Searching SNT in XML Documents Using Reduction Factor

Mary Posonia A

Department of computer science, Sathyabama University,
Tamilnadu, Chennai, India
maryposonia@sathyabamauniversity.ac.in
http://www.sathyabamauniversity.ac.in

Dr V L Jyothi

Department of computer science, Jeppiaar Engineering College,
Tamilnadu, Chennai, India
Jyothivl15@yahoo.com
http://www.jeppiaarcollege.org

Abstract

XML has become the most popular standard for data representation. In XML standard the documents represented as rooted ordered trees. The efficient query processing can be performed on the labeled document structure. The major advantage of keyword search is the user need not understand the complex query language and the structure. This paper proposes a modified H –Reduction factor which solve the inconsistency and the abnormality problem in the Xreal and Dynamic- infer method. The experimental result shows the effectiveness of the proposed approaches.

***Keywords*:** Keyword Search; Dynamic Reduction factor **;** H –reduction factor ; XML database.

1. Introduction

The XML keyword search becomes the most efficient method in text document especially for the web applications so that the user input the intended keyword for querying XML document instead of having the complete knowledge about XML schema and query. In the information retrieval area most of the information is stored in an XML format hence the XML becomes the standard format for storing and transforming data. The standard query language for querying XML document as Xquery allows the user to retrieve the exact semantically meaningful information, but the user could have the proper knowledge to raise the query so that it is not suitable for all types of end users.

The keyword search engine or the prototype has designed by Xreal[11] which identified the user search intension based on the three guidelines.

Guideline 1: For every query keyword there should be at least some ‘T’ typed node present in the XML subtree.

Guideline 2: The ‘T’ typed node in the XML tree should be capable of holding sufficient information about the user query.

Guideline 3: The ‘T’ typed node in the XML document should not be overpowering to contain irrelevant information with regard to the user query.

In the Dynamic reduction factor method[4] they discussed the abnormality problem due to the improper reduction factor value, so that the confidence score of the search for a node is lower than the confidence score of any ancestor node in the XML tree, hence the ancestor node has chosen as the search for node (SNT).

Keyword search in an XML document is used to identify the nodes which contain matching keyword and also it shows the interlink between the query keyword based on LCA[9]. However LCA fails to find the meaningful answer in all possible cases.

The issues found in XReal was discussed in Dynamic infer method but the abnormality problem has not solved because, the low frequency queries the value of the reduction factor exceeds the maximum value of ‘r’(r<=1).

In this paper we devise the confidence score using an appropriate reduction factor value which is the fixed one, whereas the dynamic factor value changes on the fly. Hence the modified reduction factor value is the

combination of the reduction factor evaluated in real and the reservation space of the Dynamic reduction factor. Using modified confidence score find the appropriate Search For Node(SNT) for the user keyword query to solve the abnormality problem.The experiments prove the effectiveness of the proposed approach.

We summarize the contribution of this paper as follows:

- We reconsider the problem (i.e.,inconsistency and abnormality problems) of Xreal for inferring SNTs identified by Jiang Li et.al[4].
- Proposed an H deduction factor r' to solve the identified problem.
- Provide an algorithm to H – infer to effectively infer the SNT.
- Conducted a comparative study between Dynamic Reduction method and Xreal to prove the effectiveness of the proposed approach.

The rest of the paper is organised as follows. Section II presents related work. Section III describes background ,definitions and terms used in the paper.Section IV proposes the modified H – reduction factor and the algorithm to infer SNT. Experiments are implemented and the results verified in section V. Finally, section VI presents the conclusion and remarks.

