ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Managing XML Retrieval through Personalization Using Search Engine

B. Kurinchi Meenakshi [1], Dr.C.Nalini [2]

PG Scholar, Dept of C.S.E, Bharath University, Chennai, Tamil Nadu, India [1]

Associate Professor, Dept of C.S.E, Bharath University, Chennai, Tamil Nadu, India [2]

ABSTRACT: The number of resources of information has increased significantly and ***Information Retrieval (IR)*** based on keyword in web has become very significant.XML has become the widely used format for sharing of information. As the number of resources of information has increased significantly and retrieval of correct data according to user preference may not be achieved efficiently. In order to improve the search of XML documents according to the user requirement and preference we use personalized search based on user preference stored as an XML document. The problem in personalized search is in identifying the correct preferences based on the search text. We have a proposed solution, in which the user preferences stored as keywords in an XML document are identified based on the query the user enters and is ranked based on ranking function of top k algorithm. The documents will be listed as two separate list one based on keyword and the other based on user preference with the relevance status value. The documents in both the lists can then be listed based on reranking strategies. The documents matching in both the list will be listed first and then reranked based on user preference score and relevance status value .The remaining documents will be included for listing based on reranking strategies [1].In this paper we have shown the simulation of identifying the user preference based on the user input and displaying the xml documents based on user query and preference and ranking them based on reranking strategies.

KEYWORDS: Relevance Status Value, XML, Ranking, top k algorithm, Re ranking, Personalization.

I. INTRODUCTION

Information search has become very crucial with the growth of internet. There are wide range of information available in different formats. Handling wide range of information in different format will be difficult to handle.XML helps in overcoming this issue as it is the standard and universally accepted format. Most of the applications currently handle their data in XML format so that it can be widely used and transferred. Handling data in XML format facilitates search engines to retrieve data. The XML documents can be retrieved and displayed based on the user search term.

As there are wide collection of information available the resultant set of information which are retrieved may be huge. The user may be interested only in some of the resultant documents out of the huge set of retrieved information. In order to make the user search more effective user personalization has been considered. The user personalized data can be captured based on personalization techniques [7][8] or by getting the preferences from the user.

The user preferences can be stored and can be used in retrieving the result documents according to the user query. The issue in using user preference for retrieval is that not all user preferences are required in retrieving the result documents. The user preferences retrieved should be of relevance to the search key entered. So the issue is to identify the user preferences according to the given query and derive the score of the preferences according to the given query. In this paper an approach for identifying the user preference and their score based on user search text

for retrieving documents has been discussed in the steps as follows. In step 1, the user preferences are stored in XML with preferences as nodes having keywords. In step2, the user preferences based on the user search text are retrieved and ranked based the relevance of the keyword with the preference node using top k algorithm. In step 3 the documents retrieved based on user preferences compared with the original search key document list and the final xml document list is retrieved based reranking strategies.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

In this paper we will discussing on the related work in section 2 in which all the referred papers have been mentioned In section 3 proposed system has been discussed. In section 4 the Experimental setup for the simulation and in section 5 the Experiment Result has been discussed. In section 6 the conclusion and future work has been discussed.

II. RELATED WORK

The works which had been carried out in the areas related to user personalized search have been discussed below.In [1], authors have proposed an approach for making the user search more effective by using the concept of weighted personalized parameters. The results of original query and expanded query have been compared to identify the documents based on user's interest based on ReRanking approach. In [2], authors have presented an approach of identifying the XML nodes according to the keyword and ranking them based on score of relevance. The problem of fuzzy typ-ahead search and the efficiency of methods such as LCA Based interactive search and top-k algorithm have been discussed to complete the search and ranking of XML documents more efficiently. In [3],Varun Varma Sangaraju has proposed a system of using Adaptive Search for searching and ranking XML based documents. In this system users can select the search algorithm based on its benefits. The search algorithms like Boolean Retrieval algorithm and LCA based algorithm have been discussed along with ranking algorithm. In [4], authors have proposed evaluation measures which makes use of score of a document calculated based on highlighted text and full text of the document which will be a value between 0 and 1.This helps to identify the document parts based on relevance and content. In [5] ,authors have presented the advantages of using Fuzzy search techniques in XML search over Xpath and Xquery and the efficiency of Minimum cost tree and LCA based search algorithms. The user need not know about the XML data when using this search. In [6],authors have presented on ranking the XML query results according to the user search intention and relevance using XML TF*IDF and KWSearch algorithm. Using XML TF and XML IDF the confidence level of each node is computed for it to be searched. They have also proposed a ranking scheme based on XML TF and XML IDF to arrive at the hierarchical structure of XML data. In [7][8],the authors have discussed about the different personalization techniques for capturing the personalized information. In [9], the authors have done a survey on managing XML retrieval through personalization..

