Keyword Search Diversification Model Over XML Data

Miss. Nita D.Bankar[1] Prof.A.N.Nawathe[2]
[1]M.E. Student [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]Amrutwahini COE, Sanmahner, Maharashtra, India

Abstract— To search large amount of data keyword searching technique would be more beneficial. Due to ambiguous nature certain barriers are occurred with query searching. There may also uncertain problems of keyword diversification that is not capable of answering correct solution in searching expected results from large collection of data. Whereas, in XML data search some short and vague keyword query is used. It dynamically diversifies keyword searching. The complications of diversifying keyword search are essentially calculated in IR association. To better the attention of query diversification in formed databases or semi structured data, it is charming to consider both formed and content of data in diversification paradigmatic.

Key words: XML; keyword search; context-based diversification

I. INTRODUCTION

KEYWORD search on structured and semi-structured information has appeal to much research interest just now, as it enables users to fetch information without the need to learn experienced query languages and database structure [1]. Co-related with keyword search methods in information retrieval (IR) that wish to find a list of consistent documents, keyword search techniques in structured and semi-structured data (denoted as DB and IR) focused more on definite information contents. Especially, if a node is an SLCA, then its ancestors will be absolutely excluded from being SLCAs, by which the nominal information content with SLCA definition can be used to represent the specific results in XML keyword search. The well-confirmed SLCA definitions [2], [3], [4], [5] as a result metric of keyword query over XML data. Almost, the more keywords user's query contains, the easier the user's search aims with commendation to the query can be analyze. After all, when the given keyword query only have a small number of uncertain keywords, it would turn into a very challenging problem to acquire the user's search intention due to the high ambiguity of this type of keyword queries. Even if sometimes user absorption is helpful to examine search goal of keyword queries, a user's collective process may be time-consuming when the size of significant result set is wide. The complications of diversifying keyword search are essentially calculated in IR association [6], [7], [8], [9], [10]. Most of them achieve diversification as a post-processing or re-ranking step of document fetched based on the search of result set and/or the query block. In IR, keyword search diversification is arranged at the point or document level. For e.g., Agrawal et al. [7] model user aims at the modern level of the taxonomy and Radlinski and Dumais [11] obtain the possible query design by mining query blocks. Even though, it is not always simpler to get this useful genetics and query logs. In extension, the diversified results in IR are often shaped at document levels. To better the attention of query diversification in formed databases or semi structured data, it is charming to consider both formed and content of data in diversification paradigmatic. So the complexity of keyword search diversification is essential to be reevaluating in structured databases or semi structured data. Liu et al. [12] is the first work to measure the diversity of XML keyword search results by analyzing their feature sets. However, the collection of feature set in [12] is finite to metadata in XML and it is also a method of post-process search result reasoning.

II. LITERATURE REVIEW

Y. Chen, W. Wang, Z. Liu, and X. Lin [1], represents techniques that support a keyword search on structured as well as semi-structured data. It contains query result definition, top-k query processing, snippet generation, result clustering, query cleaning, performance optimization, and search quality evaluation. This system provides overview of state art methods. This system contains XML data, graph structures as well as data streams etc. this system searches keyword from selected database. This system addresses various problems such as, Diverse Data Models Query Forms, Search Quality Improvement Evaluation.

L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram [2] designed a XRANK system. This system searches keywords over XML documents. This system outputs a specialized index structures as well as a query evaluation techniques. It provides significant space savings and performance gains. XRANK is generally search on HTML search engines i.e. Google. The query of XRANK can mixed with HTML and XML documents. In this system, for keyword search structured as well as semi-structured data documents must have normalized form.

C. Sun, C. Y. Chan, and A. K. Goenka[3] proposed a new approach called as SLCA-based keyword search query. This approach is useful to support keyword search over traditional AND semantics. It contains both AND as well as OR operators. This system also analyzed attributes of LCA computation. It also proposed algorithms to solve the problem in traditional keyword search. This system manages keyword search containing integration of AND and OR boolean operators.

Y. Xu and Y. Papakonstantinou[4] suggests two algorithms namely, Indexed Lookup Eager and Scan Eager, to search keyword from XML documents. It searches the keyword based on SLCA semantics. This algorithms work fast to produced quick result of query search. This system implements architecture of XKSearch. This system required a list of keywords in the form of input and it produced set of Smallest Lowest Common Ancestor nodes,

J. G. Carbonell and J. Goldstein [5] suggest a method for integrating query-relevance with information-novelty. MMR i.e. Maximal Marginal Relevance minimizes redundancy in query relevance and it also maintains re-ranking retrieved documents. This model allow user to provide information to end user by allowing user to reduced redundancy.

