International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 32 No. 1 Aug. 2017, pp. 1-4

http://www.ijisr.issr-journals.org/

An Overview of Uniqueness and Novelty of XDMA for Data-Centric XML Datasets

S. Selvaganesan and G.V. Shrichandran

Department of Information Technology, J.J. College of Engineering and Technology,
Tiruchirappalli-620009, Tamil Nadu, India

ABSTRACT: In XML keyword search, the exact detection of user's intention while searching, and grading of the result in the existence of query keyword ambiguities have been difficult problems. In recent times, many keyword search approaches for XML databases have been developed to resolve these problems. XML keyword search using Dual indexing and Mutual summation based Algorithm (XDMA) is one among the prominent keyword search approaches for data-centric XML Datasets. Also, it is proved that XDMA is more effective in keyword search for data-centric XML datasets. In this paper, we present precisely the uniqueness and novel features of XDMA in comparison with other keyword search approaches for XML databases.

KEYWORDS: XDMA, Keyword Search, XML Databases, Data-Centric XML Datasets.

1 INTRODUCTION

In recent past, a number of keyword search approaches for XML datasets have been designed and developed. Among these approaches, XML keyword search using Dual indexing and Mutual summation based Algorithm (XDMA) is a notable keyword search approach for data-centric XML datasets [1], [2], [3]. In this work, the algorithm used is based on unique combination of two indices and mutual addition concept.

The overview of XDMA is illustrated in Figure 1 which clearly shows the primary phases of XDMA. These phases are Construction of two indices, Utilization of Search Technique, Selection of desired T-typed nodes, Retrieval of the exact data and Grading of query results. Hence, effective keyword search in XML databases can be achieved using XDMA.

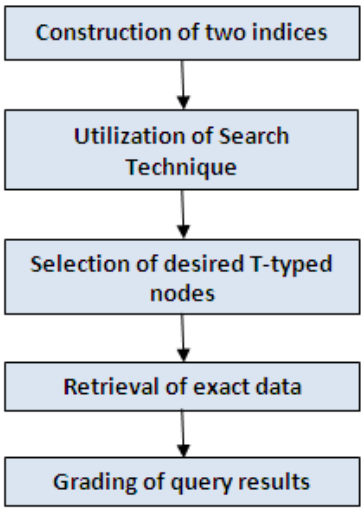

Fig. 1. Phases of XDMA

Corresponding Author: S. Selvaganesan

It is noted that XDMA is well suited for data-centric XML datasets [1], [2]. In this paper, the uniqueness, novel features and effectiveness of XDMA have been presented.

The rest of this paper is structured into the following sections. In Section 2, dual indexing, mutual dependence between two indices and utilization of frequency are discussed. In Section 3, the unique search technique is briefly discussed. Furthermore, Section 4 briefs the mutual summation concept. In Section 5, new keyword ambiguity and space requirements are presented. Finally, Section 6 presents the conclusion.

2 UNIQUE DUAL INDICES

XDMA incorporates distinctly the two indices and concept of mutual summation together in order to achieve the effective keyword search in XML databases. In this context, XDMA is the pioneer keyword search approach data-centric XML datasets. The two indices are in the form of tables, namely, tag_info index and data_info index, for each every structural nodes and data nodes of XML datasets [1], [2], [3], [4], [5].

Table 1 illustrates the portion of tag_info index for dblp XML dataset [6]. As shown in Table 1, each structural node's information, namely, name of the tag, their path-wise frequency of the instance of occurring and name of the concerned path is stored. Similarly, Table 2 demonstrates the portion of the data_info index for dblp XML dataset [6]. As shown in Table 2, for each data node, text (data) values contained in data nodes, tag name of leaf node with the same data value and frequency of occurrence of the data (text) value in the leaf node's tag are stored.

Moreover, the two indices built and used in XDMA are absolutely different from those two indices, i.e., keyword inverted list and frequency table, of another prominent XML keyword search approach XReal [7], [8]. Of the indices of XReal, keyword inverted list is not a table but is in the form of tuple adopting three indices, namely, "dup", "dupType" and "dupTypeNorm", and frequency table is the only table in XReal [7], [8]. Consequently, the concept and methodology of XDMA are entirely different from XReal.

Table 1. Portion of Tag_info index for dblp XML dataset

Sl. No.	Tag	Frequency	Path
287	mastersthesis	5	dblp,mastersthesis
289	Title	5	dblp,mastersthesis
291	School	5	dblp,mastersthesis
293	Editor	1	dblp,article
294	Title	106834	dblp,article

Table 2. Portion of Data_info index for dblp XML dataset

Sl. No.	Data (text) Value	Tag Name	Frequency
1	Kurt P. Brown	author	1
2	PRPL: A Database Workload Specification Language, v1.3.	title	1
3	1992	year	3996
4	Univ. of Wisconsin-Madison	school	16
5	Tolga Yurek	author	1

2.1 MUTUAL DEPENDENCE BETWEEN INDICES

As illustrated in Table 1 and Table 2, XDMA specifically deals with each and every tag (structural node) and data node of XML datasets independently. Leaf nodes' tag names are common in the two indices so as to share information among the indices. Moreover, data_info index relies on tag_ info index mainly to get the information about leaf node. In such a way, mutual dependence exists between data_info index and tag_info index of XDMA [1], [2]. In XReal [7], [8], there is no definite dependence between keyword inverted list and frequency table.

2.2 EXPLOITATION OF FREQUENCY

Compared with other keyword search approaches, the frequency of data (text) value in a leaf node is distinctively taken into consideration in data_info index of XDMA. As shown in Table 2, data_info index contains the frequency of occurrence of data value of leaf node of XML datasets. Hence, uncertainty in the processing of given keyword query is reduced to greater extent.

