Multi-keyword Context-Oriented diversification search on XML Data using Map-Reduce Framework

Sneha B. Mandlik[1], Prof. I.R.Shaikh[2]

[1]Student, Department of Computer Engineering, SND COE & RC, Yeola, Maharashtra, India
[2]Professor, Department of Computer Engineering, SND COE & RC, Yeola, Maharashtra, India

ABSTRACT

Recently lots of work is to be done on searching techniques. In searching process user enter particular candidate searching keyword and with the help of searching algorithm respective searching query is executed on targeted dataset and result is return as an output of that algorithm. In this case it is expected that meaningful keyword has to be entered by user to get appropriate result set. In case of confusing bunch of keywords or ambiguity in it or short and indistinctness in it causes an irrelevant searching result. Also searching algorithms works on exact result fetching which can be irrelevant in case problem in input query and keyword. This problem statement is focused in this system. By considering the keyword and its relevant context in XML data, searching is done using automatically diversification process of XML keyword search. This system gives maximum number of results as compare to regular and diversified keyword search using synonym diversification. This synonym diversification gives top-k results having maximum relevance factor. This system reduces time required to search large XML data using HADOOP. Finally Synonym diversification search gives comparatively more results than regular and diversified search in less time. Using HADOOP efficiency of the system is improved.

Keyword: *- Candidate Keyword ,XML Keyword search, feature selection, diversification process.*

1. INTRODUCTION

Keyword based searching is the main part of research domain. The search can be applied on structured and /or semi-structured information. Keyword searching is a feature which provide data abstraction to the user. User does not need to know the exact structure and /or query language to fetch information. We are mainly focusing on keyword searching on XML data. To search for individual word or group of co-related words in a set documents and fetch the most mapped results as an output is the technique of IR[1].

A keyword search looks for words anywhere in the record. It is appeared as most effective paradigm for finding information. The advantage of keyword search is its simplicity. The most important requirement for the keyword search is to rank the results of query so that the most relevant results appear. Keyword search provides simple and user friendly query interface to access xml data in web.

Keyword search over xml is not always the entire document but deeply nested xml. Xml was designed to transport and store data. It does not do anything, it is created to structure, store, and transport information.xml document contains text with some tags which is organized in hierarchy with open and close tag. xml model addresses the limitation of html search engine i.e. Google which returns full text document but the xml captures additional semantics such as in a full text titles, references and subsections are explicitly captured using xml tags[2].

In conventional keyword-search system on XML data, a user composes a query keyword, submits it to the system, and retrieves relevant information. In the case if the user doesn’t know how to issue queries, he tries multiple queries and sees multiple times what the result is[2].

A query may contain many words or small number of uncertain keywords. When query contains small number of keywords it is very challenging problem to identify interested keywords of user and their search intension. In this scenario ambiguity is generated in query generation process. To avoid this problem it is always beneficial to involve user in search process and provide multiple options or query suggestions to the user based on the context of search input keywords. User can select choose a query based on these suggested options and can get the appropriate result.

To identify suitable results we have to first identify key-words in query. Then for each keyword extract correlated feature terms of keywords from a given XML data set based on predefined metadata and its probabilistic features .This process is similar to the feature selection. The selected feature terms is not same as the labels of XML elements. Each separate combination of the feature terms and query keywords may represents one of diversified contexts. After identifying the context of diversified query in terms of its relevance with original query and novelty of produced result we will get appropriate queries.

To work on large XML data T, our main aim is to derive top-k expanded query candidates from a given query Q with more relevance and also maximal synonym diversification where every candidate in candidate list represent the search intention of q in T.

2. LITERATURE SURVEY

In this, by considering keyword and its relevant context in XML data, searching should be done using automatic diversification process of XML keyword search is the large area of concern [1].

For structured and semi-structured data, various techniques are discussed for keyword search. In this query optimization , ranking phases , top-k main query processing is reviewed. Different data models such as XML graph-structured data is reviewed. Application of these concepts are also discussed in which keyword based searching is having main importance. Problems like Diversified Data Models, Query Formation: Complexity versus Expressive Power , Search Quality Improvement , Evaluation are also discussed [3].