III. PROPOSED SYSTEM

The proposed system stores the user preference as nodes and sub nodes with keywords in an xml document. The user preference nodes related to the search key is identified by parsing the user preference XML using DOM (Document Object model). The rank of these nodes will be calculated using the ranking function in top-k algorithm. The xml document list based on the user search text will be retrieved initially with relevance status value. The xml documents based on the user preferences nodes will be retrieved as the second list of xml documents. Both the list of documents will be compared and returned as the final resultant document list after considering the reranking strategies and preference node scores.The overall architecture of the proposed system is given below.

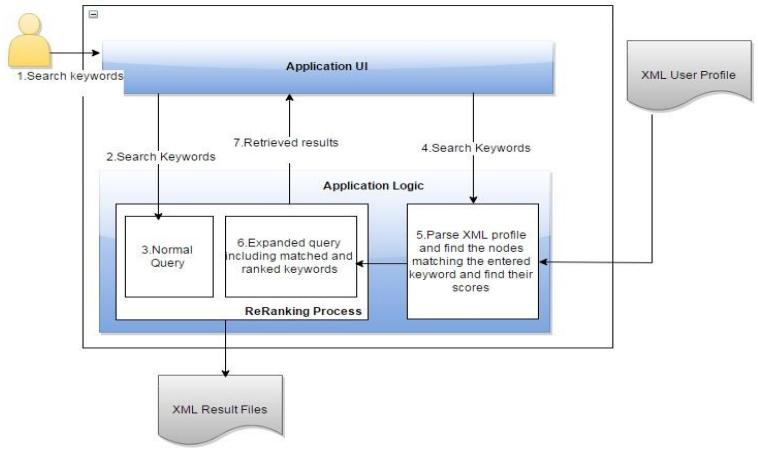

Fig 1.Overall architecture diagram

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

The proposed system has the following modules which have been explained below

1. User Preference XML
2. Normal Search XML document retrieval
3. Identification of user preference and score
4. Personalised Search XML document retrieval
5. Final Document result set based on HRR, SRR, IRR

1. User Preference XML

This module gets the user preferences at the time of user registration the user preferences are captured as nodes and sub nodes with keywords in an xml format.

2. Normal Search XML document retrieval

The xml documents with the user search text will be retrieved with the Relevance Status Value and ordered based on RSV value.

3. Identification of user preference and score

The user preference nodes will be identified based on the presence of search text present as keyword in the user preference xml. The score of those user preference nodes is calculated based on the ranking function of top -k algorithm.

The ranking function will calculate the rank based on the below calculation

If node n contain keyword ki
The score or relevance of node n and keyword ki is calculated by

$$\text{SCORE1 (n, ki)} = \frac{\ln(1+tf(ki,n))*\ln(idf(ki))}{(1-s)*s*ntl(n)}$$

Where

tf (ki, n)	-	no of occurrence of ki in sub tree rooted n
idf (ki)	-	ratio between number of node in XML to number of nodes that contain keyword ki
ntl (n)	-	length of n/length of nmax
nmax	-	node with max terms
s	-	constant set to 0.2

If there are more number of search text then we first evaluate the relevance between node n and each input keyword, and then the overall score can be calculated by combining the individual relevance scores. We then rank the nodes based on the score.

4. Personalized search XML document retrieval

The xml documents will now be retrieved based on user preferences. The documents will be retrieved with the relevance status value and order based on relevance status value.

5. Final Document result set based on HRR, SRR, IRR.

This modules retrieves the documents considering the normalised search list and personalised search list.
The final result set can be retrieved either base don HRR or SRR or IRR

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Reranking strategy using HRR

The documents matching normalised and personalized list will be listed first and will be sorted based on personalized document list RSV. The user preference node will now be sorted based on their descending scores and will be used to retrieve the documents from the matching list considering one node at a time. This will now result in the xml documents being retrieved based on user preference and based on RSV. The documents unmatched in the normal search will be appended to the final list.

Reranking strategy using SRR

This will be similar to HRR except that the matching list will be sorted based on the sum of RSV value of matching normal search document and RSV value of user preference search document.

Reranking strategy using IRR

This is similar to SRR except that in addition to the unmatched records from the normal search, the unmatched records from user personalized search will also be added to the final list.

We can view the content of the xml document from the final list of document one at a time.

IV. EXPERIMENTAL SETUP

Experiment has been done with the following design considerations.