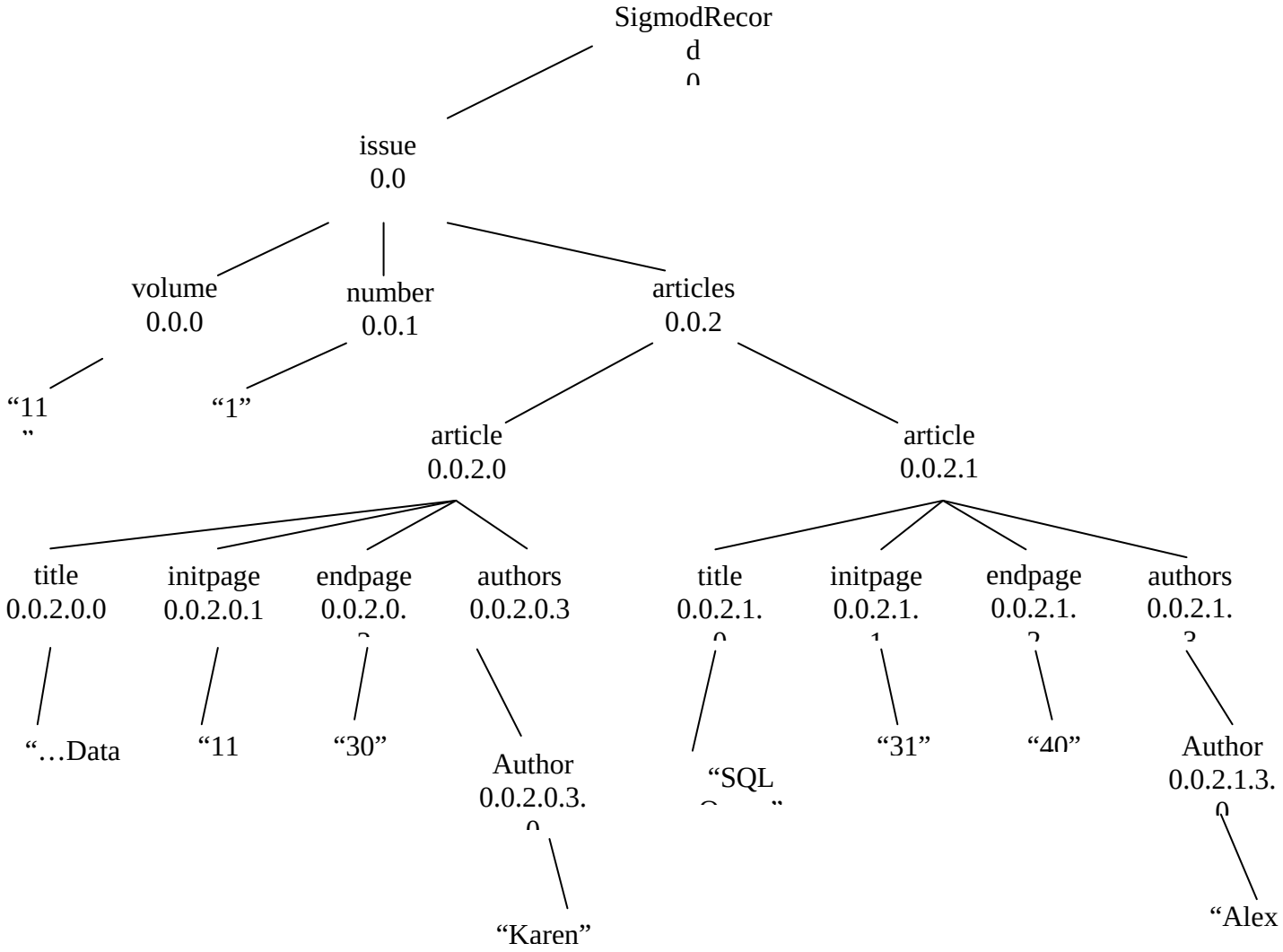

Fig 1 . XML data tree of Sigmoid Record

2. Related work

In XML there has been a lot of work done in association with keyword search, Xreal [4,12] proposes a keyword ambiguity problem and developed an approach for finding the search for the node. In 2010 Jian Li et.al