XRANK system is main discussion of this paper. Ranked search technique over XML data is considered here. Space saving, performance gaining techniques like index structure and query evaluation are also focused. XRANK can help in searching for HTML and also XML documents. Disadvantage: Authors have currently taken a central view of document, where it is assumed that query results are strictly in hierarchical manner. Index maintenance is major problem for effective search and which is main blockage in search area [3].

In this SLCA-based keyword searching is discussed. In this Multiway - SLCA approach (MS) Queries are helpful to support the keyword searching at and old methods like AND / OR. Then LCA analysis improvement algorithms re used to solve search problems based on keywords [4].

Using previous query and its analysis provides perfect direction for diversification. Old query reformulation provides behavior which is exactly same with the query related of user. Client data request, its re-ranked structure and query is observed and analyzed at client side for perfect diversified output. Large query logs are resolved in this paper from search engine [5].

3. PROBLEM STATEMENT

In this system, keyword query is given as input to perform searching on XML data to get certain, exact and distinct result as a output. Main objective is to derive top-k expanded query candidates by considering its context in terms of high relevance and maximal synonym diversification for given query. To deal with big XML data to perform searching HADOOP platform is used. It gives synonym diversified results as output.

4. PROPOSED WORK

This system helps user to get relevant results for multi-keywords. This system focused on meaningful expansion of basic query by extracting feature terms by considering the context of basic query. This system first focused on to get diversified synonym of keyword query by extracting k additional words. Expanded query is used to search more specific documents. This system gives maximal results using synonym diversification keyword search. This synonym diversification search gives top results having maximum relevance factor value. It also shows comparative study of time required for searching and number of results of regular search, diversified search and synonym diversified search. Following figure shows block diagram of proposed system:

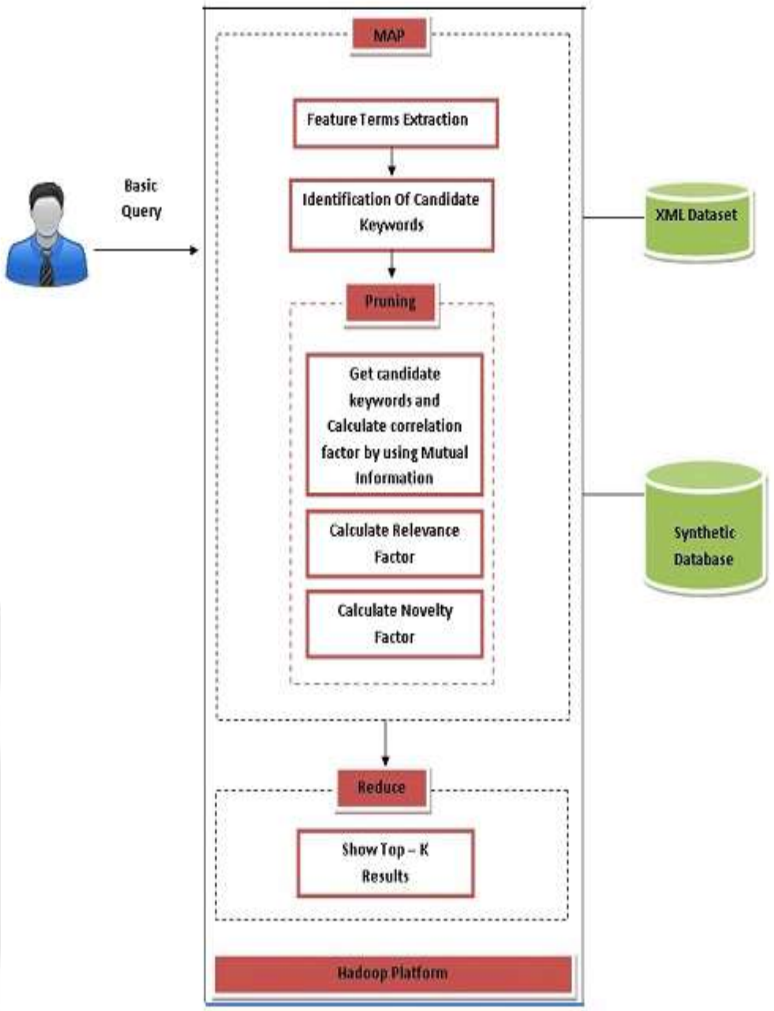
MAP
Feature Terms Extraction
Identification Of Candidate Keywords
Pruning
Get candidate keywords and Calculate correlation factor by using Mutual Information
Calculate Relevance Factor
Calculate Novelty Factor
Reduce
Show Top – K Results
Hadoop Platform
Basic Query
XML Dataset
Synthetic Database