3 UNIQUE SEARCH TECHNIQUE

For each search query in XML datasets, XDMA identifies certainly each keyword as tag keyword or data keyword. Hence, searching happens obviously in the tag_info and data_info indices. For combination of tag and data keywords in queries, a distinctive search technique is introduced using the two indices to search all possible T-typed nodes [1], [2].

For a keyword query issued on the XML database, the search technique will search the keyword in the tag_ info table and if the keyword finds matches, it is regarded as a tag keyword. If the keyword does not find matches in the tag_info table, searching will be carried out in the data_info table. If the keyword matches with data (text) value(s) in the data_info table, the keyword is regarded as a data keyword.

For queries having different number of tag keywords and data keywords, these various cases are taken into consideration and given in Table 3.

Table 3. Search Technique based on dual-indexing

Case No.	Tag_info index	Data_info index	Type of Keyword	
			Tag	Data
Single keyword Query				
Case 1	Match	No Match	1 tag	-
Case 2	No Match	Match	-	1 data
Two Keyword Query				
Case 3	2 Matches	No Match	2 tag	-
Case 4	No Match	2 Matches	-	2 data
Case 5	Match	Match	1 tag	1 data
Three Keyword Query				
Case 6	2 Matches	Match	2 tag	1 data
Case 7	3 Matches	No Match	3 tag	-
Case 8	Match	2 Matches	1 tag	2 data
Case 9	No Match	3 Matches	-	3 data
Four Keyword Query				
Case 10	No Match	4 Matches	4 tag	-

4 MUTUAL SUMMATION CONCEPT

Furthermore, a novel terminology and concept, namely, mutual summation is employed for a pair of random discrete variables. In case of more number of occurrences of query keyword matching tag, the mutual score between two indices has to be determined. Hence, XDMA has made use of both the mutual summation concept and dependence among indices primarily to define and determine the mutual score between keyword matching tags and tag containing keyword matching data (text) values in order to find out the desired nodes [1], [2]. Subsequently, this concept has been utilized in XDMA for grading of query results.

5 OTHER UNIQUE FEATURES OF XDMA

5.1 IDENTIFICATION OF NEW AMBIGUITY

In XML databases, the occurrence of (query) keyword matching tag of structural node as well as (query) keyword matching data value might be once or several times. Besides the three ambiguities of keyword [7], [8], a new ambiguity of keyword is discovered and evidently emphasized in XDMA. This new ambiguity which is regarded as **Ambiguity 4** [1], [2] is "A keyword can exist as the name of a tag for node types having different data (text) values and vice versa". Also, XDMA

addresses all the ambiguities of keyword including ambiguity 4. It is noted that XDMA is the first of its kind to identity and address this new ambiguity of keyword in XML datasets [1], [2].

5.2 SPACE REQUIREMENTS

To the best of our knowledge, XML keyword search using Dual indexing and Mutual summation based Algorithm (XDMA) is a pioneer keyword search method for XML databases using both dual indexing and mutual summation concept. XDMA occupies less space in comparison with other keyword search approaches, namely, XRank and XReal. In particular, size of indices in XDMA for larger XML dataset like dblp [6] is 21 MB whereas size of indices is 2.2 GB in XReal and, 400 MB in XRank [9]. Space requirement of XDMA is considerably low [1], [2].

6 CONCLUSION

The dual indexing and mutual summation based keyword search method for XML databases, i.e., XDMA, is unique and novel and, distinctively different from other existing keyword search methods for XML databases [1], [2]. In this paper, the uniqueness and novel features of XDMA have been precisely presented. Moreover, the experiments on different XML datasets with various XML keyword search algorithms and evaluation of the experimental results prove the effectiveness of XDMA [1], [2], [3].

REFERENCES

[1] S. Selvaganesan, S. C. Haw, and L. K. Soon, "XDMA: A Dual Indexing and Mutual Summation based Keyword Search Algorithm for XML Databases," *International Journal of Software Engineering and Knowledge Engineering (IJSEKE)*, vol. 24, no. 4, pp. 591-616, 2014.

[2] Selvaganesan S., *Dual Indexing and Mutual Summation Based Keyword Search Method for XML Databases*, Ph.D. thesis 2014. Multimedia University, Cyberjaya, Malaysia.

[3] S. Selvaganesan, S. C. Haw, and L. K. Soon, "Effective Keyword Search Approach XDMA for XML Databases," *Mitteilungen Klosterneuburg Journal*, vol. 64, no. 10, pp. 43-53, 2014.

[4] S. Selvaganesan, S. C. Haw, and L. K. Soon, "Effective XML Keyword Search Using Dual Indexing Technique," *Information Technology Journal*, vol. 13, no. 4, pp. 643-651, 2014.

[5] Selvaganesan, S, Haw, S.C., and Soon, L.K., *Towards developing an Efficient Approach to Keyword Search for XML Documents*, Proceedings of International Conference on System Engineering and Modeling, pp. 78-83, 2012.

[6] Miklau G. XML Data Repository, 2002. [Online] Available: http://www.cs.washington.edu/research/xmldatasets/www/repository.html (accessed 15/01/2013)

[7] Bao, Z., Lu, J., Ling, T.W., and Chen, B., *Towards an Effective XML Keyword Search,* IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 8, pp. 1077-1092, 2010.

[8] Bao, Z., Ling T.W., Chen, B., and Lu J., *Effective XML Keyword Search with Relevance Oriented Ranking*, Proceedings of the IEEE International Conference on Data Engineering, pp. 517-528, 2009.

[9] Guo, L., Shao, F., Botev, C., and Shanmugasundaram, J., *XRANK: Ranked Keyword Search over XML Documents*, Proceedings of 2003 ACM SIGMOD International Conference on Management of Data, pp. 16-27, 2003.