[4] proposed a novel approach for dynamic search which solves the drawback like inconsistency and abnormality problem in Xreal. In 2010Liang Jeff Chen et al. [6] developed a method of top-K keyword search using tree pruning with the combination of join-based algorithm and implemented the algorithm in relational databases. In ZhifengBao et al. [11] developed an IR – style approach to address user search intension, Resolve keyword ambiguity problem, and relevance oriented ranking. They designed the specific guidelines and formulated the novel approach XML TF*IDF strategy to find the possible matches for the user search intension. In Ziyang Liu et al. [12], proposed an axiomatic framework that includes two non-trivial property like monotocity and consistency, with respect to the user and data query this algorithm satisfies both properties including semantics in the query. For users, each result proportional to possible user search intention. To improve the efficiency it is desirable to find the most relevant results from the retrieved data. In 2010 Jianxin Li et al. [13] addressed the problem for 1) the retrieved data does not contain the sufficient schema information 2) relevant results may be computed for before query evaluation. They proposed estimation type of query evaluation to speed up the user required result retrieval and they the efficient algorithm to explore the semantics. In 2009 Lingbo Kong et al. [7] Proposed the technique of Relaxed Tightest Fragment (RTF) as the initial result segment and they also proposed filtering technique to reduce the redundancy. Keyword based search in an XML document using Lowest Common Ancestor (LCA) made real interest in keyword search [9,15]. In 2008 Yu Xu et al. [9] Introduced Index Stack algorithm to effectively find the keyword query with the help of XRANK.They analysed the query complexity based on number of keywords,the depth of the tree and the number of occurrences of keyword present in the tree.

3. XML Data model and definitions

The Xml document has modelled as rooted ordered labelled trees and elements in the tree are also represented as nodes. The bellows DTD representation shows the internal format of elements in the XML document.

<!ELEMENT Sigmoid Record(issue)>

<!ELEMENT issue(CDATA)>

<!ELEMENT issue(Volume,number,article)>

<!ATTLIST Volume(PCDATA)>

<!ELEMENT (number,article)>

<!ELEMENT number(CDATA)>

<!ELEMENT article(CDATA)

<!ELEMENT initpage,endpage(CDATA)

The root element in the DTD as Sigmoid record and it's also called an ancestor node. The sub element issue has the associated elements like author,title,initpage and endpage.

3.1. *Definitions*

Entity node : XML document modelled as rooted labelled ordered tree and every node present in the tree called as entity node. The nodes in XML tree are related to the ancestor - descendent relationship. For example in the data tree fig 1. nodes issue (0.0), article (0.0.2.0) and article (0.0.2.1) are inferred as the entity nodes. The entity nodes article (0.0.2.0) and article (0.0.2.1) have the same entity-type, which is the node type SigmodRecord.issue.articles.article.

Ancestor Node Type:

Every XML document has root node and this root node also called as ancestor node. Every node in the XML tree must be rooted with the ancestor node type.

Neighbour node types:

The nodes appear in the XML subtree may be related to same entity type called neighbour node.In fig 1. The node types article, title, initPage, endPage, authors and author are neighbour node types because they share the same entity-type article.

4. Proposed Modification

The modified H reduction factor with the associated confidence score which resolves the problem identified by the Jiang Li et.al.[4]. The lower node in the XML tree has enough information to find the SNT for the user query. The reduction factor should satisfy the bellow condition.

rf ≤ 1 , the range of rf should be in 0 to 1.

According to Jiang Li et.al.[4], the reduction factor must be determined from number of occurrences of search for node for the respective keyword in given user query and the reservation space. In the proposed method, consider the reduction factor as a fixed value with reference to Zhifeng Bao et.al. [11] and Jiang Li et.al.[4],which is rf=0.85. In the proposed method we added the value of reduction factor r=0.8, and the reservation[4] space rs = 0.05.

rf = (Reduction factor + Reservation space) ≤ 1

The confidence score formula for search for node with respect to the dependent variable.

$C(F , Q) = In[1 + P(F \cap K)] * rf * depth (F)]$

Where, $P (F \cap K) = P(F) * P(K / F)$

P(F) = Frequency/ Probablity of node containing 'K'

P(K/F) = Frequency/ ocrrences of keyword in node type

Depth (F) = Depth of search for node.