Fig -1: System Architecture

4.1 System Description

Following is the flow of process:

1. First user query is analyzed and searching keywords are traced.
2. After finalizing the searching keywords of user, system used mutual information model and calculate the correlation values so that it will be easy to get new query keywords.
3. After finalizing the mutual information amongst the keywords , their context based relevant keywords or featured term for new query is searched over XML dataset.
4. Original keywords and fetched keywords has some common information hence their relevance factor is calculated.
5. After relevance factor calculation their novelty factor is calculated. This provides diversified result on the basis of context terms or keywords extracted.
6. After getting relevant and novelty result set , top- k results are defined[7].

After getting top- k keywords, for refinement of diversified result , there are two algorithms. One is Baseline algorithm and other is Anchor- based pruning algorithm. In Base line algorithm it first retrieve the relevant feature terms with high mutual scores from the term correlated graph of the XML data T; then generate list of query candidates that are sorted in the descending order of total mutual scores; and finally compute the SLCAs as keyword search results for each query candidate and measure its diversification score[4]. As such, the top- k diversified query candidates and their corresponding results can be chosen and returned. But in this case by analyzing the baseline solution, it is found that the main cost of this solution is spent on computing SLCA results and removing unqualified SLCA results from the newly and previously generated result sets[7].

7. Hence anchor- k based Algorithm is used for refinement of diversified result. The basic idea of this algorithm is described as follows. It generate the first new query and compute its corresponding SLCA candidates as a start point. When the next new query is generated, we can use the intermediate results of the previously generated queries to prune the unnecessary nodes according to the above theorems and property. By doing this, we only generate the distinct SLCA candidates every time. That is to say, unlike the baseline algorithm, the diversified results can be computed directly without further comparison [7].

All this calculation and getting of top- k results will be executed on Hadoop Platform.

- **Feature Selection Model:**

To extract meaningful feature terms with respect to an original keyword query ,top-k interesting and meaningful expansions to a keyword query is produced by extracting k-additional words with high specific values[6][7]. To identify feature terms from dataset we assume a XML tree T, its gives sample result set as R(T). For feature selection, mutual information is used. This must have minimum redundancy and maximum relevance. Probability of term x appears in R(T) is:

$$Prob(x,T) = \frac{|R(x,T)|}{|R(T)|}$$

Similarly Probability of terms x and y co-occurring in R(T) is:

$$Prob(x,y,T) = \frac{|R(x,y,T)|}{|R(T)|}$$

Finally mutual information is calculated as follows:

$$MI(x,y,T) = \frac{Prob(x,y,T)* \log (Prob(x,y,T))}{Prob(x,T)* Prob(y,T)} \qquad (1)$$

- **Keyword Diversification Model:**

In keyword diversification model, Relevance is calculated to get relevant result and Novelty is calculated to get new and distinct results. To include the relevance and novelty of keyword search together, it must satisfy two criteria: 1) the newly generated query qnew has maximum probability to determine the contexts of original query q with respect to data which is to be searched and 2) the generated query qnew has a maximum difference from the previously generated query set Q. So to calculate this a combined score is calculated as

$$Score(q_{new})= Prob(q_{new}|q,T) * DIF(q_{new},Q,T) \qquad (2)$$

Here, Prob(qnew |q,T) represents probability that qnew is search applied when original query q is issued over the data T which is a relevance factor and DIF(qnew ,Q,T) represents the percentage of results that are produced by qnew, but not by any old generated query in Q which is a novelty factor. By applying query 'Q' on a dataset 'D', using Base Line algorithm query candidate keywords are generated and after processing on these keywords, the top-k diversified query candidates are generated. For this two algorithms are used Baseline and Anchor based pruning algorithm.