- The user preferences is captured at the time of registration.
- The user preference is stored as an xml file and is named based on the user id which will be unique.
- The user registration details such as username, password, contact number and DOB has been stored in a registration table which will be used for validating the credentials during login.
- The list of resultant xml documents is maintained as a file list storage.
- A predefined list of user preferences has been used for simulation which needs to be filled by the user at the time of registration. Fig 2.shows the screen shot of the screen with some predefined list of preferences where the user can specify their preferences as comma separated text.

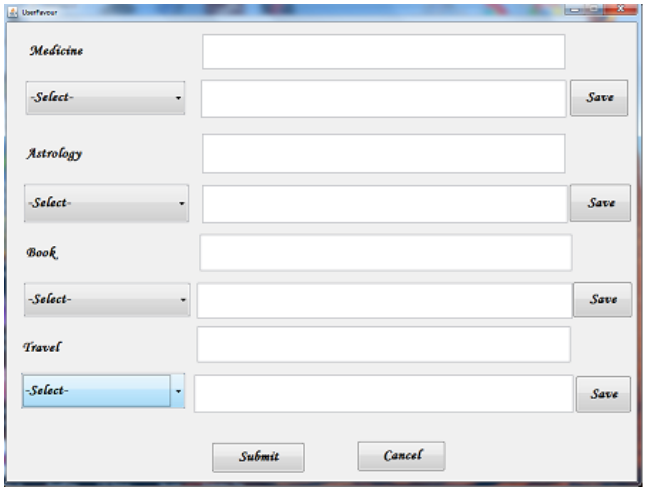

Fig 2.User Preference Screen

- The search documents can be listed either based on HRR, SRR or IRR by selecting the option one at a time.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

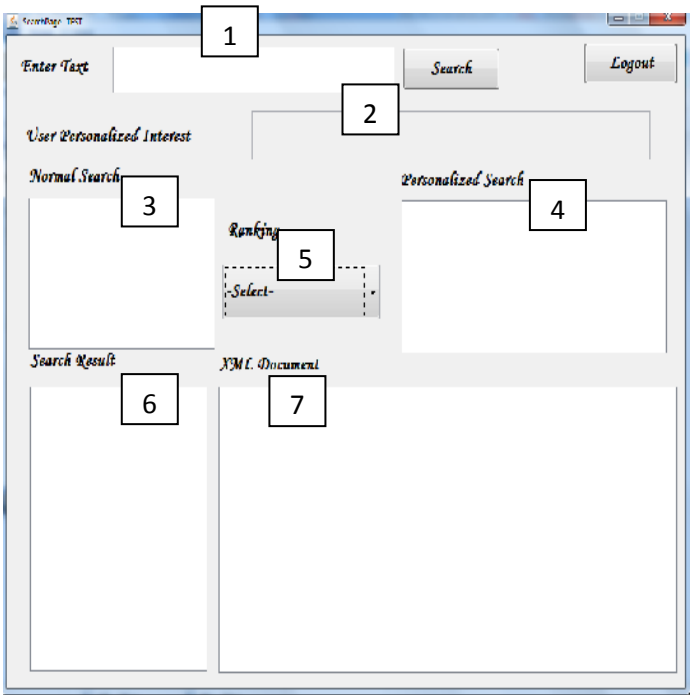

Fig 3.Search Engine screen.

Fig 3 shows the search screen which can be used by the user for listing the xml documents based on search text.
The user will enter the search text in 1.The user preference nodes based on search text will be listed in 2.XML documents retrieved from using normal search will be listed in 3.The documents retrieved using personalized user preference nodes will be listed in 4.The xml documents from both the list can be reranked based the reranking option selected in 5.The final xml document result set will be listed in 6.The user can view the document in section 7 by clicking on the document name in 6.

The simulation has been done using a application which is developed using Java Swing as the front end and MYSQL as the backend. The other system requirements are 128 mb RAM, Windows Operating system and JDK1.5.

V. EXPERIMENT RESULTS

The experiment has been carried out by entering a search text which returns four user preferences nodes. The scores for these three nodes were calculated based ranking function of top k algorithm.

6 xml result files were retrieved from the xml document list sorted based on RSV value for the user entered text. The preference nodes were retrieved based on the search text and the xml result files which contain the preferences node text were retrieved with RSV as a separate document list. There were some 10 xml result files.

