International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 3 (2015) pp. 7325-7334

http://www.ripublication.com

Duplicate Detection in XML Data Using Extended Sub Tree Similarity Function

G.Bharathi Mohan

Research scholar Anna University
gbharathi80@gmail.com

Dr.T.Ravi

The Principal Srinivasa Institute of Engg Tech
travi679@yahoo.com

J.Lin Eby Chandra

Asst.Professor Jaya Engineering college
ebyworld@gmail.com

M.A.Mukunthan

Asso.Professor Jaya Engineering College
mamukunthan@gmail.com

Abstract

Electronic data acts as a major role in many applications such as banking, catalog maintenance, and Library management etc.. While collecting such large amount of data from many, distributed and different sources causes data quality problems such as duplicates. Those data's are normally in the form of relational or in hierarchical manner. An XML is one of the hierarchical ways of representing the data. There are not too many solutions available for duplicate detection in hierarchical data. A recent approach for XML duplicate detection, called XMLDup uses a Bayesian network to determine the probability of two XML elements being duplicate. It consider both the similarity of attribute content and the relative importance of the descendant elements with respect to the overall similarity score calculated using the Edit base distance function. Even though Edit base distances are a well-known family of tree distances function, however it has several drawbacks in its mapping rules. This paper proposes a new similarity function for XML data comparison, namely Extended Sub Tree

(EST), a new Sub Tree mapping is introduced in order to identify duplicates between two different XML data.

Keywords Data cleaning, Duplicate Detection, Similarity function, Edit distance, Extended Sub tree, Bayesian Network, XML.

1. INTRODUCTION

Data cleaning, mainly deals with detecting and removing errors and inconsistencies from data for improving the quality of data. Duplicate detection is one of the processes of data cleansing which mainly used to identify the duplicates. Duplicates are the different representation of the same real world entity in multiple times. In order to provide access to quality data, consideration of different data representations and elimination of duplicate information become necessary.

Duplicates are not easy to detect, especially in large volumes of data because of the fact that it is represented in different structure. Simultaneously, they decrease the usability of data, cause unnecessary expenses, customer dissatisfaction, incorrect performance indicators, and their inhibit comprehension of the data and its value. The effects of such duplicates are detrimental; for instance, Consider E-shopping application, the shop keeper maintains a details for each customer .if their exists multiple representation of details of the same customer ,catalogs will mailed multiple times to the same household, etc. Let us consider catalog maintenance for a paper in which each object has the following element such as title, author, first name, last name, journal, volume, pages. Duplicate detection in XML data is to identify whether two objects in the xml file represent the same real world entity or not. These can be constructed by comparing the two xml tree as input as shown in the fig1.1

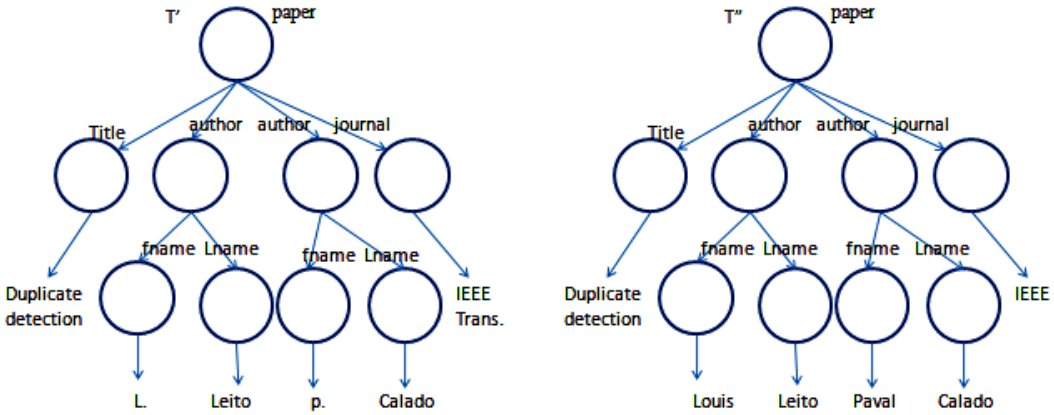

Fig1.1XML Tree T' and T"

Contribution:
One cannot apply the duplicate detection method introduced for identifying the duplicates in relational data to that of the hierarchical data like XML. In order to identify the duplicates between two XML data the similarity scorer acts as the major role .In the recent approach of duplicate detection involves XMLDup algorithm deals with the edit distance (.ned) similarity scorer to calculate the probability of each node in a XML tree to identify whether a root node is duplicates or not. If the overall similarity score for two records exceeds the threshold then it is set as duplicates. However, Edit base distance approach has several drawbacks in its mapping. First ordering may prevent mapping between similar nodes. Second every single node in a tree can be mapped to a single node in another tree leads to inappropriate mapping. Finally Edit base distance is mainly a single node mapping function rather than sub tree mapping. The benefits of sub tree mapping are mainly to overcome these problems and reduce the number of comparison strategy.

2. Related Work

S.Guha et al [7] proposed Duplicate detection using XML join is one of the efficient way of identifying the duplicates in which main focus is on how to efficiently join two sets of similar elements and the accuracy if its joining process. So they targeted on tree edit distance which to be applied on Xml join further.

A later approach was based on accuracy developed by Carvalho et al [3].In this work they mainly concerned with the integration of the tree-structured data extracted from the web. Uses a cosine measure of similarity to identify duplicates by comparing the two object representation by transforming its hierarchical representation of two person elements into a vector of terms. Mostly ignoring the concept of hierarchical structure object and weighted similarity is taken into account for the related fields within the vectors.

Only the recent work focus on the appropriate goals of identifying the duplicates objects representation in XML databases. These works are different from the older approach in which they were designed specifically to exploit the distinctive characteristics on XML object such as the textual content, and the semantics implicit of XML tags.