- **Mathematical Model:**

S = {I,F,O }
Here,
I = {J,Q,S }
Set of inputs j = {j1,j2,…,jn}
set if json data objects Q = {q1,q2,..qn}
set of query words S = {s1,s2,..,sn }
set of synthetic queries database

F={F1,F2,F3,F4,F5,F6,F7,F8 }
set of functions
F1 = Feature text extraction

F2 = Identification of candidate keywords
F3 = Calculation of co-relation factor
F4 = Calculation of Relevance factor
F5 = Calculation of Novelty factor
F6 = data pruning
F7 = sort result
F8 = display top result

O = {O1,O2} set of output.
O1 = exact mapped result.
O2 = context based diversifiable mapped results.
Following figure shows functional dependency of system:

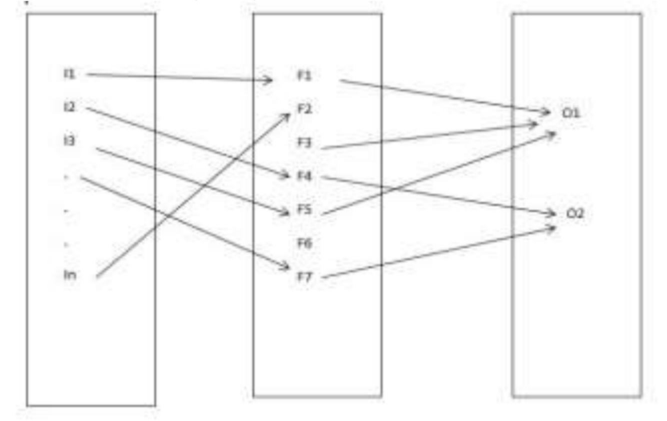

Fig -2: Functional Dependency of system

5. ALGORITHM

• **BaseLine Algorithm*:***
Baseline algorithm is used to retrieve the diversified keyword search results. Steps of baseline algorithm are as follows:
1. It first retrieve the relevant feature terms having high mutual scores with the terms in XML data T.
2. To calculate mutual score, equation (1) is used.
3. Then it generates a list of query candidates that are sorted in a descending order of total mutual scores;
4. Finally it computes the SLCAs using probability Prob(q| qnew ,T) is result of keyword searching for each query candidate and then measure its diversification score using equation(2).This top-k mixed query candidates and their identical results are output of Baseline algorithm.

• **Anchor Based Pruning Algorithm:**
Baseline algorithm takes more time in computing SCLA results. In Anchor based pruning solution, it can avoid the unneeded computational cost of unqualified SLCA results which is duplicated and ancestors. For this interrelationship between intermediate SLCA candidate is first analyzed using equation(1) and then newly generated query candidates are used to get final top-K results.

6.RESULT AND DISCUSSION

As per paper [1] feature selection model and keyword diversification model used to get feature keywords and top-k diversified result as output respectively. This diversified result has maximum novelty which gives new and distinct result. And by using HADOOP for this implementation. Using HADOOP efficiency of system is going to be improve. Now using proposed system first feature terms are extracted. For example, Query q is applied over DBLP database T as, q={system, list} over T. First feature terms related to q are extracted. Terms related to keyword system are extracted and terms related to keyword list are also extracted. This each combination of keyword and feature term is one of the diversified result.

Table -1: Feature Term Extraction

Keyword	Feature terms
parallel computing	Analog, analogue, calculate, cipher comput, comput network, comput science, comput parallel, parallel integ, parallel sort, twin
Keyword database	Database, database relate, keywords database

After getting feature terms corresponding regular search, diversified search and synonym diversified search is performed. It is shown with and without HADOOP. With HADOOP this comparative study is shown in this system. It shows a graph having number of keywords versus execution time (t). It shows time required to search feature term in database.

Following graph shows execution time required for regular search, diversified search, synonym diversified search without HADOOP.