R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong[6]proposed systematic approach for altering results for reducing the risk of unsatisfied average user. This system generalizes several classical IR metrics. It contains NDCG, MRR, and MAP, to obvious account for the value of diversification. This system used a Greedy Algorithm for Diversification. In this system diversification can be viewed as a reactionary metric. It tends to increase the probability of average user to find some useful information from the search results.

H. Chen and D. R. Karger [7] optimized a probabilistic approach to evaluate retrieved information. It is appropriate to rank the keywords. This system used greedy algorithm for searching TREC queries. It results into standard approach to base on probabilistic principle of ranking.

C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova et al[8]represent a systematic evaluation of novelty and diversity. This system uses TREC question answering track to demonstrate feasibility of it. This system tends to express their ideas into coherent foundations to manage the redundancy. Experimental result of this system has limited scope. In this system nDCG only represents possible paths.

F. Radlinski and S. T. Dumais [9] suggested a number of methods to gathered diverse results. It is required for a given query. For that it uses a past query reformulations. This system focused on alternative approach that runs whole client-side. In this client requests a huge number of search results as well as re-ranks them in such way that documents are more likely to interest the user are presented higher. In this system diversification methods obtain query re-formulations. In this system authors analyzed large number of samples of the query logs from a popular web search engine over about 6 weeks.

Z. Liu, P. Sun, and Y. Chen [10], proposed a technique for comparison as well as differentiation of structured search results. In this system algorithm required input as, a set of structured result. DFS i.e. Differentiation Feature Set is highlights their difference bounds. This system initiates a problem of differentiate search results. In this they defined XRed systems. This system used two optimality algorithms known as, single-swap and Multi-swap algorithm. This system selects an attribute from tables.
E. Demidova, P. Fankhauser, X. Zhou, and W. Nejdl[11], proposed an approach to search a results for the diversification of structured data. In this system author proposed query similarity search using greedy search. This system mainly focused on interpretation of non-empty result set. This system also proposed α- nDCG-W and WS-recall, an adaptation of α-nDCG and S-recall.

N. Sarkas, N. Bansal, G. Das, and N. Koudas[12] introduced a new data analysis and exploration model. It enables the continuous clarification of a keyword-query result set. This process is forwarded by implying development of the original query with additional search terms and it is supported by an efficient framework, grounded on Convex Optimization principles. This system is implemented based on Full Materialization, No Materialization and Partial Materialization

N. Bansal, F. Chiang, N. Koudas, and F. W. Tompa [13] present fast algorithms for the identification of sets of associated keywords i.e. keyword clusters in the blogosphere at any specified temporal interval. This system formalizes as well as represents algorithms for the notion of temporal stable keyword clusters. This system gives solution for the problem that are related to the temporarily associations of keyword sets. This system consists of generating keyword clusters, and identifying stable clusters.
S. Brin, R. Motwani, and C. Silverstein [14] introduced a generalization of association rules, known as correlation rules. This is particularly useful for applications beyond the standard market basket setting.

W. DuMouchel and D. Pregibon[15]"Unusually frequent" involves estimates of the frequency of each item set divided by a baseline frequency computed as if items occurred independently. The focus is on obtaining reliable estimates of this measure of interestingness for all item sets, even item sets with relatively low frequencies. This system represents three variations on the market basket problem that are drawn from statistical considerations. Finally this system built on earlier work that considers interestingness measures that assess departures of observed frequencies from baseline frequencies.

R. L. T. Santos, J. Peng, C. Macdonald [16]suggested, xQuAD, a novel framework for search result diversification that builds such a diversified ranking by explicitly accounting for the relationship between documents retrieved for the original query and the possible aspects underlying this query, in the form of sub-queries. This system evaluates the effectiveness of xQuAD using a standard TREC collection.

III. PROBLEM STATEMENT

In case of ambiguous keyword input for searching may cause an irrelevant result and also searching intentions are not matched with output. Hence this system framed a process having context based diverse keywords query suggestions to user. This context is of keywords provided by user for searching. System suggests the candidate keywords and user is having decision power to use suggested keywords or reject the keywords for searching. In case of multi-keyword search if keywords are having common mutual information then result will be having repetitive data which may frustrate user.