The final number of documents retrieved using HRR was same as the no of documents returned using normal search but the order is based on RSV of personalised result RSV. Initially the matching records were listed .The same list was looped through by the set of preference nodes based on the score in descending order. So the documents were listed based on first preference node which has the highest score and within that list the documents were listed based on RSV. This was repeated for all set of preference nodes. Then the unmatched documents from normal search will be appended.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

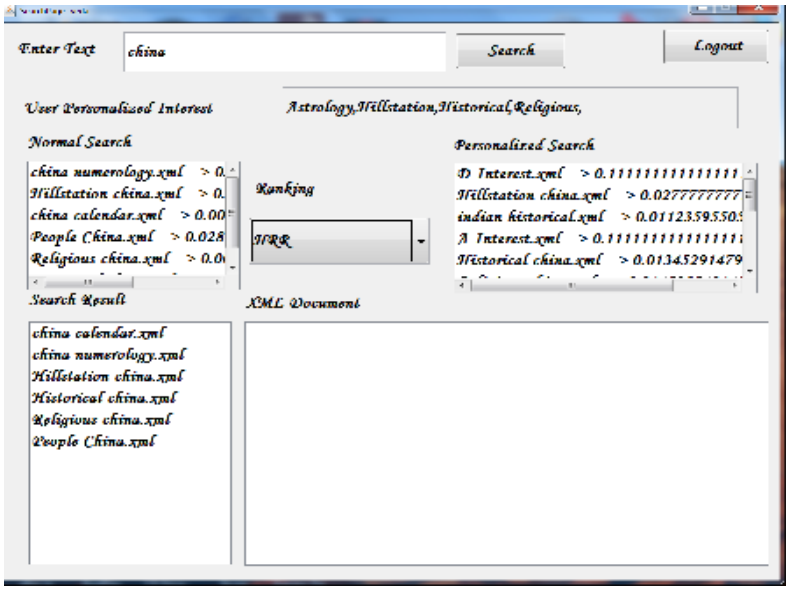

Fig 4.Search Results for HRR Ranking

The number of documents retrieved using SRR is same as the number of documents returned using normal search but the order is based on RSV of sum of preference and normal search xml documents.

The documents will be retrieved similar to SRR, but the resultant document list will also include the unmatched documents from user preference list.

VI. CONCLUSION

We have proposed an approach which will make the retrieval of XML documents more effective by making use of the personalized information. The approach involves calculating the score of the user preferences which are stored as xml document based on user query by using the ranking function of top -k algorithm. These calculated user preferences will be used in retrieving the documents. The retrieved documents will be re ranked based on HRR, SRR and IRR. In this simulation we have computed the score of user preference node which contain the search text. We have not considered the user preference nodes which do not contain the search text but has a descendant node which contains search text. In future work we would like to implement this personalized search by computing the score for nodes which do not contain the search text but has a descendant node which contains search text.

REFERENCES

1. Luis M. de Campos, Juan M. Fern´andez-Luna, Huete, and Eduardo Vicente-L´opez Juan F. "Using Personalization to Improve XML Retrieval", IEEE Transactions on Knowledge and Data Engineering, Vol 26 No.5, May 2014
2. Harshal R Aher , Anupkumar Bongale "Create XML Document and Efficient Interactive Keyword Search Technique over XML Data", IJSCE , ISSN: 2277 128X/ Volume 4, Issue 5, May 2014
3. Varun Varma Sangaraju "Ranking Of XML Documents by Using Adaptive Keyword Search", IJCSIT ,Vol. 5 (2) ,1619-1621/ISSN:0975-9646, 2014
4. Jovan Pehcevski1, Jaap Kamps2, Gabriella Kazai3, Mounia Lalmas4, Paul Ogilvie5,5 Benjamin Piwowarski6 and Stephen Robertson3 "INEX 2007 Evaluation Measures", Springer Berlin Heidelberg, Volume 4862, pp 24-33, 2008
5. Laxman Dethe, Prof. R. M. Goudar, Prof. Sunita Barve "Performance Evolution of XML Data Searching by Using Fuzzy Type a head Search", IJIRCCE, Vol. 2, Issue 11, November 2014
6. Pradeep Kumar Reddy Gade, N Prasanna Balaji, U Sreenivasulu "An Effective XML Keyword Search with User Search Intention over XML Documents" , Global Journals Inc. (USA), Volume 11 Issue 16 Version 1.0, September 2011.
7. Rutuja S .Lachake Prof. G. P. Potdar " A Survey on Personalized Search: An Web Information Retrieval System" , IJCSIT, Vol. 5 (2) , 1120-1127. 2014
8. Dr. M.Thangaraj, M. Chamundeeswari "A Survey of Agent-based Personalized Semantic Information retrieval" , IJCSIT Vol. 2, Issue 3, September 2011
9. B. Kurinchi Meenakshi Dr. C. Nalini, "Survey on Managing XML Search through Personalization", IJRITCC, Vol 3 Iss 3 ISSN: 2321-8169 ,March 2015