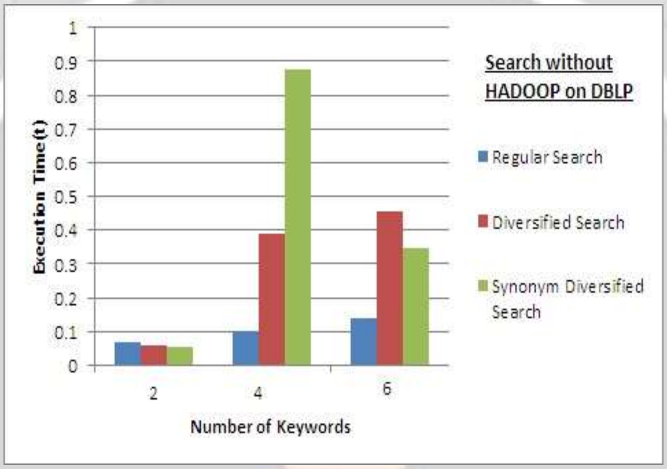

Chart -1: Searching without HADOOP on DBLP dataset.

Following graph shows execution time required for regular search, diversified search, synonym diversified search with HADOOP. It shows that time required to search on HADOOP is less that searching without HADOOP.

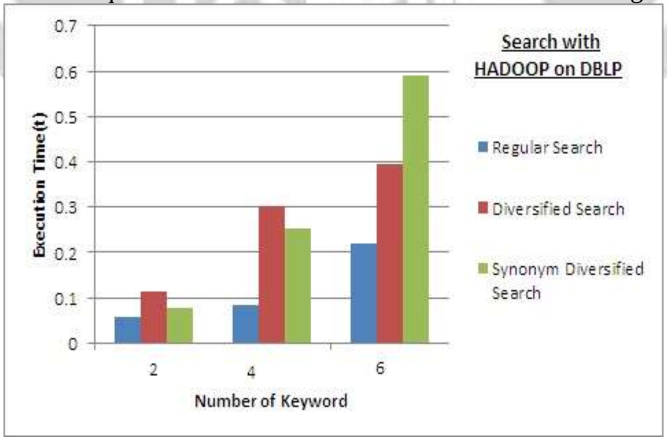

Chart -2: Searching with HADOOP on DBLP dataset.

Following graph shows comparative study of execution time required for regular search, diversified search, synonym diversified search without HADOOP and with HADOOP. It shows that time required to search on HADOOP is less that searching without HADOOP.

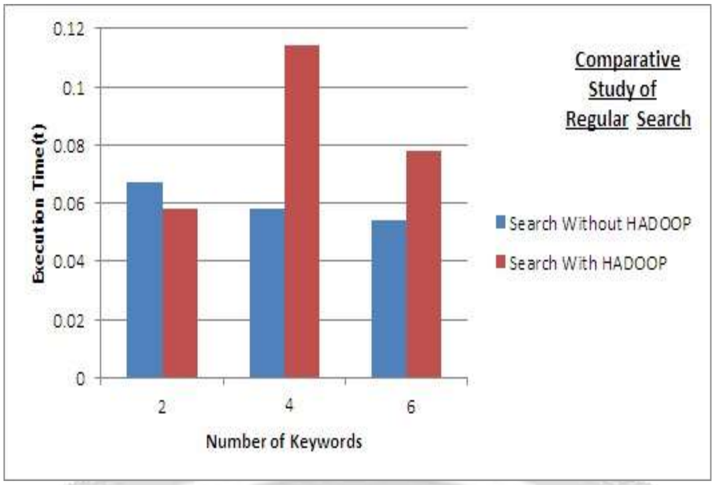

Chart -3: Comparative study of regular Searching without HADOOP and with HADOOP on DBLP dataset.