IV. SYSTEM ARCHITECTURE

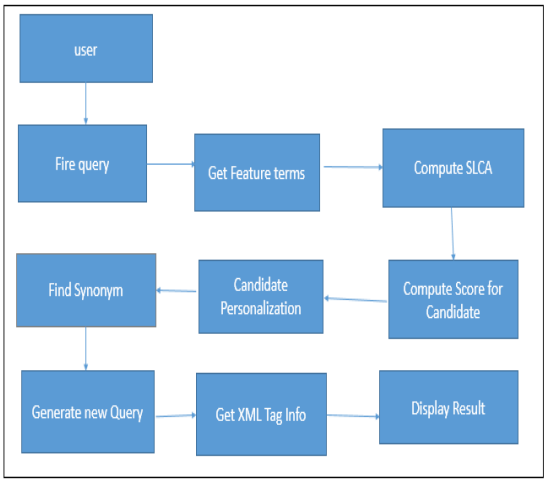

Fig. 1: System Architecture

Above is the system architecture diagram of our proposed system. It also depicts the working flow of system. To get access for the system user have to login to the system with valid credentials.

1) Login: To login the system user needs valid credentials i.e. username and password.
2) Upload XML Dataset: After successful login user comes to control panel. Firstly user have to upload XML documents i.e. DBLP dataset. Dataset is mandatory to perform keyword search. We have tested our system on DBLP dataset.
3) Fire Query: In this module user gives input to the system, user input is nothing but a phrase or keywords. With the given input query user also has to specify search type i.e. regular search, synonym search or diversified search.
4) Get feature terms: When user specify search type over a selected dataset. System proceeds for specified search i.e. regular, diversified, synonym search and display feature terms i.e individual keywords, synonym, or diversified result.

V. ALGORITHM

A. *Algorithm: Anchor-Based Pruning Algorithm*

Input: A query q with n keywords,
XML data T
and its term correlated graph G.
Output: Top-k query intentions Q
and the whole result set Φ.
Steps:

1) Calculate Matrix of feature terms M_{m*n} from q and G
2) For K feature terms
3) Travel Graph G
4) Read Node List
5) Evaluate probability of each node pi
6) Read all nodes from a selected partition and generate Phrase P
7) Compute SLCA for P
8) Remove redundant terms
9) Remove anchor node from Φ
10) Evaluate probability score S for P
11) If score >Threshold T
12) Put P in Qnew
13) Add anchor node in Φ
14) If score > Prev Q
15) Replace PrevQ by Qnew
16) Return Qnew and result set Φ

VI. MATHEMATICAL MODEL

S = {I, F, O}
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

VII. EXPERIMENTAL RESULTS

Keywords in Search Query	Keyword Count For Diversified Result	Keyword Count For Synonym Diversified Result
2	4	12
4	23	23
6	46	48
8	57	64
10	64	82

Table 1: Keyword in search query

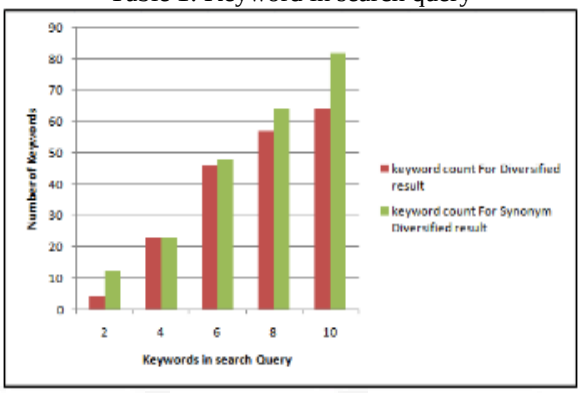

Fig. 1: Graph of keyword in search query

Keywords in Search Query	Regular Search Without Hadoop(Time in Milli Sec.)	Regular Search With Hadoop(Time in Milli Sec.)
2	10141	6468
4	13553	9353
6	15862	9821
8	16175	12853
10	16912	14682

Table 2: Regular search query with and without hadoop

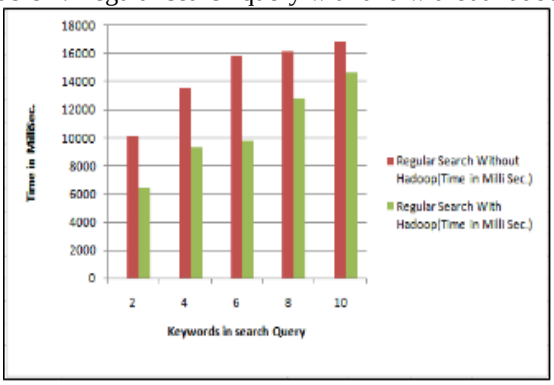

Fig. 2: Graph of regular search query with and without hadoop

Keywords in Search Query	Diversified Search Without Hadoop(Time in Milli Sec.)	Diversified Search With Hadoop(Time in Milli Sec.)
2	14149	8864
4	17593	10823
6	20869	11945
8	22073	13546
10	23017	15473

Table 3:Diversified search result without and with hadoop.