5. Results and Discussion

The experimental results of the proposed approach is implemented in JAVA (jdk 1.6). The input data set taken from Sigmoid record[14]. The below table represents the effective comparison of real , Dynamic infer method and the proposed approach. In the resultant table even the Dynamic infer method has effectively identified the search for node but the reduction factor (rf = 1.05) which exceeds the maximum value of '1'. Hence in the proposed approach considered the fixed value of the reduction factor rather than adjusting reduction factor which may lead to the program complexity.

Table : 1. C (F, Q) for abnormality problem (Query q : { Karen})

Query	SNT Xreal	rf	C(F,Q)	SNT Dynamic	rf	C(F,Q)	SNT Proposed	rf	C(F,Q)
{ Karen}	issue	0.8	0.8789	issue	1.05	1.1535	issue	0.85	20.423
	article	0.8	0.5625	article	1.05	1.278	article	0.85	27.571

6. Conclusion

In this paper, we identified the undesired reduction factor which increases the program complexity in the Dynamic infer algorithm. To overcome these problems the proposed method uses the modified confidence score and the H- reduction factor to reduce the program complexity. Finally we conducted experiments to evaluate and compare the proposed approach against existing methods like Xreal and Dynamic infer method for measuring the effectiveness of the approach. The experiments prove the better performance of the proposed approach .

References

[1] Bao, Z., et al.,(2009) : Effective XML Keyword Search with Relevance Oriented Ranking, in Proceedings of the 2009 IEEE International Conference on Data Engineering , IEEE Computer Society. p. 517-528.

[2] http://www.cs.washington.edu/research/xmldatasets

[3] Ilyas, I.F., G. Beskales, and M.A. Soliman, (2008) :A survey of top-k query processing techniques in relational database systems. ACM Comput. Surv., p. 1-58.

[4] Jiang Li, Junhu Wang, (2010) : Effectively Inferring the Search-for Node Type in XML Keyword Search, Database Systems for Advanced Applicationsm, Lecture Notes in Computer Science,Volume 5981, pp 110-124.

[5] Jianxin Li, Chengfei Liu, Rui Zhou and Wei Wang,(2011): Top-k Keyword Search over Probabilistic XML Data, Data Engineering (ICDE),IEEE 27th International Conference.

[6] Liang Jeff Chen, YannisPapakonstantinou, (2010) : Supporting top-K keyword search in XML databases, Data Engineering (ICDE), IEEE 26th International Conference , pp 689- 700

[7] Lingbo Kong, RémiGilleron, Aurélien Lemay Mostrare, (2009) : Retrieving meaningful relaxed tightest fragments for XML keyword search," EDBT '09 Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, pp 815-826.

[8] World Wide Web Consortium, http://www.w3.org.

[9] Yu Xu, YannisPapakonstantinou, (2008) : Efficient LCA based keyword search in XML data, EDBT '08 Proceedings of the 11th international conference on Extending database technology: Advances in database technology, pp 535-546.

[10] Y. Xu and Y. Papakonstantinou,(2005) : Efficient keyword search for smallest LCAs in XML databases, in proceedings of SIGMOD, pp. 537–538.

[11] ZhifengBao, Jiaheng Lu, Tok Wang Ling, Bo Chen,(2010) : Towards an Effective XML Keyword Search, Knowledge and Data Engineering, IEEE Transactions, Volume: 22 , Issue: 8, pp 1077- 1092.
[12] Ziyang Liu, Yi Cher,(2008) : Reasoning and identifying relevant matches for XML keyword search, Journal Proceedings of the VLDB Endowment, Volume 1, Issue 1, Pages 921-932.
[13] ZhifengBao, Tok Wang Ling , Bo Chen , Jiaheng Lu,(2009) : Effective XML Keyword Search with Relevance Oriented Ranking, Data Engineering, ICDE '09.IEEE 25th International Conference, pp 517- 528.
[14] Ziyang Liu, Yi Chen, (2010) : Return specification inference and result clustering for keyword search on XML, Journal ACM Transactions on Database Systems (TODS) Volume 35, Issue 2, Article No. 10.
[15] Ziyang Liu and Yi Chen,(2007) : Identifying Meaningful Return Information for XML Keyword Search, in SigmodConferenceACM , pp 329--340.