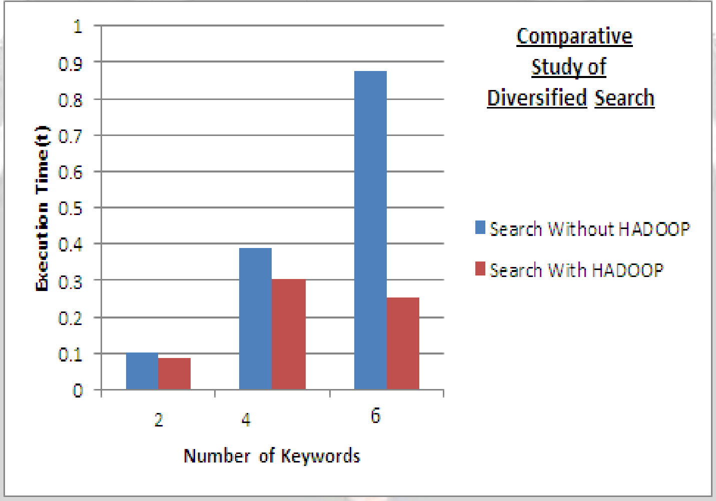

Chart -4: Comparative study of Diversified Searching without HADOOP and with HADOOP on DBLP dataset.

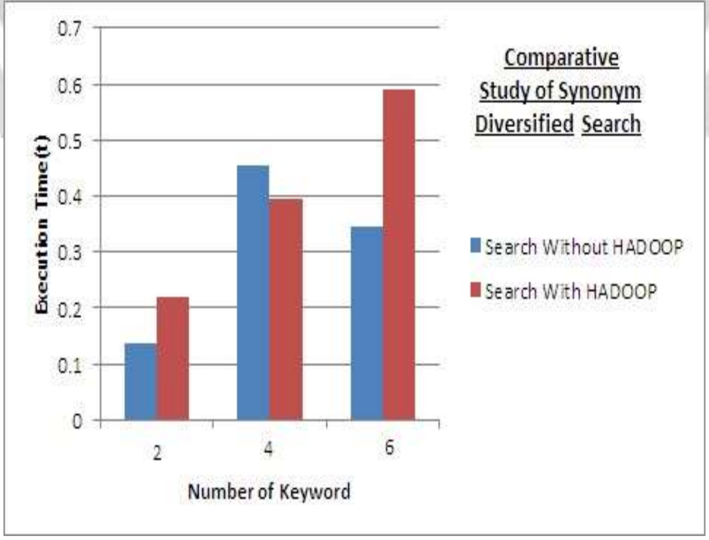

Chart -5: Comparative study of Synonym Diversified Searching without HADOOP and with HADOOP on DBLP dataset.

Diversified Search:

Diversified search is a searching which gives result set which contains result which are very relevant elements to the query and at the same time, as diverse as possible to other ones in the result set R. When we search on HADOOP, it gives 3 types of results: 1.Regular search result 2. Diversified search result 3. Synonym diversified search. Out of this 3 searching technique synonym diversification gives more number of results than number of results in regular search and diversified search, as number of searching keywords increases in synonym diversification.

Following table shows number of results obtained in diversified search and Synonym diversified search:

Table -2: No. of results in Diversified search and Synonym diversified search

Keyword	**No. of results in diversified search**	**No. of results in synonym diversified search**
parallel computing	8	16
Semi-structured database system	26	31

After analysis of this system it is concluded that this system gives comparatively more results in less time than regular search and diversified search.

This system is also used for searching over unstructured and semi-structured data. It performs searching over text documents also. To search on text documents it used Enron dataset. The size of this dataset is 1.7GB. This dataset is a collection of email text files.

7. CONCLUSION

In this system the main focus is on the searching over large XML dataset and provide a synonym diversified result forms given keyword query based on the context of query keywords. It gives new and distinct result set as output. This proposes an effective solution that provides efficiency in searching process by distributing its work on HADOOP. It shows comparative study between regular search, diversified search and synonym diversified search with relevance score. It also shows time required for execution of regular search, diversified search and synonym diversified search. From this it is analyzed that the number of results in synonym diversified search is greater than regular search and diversified search. The number of results in regular search and diversified search are less as compare to synonym diversified search.so the time required for such large data searching in synonym diversification is comparatively less than time required to search small number of results in regular and diversified search. Also this system performs searching over unstructured and semi-structured data also.