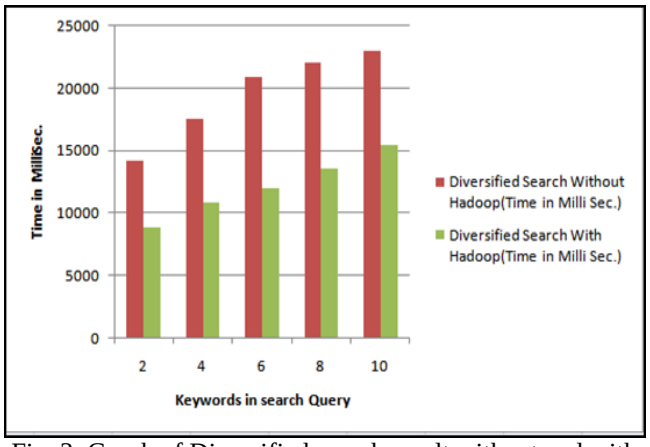

Fig. 3: Graph of Diversified search result without and with hadoop

Keywords in Search Query	Synonym Diversified Search Without Hadoop(Time in Milli Sec.)	Synonym Diversified Search With Hadoop(Time in Milli Sec.)
2	14572	1049
4	18907	12468
6	21875	13460
8	22662	14279
10	24186	15935

Table 4: Synonym diversified search with and without hadoop

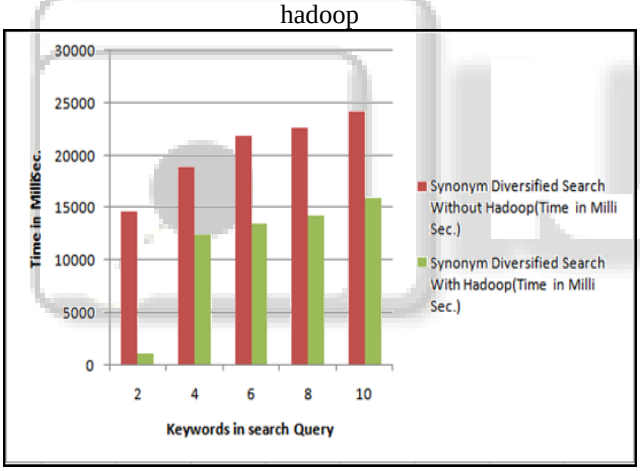

Fig. 4: Graph of synonym diversified search with and without hadoop

Keywords in Search Query	Regular Search Without Hadoop(Time in Milli Sec.)	Diversified Search Without Hadoop(Time in Milli Sec.)	Synonym Diversified Search Without Hadoop(Time in Milli Sec.)
2	10141	14149	14572
4	13553	17593	18907
6	15862	20869	21875
8	16175	22073	22662
10	16912	23017	24186

Table 5: comparative analysis 1

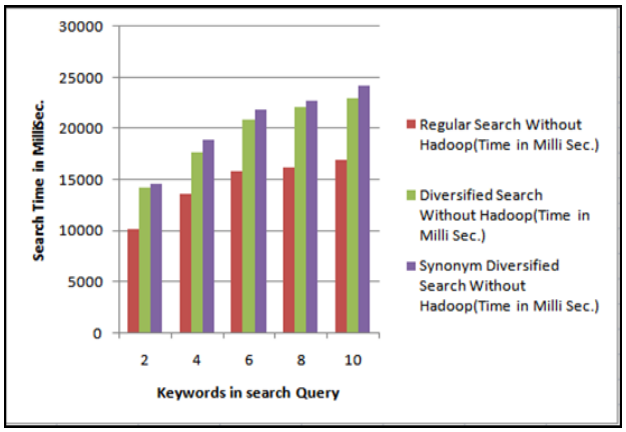

Fig. 5: Graph of comparative analysis 1

Keywords in Search Query	Regular Search Results	Diversified Search Results	Synonym Diversified Search Results
2	5	8	12
4	10	23	23
6	13	36	43
8	2	153	283
10	1	73	93

