ISSN:0975-9646

A Review on Multi-keyword Context–Oriented diversification search on Map-Reduce Framework over XML Data

Ms.Sneha B. Mandlik[#1], Prof.Santosh Durugkar[#2]

[#1] *Computer Engineering Department, Savitri Bai Phule University,Pune,India,*

[#2] *Computer Engineering Department, Savitri Bai Phule University,Pune,India,*

***Abstract*— 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 should be done using automatically diversification process of XML keyword search. In this way system may satisfy user, as user gets the analytical result set based on context of searching keywords. For more efficiency and to deal with big data, HADOOP platform is used.**

***Keywords*— Candidate Keyword ,XML Keyword search, feature selection, diversification process.**

I. INTRODUCTION

Keyword based searching is the important part of research domain. The search can be applied on structured and /or semi-structured information. The keyword search feature provide data abstraction to the user i.e. user do no need to know the exact data structure and /or query language to fetch informa-tion. We are mainly focusing on keyword search over XML data. To search for particular 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(information retrieval).

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

To identify appropriate results we have to first identify keywords in query. Then for each keyword extract the co-related feature terms 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 individual combination of the feature terms and query keywords may represents one of diversified contexts . After analysing 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 with large xml data T, our basic aim is to derive top-k expanded query candidates from a given query Q with more relevance and maximal diversification where every candidate in candidate list represent the search intention of q in T.

II. LITERATURE SURVEY

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

In this for structured and semi-structured data, various state-of-the-art techniques are discussed for keyword search. In this query optimization , ranking phases , top k important query processing is discussed. Different data models such as XML , graph-structured data is discussed. Application of these concepts are also discussed in which keyword based search is having prime importance. In this paper some problems like Diverse Data Models, Query Forms: Complexity versus Expressive Power , Search Quality Improvement , Evaluation are also discussed [2].

XRANK system is discussed in this paper. Ranked search technique over XML data is considered here. In this paper space saving and performance gaining techniques such as index structure and query evaluation are also focused. XRANK can help in searching for HTML as well as XML documents.
Disadvantage: For instance, authors have currently taken a document-centric view, where they assume that query results are strictly hierarchical.
Index maintenance is major problem for effective search and which is bottleneck area [3].
In this SLCA-based keyword search approach is discussed. Queries called the Multiway - SLCA approach (MS) is helpful to promote the keyword search beyond and old methods like AND / OR. After LCA analysis improved algorithms are put to solve search problems based on keywords [4].

In this Indexed Lookup Eager and Scan Eager, algorithms are discussed. XML search based on keyword according to SLCA semantics is prime topic of discussion and for this these algorithm are used. Instant search result is the beauty of theses algorithm. XKSearch architecture implementation is discussed in it. The XKSearch system inputs a list of keywords and returns the set of Smallest Lowest Common Ancestor nodes [5].

Query and information relevance is calculated so that unnecessary checks are avoided and effective search is achieved. Hence effective text retrieval and summarization is achieved. The Maximal Marginal Relevance (MMR) achieves the stopping of redundancy. This approach provides very much relevant data in terms of search result to the end user by effectively minimizing the redundancy [6].

In this paper Risk of dissatisfaction of user is major area of concern. To minimize it systematic approach to diversifying results is discussed in it. For this several techniques such as NDCG, MRR, and MAP are discussed in detail in it. A Greedy Algorithm for Diversification used in it. Among the search result user should find most relevant data is the aim of diversification. Also another aim of this paper is to minimize the rank of best fitted result [7].

This paper also uses greedy approach. Different datasets are considered in this to get approach tested thoroughly and relevant document in terms of search result is expected as search result [8].

In this using test collection based on TREC question answering track this paper discussed the framework which achieves novelty and diversity. In this approach document is linked with the relevant information in it. Chunk of information is in this way get attached with document and which is helpful in at time of search. This piece of information is having content as well as document properties. The major drawback of this approach is that unusual features of document may cause judging error. Some raw data related with the document may delayed the search result [9].

Using past query and its analysis provides proper direction for diversification. Past query reformulation provides exact query related behaviour of user. Client data request, his re-ranked structure and query is observed and analysed at client side for proper diversified result. Large query logs are analyzed in this paper from search engine [10].

In this single swap and multi swap algorithms are used in this paper. On structured data differentiation of search results is carried out. Degree of difference is quantified so that it represents the accuracy of search result. Features from the search result is traced and this result is prominently considered in calculation [11].

In this by considering query result and its redundancy , new scheme named re-ranking query interpretations is discussed to diversify the search result. For sub-topics and relevance new proposed technique such as propose α-n DCG-W and WS-recall is promoted in it. Algorithm named as Diversification algorithm is used in it. For database query search query similar measure and greedy algorithm is used to obtain diversified query interpretation and its relevance [12].

Progressive refinement of keyword-query result set is discussed in this paper. New data analysis model and exploration model is used for it. Efficient framework based on Convex Optimization principles is used for expansions of the original query with some add on terms in it. This query expansion helps to get relevant results if and only if relevant terms are introduces in it [13].

