] X. Luo, X. Wei, and J. Zhang, "Guided Game-Based Learning Using Fuzzy Cognitive Maps," IEEE Trans. Learning Technologies, vol. 3, no. 4, pp. 344-357, Oct.-Dec. 2010.
[25] Aditya Telang, Chengkai Li, and Sharma Chakravarthy, "One Size Does Not Fit All: Toward User- and Query-Dependent Ranking for Web Databases", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 9, SEPTEMBER 2012.

AUTHORS

Sruthi Ambati is a student of DRK Institute of science and Technology, Ranga Reddy, Andhra Pradesh, India. He has received B.Tech degree in Computer Science and Engineering and M.Tech Degree in Computer Science and Engineering. Her main research interest includes Data Mining and Image Processing.

Raghava Rao.N is working as HOD and Associate Professor at DRK Institute of Science & Technology, Ranga Reddy, and Andhra Pradesh, India. He has received M.Tech Degree in Computer Science and Engineering. His Main Interest includes Cloud Computing, Software Engineering.