Table 6: comparative analysis 2

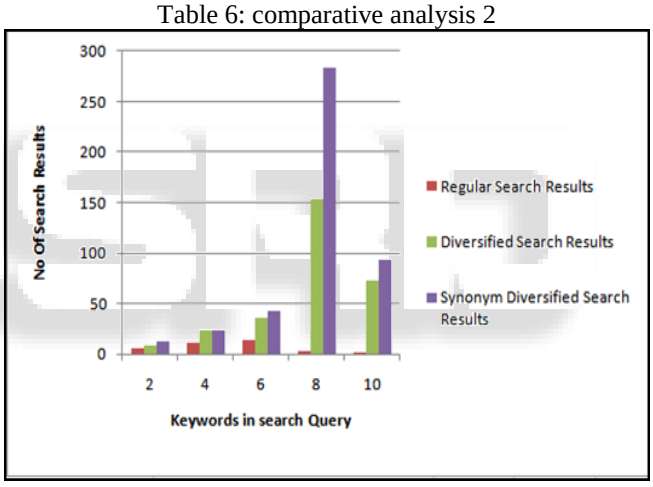

Fig. 6: comparative analysis 2

Keywords in Search Query	Regular Search With Hadoop (Time in Milli Sec.)	Diversified Search With Hadoop (Time in Milli Sec.)	Synonym Diversified Search With Hadoop(Time in Milli Sec.)
2	6468	8864	1049
4	9353	10823	12468
6	9821	11945	13460
8	12853	13546	14279
10	14682	15473	15935

Table 7: comparative analysis 3

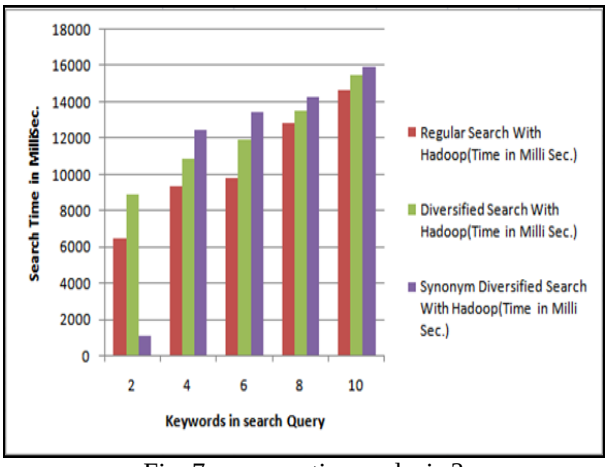

Fig. 7: comparative analysis 3

VIII. CONCLUSION

In this paper, we commence a legal study of the diversification problem in XML keyword search, which can directly measure the diversified results externally fetching all the related candidates. Against this goal, given a keyword query, we first assume the co-related feature terms for each query keyword from XML data based on collective information in the anticipation theory, which has been used as a proof for feature selection.

ACKNOWLEDGMENT

It gives me an immense pleasure to express my honest and Heartiest gratitude towards my guide Prof. ANURADHA N. NAWATHE for guidance, inspiration, honest support and affection during the course of my work. I am especially thankful of their willingness to listen and guide me to discover the finest results, regardless of the challenge. This work is also the result of the blessing guidance and support of my parents and family members and friends. I am also thankful to all who have contributed indirectly and importantly in words and deeds for completion of this work.

REFERENCES

[1] 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.
[2] L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram, "Xrank:Ranked keyword search over xml documents," in Proc. SIGMOD Conf., 2003, pp. 16–27.
[3] 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.
[4] Y. Xu and Y. Papakonstantinou, "Efficient keyword search for smallest lcas in xml databases," in Proc. SIGMOD Conf., 2005, pp. 537–538.
[5] 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.
[6] 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.
[7] H. Chen and D. R. Karger, "Less is more: Probabilistic models for retrieving fewer relevant documents," in Proc. SIGIR, 2006, pp. 429–436.
[8] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Bє€uttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," in Proc. SIGIR, 2008, pp. 659–666.
[9] F. Radlinski and S. T. Dumais, "Improving personalized web search using result diversification," in Proc. SIGIR, 2006, pp. 691–692.
[10] Z. Liu, P. Sun, and Y. Chen, "Structured search result differentiation," J. Proc. VLDB Endowment, vol. 2, no. 1, pp. 313–324, 2009.
[11] E. Demidova, P. Fankhauser, X. Zhou, and W. Nejdl, "DivQ:Diversification for keyword search over structured databases," inProc. SIGIR, 2010, pp. 331–338.
[12] 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.
[13] 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.
[14] S. Brin, R. Motwani, and C. Silverstein, "Beyond market baskets:Generalizing association rules to correlations," in Proc. SIGMOD Conf., 1997, pp. 265–276.
[15] W. DuMouchel and D. Pregibon, "Empirical bayes screening for multi-item associations," in Proc. 7th ACM SIGKDD Int. Conf.
[16] R. L. T. Santos, J. Peng, C. Macdonald, and I. Ounis, "Explicit search result diversification through sub-queries," in Proc. 32nd Eur. Conf. Adv. Inf. Retrieval, 2010, pp. 87–99.