Keyword clusters are developed in it. Algorithm to get correlated keyword is used in it. After finalizing the keywords cluster is formed and finalize. Finally stable keyword cluster is finalize. Temporal association of sets of keywords related problem is traced in it as far as blogosphere is concern. Keyword clusters and identifying stable cluster is the major area of interest [14].

Correlation rules are generalized here. Here dynamic nature of terms interrelation with each other is achieved which is beyond the current trends and market standard settings[15].

Three major variations are discussed in this paper. In first part interrelated item sets are considered and strategy is designed accordingly about transaction. In the second part empirical Bayes model is used to eliminate the recommended minimal support size. In the final section they built on earlier work that considers interestingness measures that assess departures of observed frequencies from baseline frequencies[16].

III. PROPOSED SYSTEM

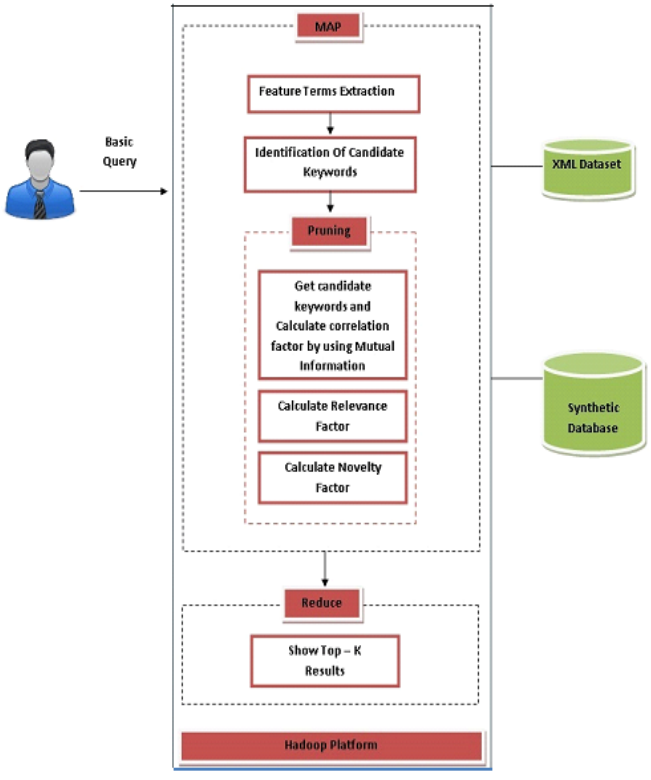

Figure 1: Block diagram of proposed system.

There is need of one system that will provide a diversifiable search over large dataset in minimal time. We introduce a HADOOP platform to work with big data. The process of system can be described as : For a given query q With the help of feature selection process this system first

finalizes and fetches the candidate keywords from the query. Then system validates and checks the quality of fetched candidate keywords for searching purpose with the help of XML keyword search diversification model. After validating the candidate keywords, using effective algorithms selective keywords are referred (top-k keywords are referred) this decides the diversified search objective. Rule is decided that these selected candidate keywords must be most relevant and should cover good amount of distinct results. In this way system may satisfy user as he gets the analytical result set based on context of searching keywords. Hence variety of results for the same context is the feature of this system.

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 to get top-k worth and meaningful expansion to a keyword query by extracting k additional words. Expanded query is used to search more specific documents.

Following is the flow of process:

A] First user query is analysed and searching keywords are traced

B] 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.

C] After finalizing the mutual information amongst the keywords , their context based relevant keywords or featured term for new query is searched over XML dataset

D] Original keywords and fetched keywords has some common information hence their relevance factor is calculated.

E] After relevance factor calculation their novelty factor is calculated. This provides diversified result on the basis of context terms or keywords extracted

F] After getting relevant and novelty result set , top – k results are defined

IV. CONCLUSIONS

In this paper we mainly focusing on the search approach over large xml dataset and provide a diversified result form given keyword query based on the context of query keywords. We have studied multiple journal papers related to this domain and identifies its need and limitations. We propose an effective solution that provides efficiency in search process by distributing its work on HADOOP platform.

ACKNOWLEDGMENT

We are glad to express our sentiments of gratitude to all who rendered their valuable guidance to us. We would like to express our appreciation and thankful to Prof. S.R.Durugkar, Head of Department, Computer Engineering., S.N.D. College of Engineering. and Research Center, Nashik. We thank the anonymous reviewers for their comments.

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] 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.

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

[4] 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.

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

[6] J. G. Carbonell and J. Goldstein, "The use of MMR, diversitybased reranking for reordering documents and producing summaries,"in Proc. SIGIR, 1998, pp. 335–336.

[7] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in Proc. 2nd ACM Int. Conf. Web Search Data Mining, 2009, pp. 5–14.

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

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

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

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

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

[13] 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.

[14] 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.

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

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

AUTHORS

Ms.Sneha B. Mandlik received the B.E. degree in Information Technology from MET BKC IOE,Nashik in 2012. She is 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 paperis published as a part of the research work done for the degree of Masters.

Prof. S. R. Durugkar 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.