ACKNOWLEDGEMENT

I would like to thanks all people who have helped us directly or indirectly for the paper presentation. I am grateful to my guide, **Prof. I.R.Shaikh,** for his guidance and support. I wish to express my sincere thanks to the PG Coordinator **Prof.V.N.Dhakane** and the departmental staff members for their support. Lastly I would like to thank Department of Computer, SND COE & RC, Yeola for making all the requirements possible and simply available whenever required

REFERENCES

[1] Jianxin Li, Chengfei Liu , "Context-Based Diversification for Keyword Queries Over XML Data" in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL. 27, NO. 3, MARCH 2015.

[2] Supriya C. Rathod, Sonali M. Tidke, "Survey on Interactive Keyword Search Over XML Data to Obtain Top-K Results",IJAR,vol.4Issue.3, March 2014, ISSN.555X,pp.158-160.

[3] Y. Chen, W. Wang, Z. Liu, and X. Lin, "Keyword search on structured and semi-structured data," in Proc. SIGMOD Conf., 2009 pp. 1005–1010.

[4] L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram, "Xrank:Ranked keyword search over xml documents," in Proc. SIGMOD Conf., 2003, pp. 16–27.

[5] C. Sun, C. Y. Chan, and A. K. Goenka, "Multiway SLCA-based keyword search in xml data," in Proc. 16th Int. Conf. World Wide Web, 2007, pp. 1043–1052.

[6] F. Radlinski and S. T. Dumais, "Improving personalized web search using result diversification," in Proc. SIGIR, 2006, pp. 691–692.

[7] Ms.Sneha B. Mandlik, Prof.Santosh Durugkar,"A Review on Multi-keyword Context-Oriented diversi_cation search on Map-Reduce Framework over XML Data",in Proc. IJCSIT Volume 6 Issue 6 ISSN:0975-9646, PP.5145-5147, International.

[8] Y. Xu and Y. Papakonstantinou, "E_cient keyword search for smallest lcas in xml databases", in Proc. SIGMOD Conf., 2005, pp. 537-538.

[9] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results",in Proc. 2nd ACM Int. Conf. Web Search Data Mining, 2009, pp. 514.

[10] H. Chen and D. R. Karger, "Less is more: Probabilistic models for retrieving fewer relevant documents", in Proc. SIGIR, 2006, pp. 429-436.

[11] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher,and I. MacKinnon, "Novelty and diversity in information retrieval evaluation", in Proc. SIGIR, 2008, pp. 659-666.

[12] Z. Liu, P. Sun, and Y. Chen, "Structured search result di_erentiation", J. Proc.VLDB Endowment, vol. 2, no. 1, pp. 313-324, 2009.

[13] E. Demidova, P. Fankhauser, X. Zhou, and W. Nejdl, "DivQ:Diversi_cation for keyword search over structured databases", inProc. SIGIR, 2010, pp. 331?338.

[14] N. Sarkas, N. Bansal, G. Das, and N. Koudas, "Measure-driven keyword-query expansion", J. Proc. VLDB Endowment, vol. 2,no. 1, pp. 121?132, 2009.

[15] N. Bansal, F. Chiang, N. Koudas, and F. W. Tompa, "Seeking stable clusters in the logosphere", in Proc. 33rd Int. Conf. Very Large Data Bases, 2007, pp.806?817.

[16] S. Brin, R. Motwani, and C. Silverstein, "Beyond market baskets:Generalizing association rules to correlations" in Proc. SIGMOD Conf., 1997, pp. 265?276.

[17] W. DuMouchel and D. Pregibon, "Empirical bayes screening for multi-item associations", in Proc. 7th ACM SIGKDD Int. Conf.

BIOGRAPHIES

	Ms.Sneha B. Mandlik receive the B.E. degree in Information Technology from MET BKC IOE, Nashik in 2012 and currently pursuing her Masters degree in Computer Engineering from S.N.D. College of Engineering and Research Centre, Savitribai Phule Pune University Former UOP. This paper is published as a part of the research work done for the degree of Masters.
	Prof. I. R.Shaikh is an Head of Department in Department of Computer Engineering, S.N.D. College of Engineering and Research Centre, Savitribai Phule Pune University.His current research interest is in data mining.