Dogmatix framework, Felix Naumann et al [1], aims for both the efficient and effective in identifying the duplicates majorly consist of three steps in it: Candidate definition, what objects in the data source should be compared. Duplicate Definition, defines when two duplicate candidates are duplicates. Duplicate detection, defines how we search duplicates within duplicate candidates.

Structure aware approaches proposed by D.Milano et al [4] for XML object identification rely on distance measures based on the tree structure of XML, like tree edit distance. This hierarchical, tree-like nature justifies the proposal of similarity measures that integrate string comparison functions with tree edit distances. which is not suited to perform approximate comparisons of XML data, In this paper, they defined a new distance for XML data, the structure aware XML distance, that identify the duplicates based on the optimal cost of overlay of two comparable XML .These

approach is different from other tree-distances, the distance defined here can be computed in polynomial time, even if the trees are unordered.

The XMLDup system initially proposed by L.Leito et al [2] uses a Bayesian Network model for XML duplicate detection and the similarity can be calculated using the edit distance function. This is one of the probabilistic approaches of identifying the duplicates.

The SXNM (Sorted XML Neighborhood Method) proposed by S.Puhlmann et al [6] adapts the concept of relational sorted neighborhood to XML data. As followed in the original SNM, they focused on reducing the useless comparisons between objects by grouping together the most similar objects using windowing technique.

Efficient XMLDup L.Leito et al [8] system which proposed the concept of the XMLDup in which it also mainly focused on Bayesian network construction a probability way of identifying the duplicates, it additionally involves the pruning techniques to prune the Bayesian network in order to improve the runtime efficiency of XMLDup. The similarity calculation used for the approach is the edit distance but it has several drawbacks in it, these leads to the development of this paper.

3. Proposed Work

XML Duplicate detection(EST)

In this paper, we propose a new similarity function with respect to tree structured data, namely Extended Sub tree (EST). We justify the need to propose a new tree comparison approach by discussing situations where previous approaches have poor performance. The new similarity function avoids these problems by preserving the structure of the trees. That is, mapping sub trees rather than nodes is utilized by new mapping rules. The motivation of proposing EST is to enhance the edit base mappings. Consequently, EST introduces new rules for sub tree mapping. This new approach seeks to resolve the problems and limitations of edit based approaches.EST similarity function has been proposed for the domain of tree structured data comparison with the aim of increasing the effectiveness of applications utilizing tree distance or similarity functions. A variety of tree comparison approaches are introduced in the previous section. Each approach has advantages and disadvantages in terms of the distance/similarity score. Based upon an empirical investigation, where the previous approaches do not give an appropriate similarity/distance score. In the following, these cases are analyzed with illustrative examples where all discussions are in terms of a normalized similarity score, $S^*(T^a, T^b)$. while , $S^*(T^a, T^b)=1$ means that the tree are similar, $S^*(T^a, T^b)$means that the trees are totally distinct.

$$S^{¿}(T^a, T^b) = \frac{S(T^a, T^b)}{Max(|T^a|, |T^b|)}$$

ADVANTAGES

- A novel similarity function to compare tree structured data by defining a new set of mapping rules where sub trees are mapped rather than nodes.

- The new approach reduces the no of comparisons than that of the previous distance function

4. System Architecture

Fig4.1 System Architecture

4.1 Extraction

The Structure Extractor automatically processes the XML document to build a structure tree and determine the multiplicity of the elements. This can be achieved through a XML SAX Parsing technique. The output of the extractor is the tree structured data that describe the nesting of the xml elements. The tree view highlights differences down to the level of elements, words or attributes. These leads to easy selection of candidate pair or objects from the dataset.

4.2 Schema Integration

Schema Integration involves the process of Integrating two xml objects through Bayesian network construction. Bayesian Networks provide a concise specification of a joint probability distribution. They can be seen as a directed acyclic graph, where the nodes represent random variables and the edges represent dependencies between those variables. When the two xml objects are said by duplicates only based on the root node of the bayesian network. if the root node takes the value 0 to represent the fact that the nodes are not duplicates. The fact that two XML nodes are duplicates depends only on if values and children nodes are duplicates.

Based on probability measured using EST similarity function duplicates in xml data can be identified. To compute the probability it is necessary to define

- Prior probabilities for leaf nodes.
- Conditional probability for inner nodes.
- Posterior Probability for root node.

4.3 Duplicate Identification

Prior probabilities

Prior probabilities are mainly to represent the likelihood that two values in the xml trees are the same i.e the leaf node of Bayesian network. This Probability can be calculated using the similarity function ESTSim (.) by setting the threshold values between 0 to 1 Defined as

$$P(t_{ij}[a]) = \begin{cases} \\ \text{Sim (Vi [a], Vj [a])}, & \text{if similarity was measured} \end{cases}$$

$$= \qquad \text{Ka} \qquad , \qquad \text{other}$$

Conditional probabilities

Conditional probabilities are mainly to represent the likelihood that the two children node in the xml trees is the same i.e. the inner node in Bayesian network. These probabilities can be calculated by applying condition in four different ways**.CP1**- probability of the values of the nodes being duplicates, given that each individual pair of values contains duplicates. **CP2-** probability of the children nodes being duplicates given that each individual pair of children are duplicates.**CP3**-probability of two nodes being duplicates given that their values and their children are duplicates **CP4** - probability of a set of nodes of the same type being duplicates given that each pair of individual nodes in the set are duplicates .

Final probability

Once all prior and conditional probabilities are defined, the BN can be used to compute the probability of two XML trees being duplicates. This can be achieved by any probability propagation algorithm in which its include the overall probability value of a root node .These can be calculated by taking account of the prior and conditional probability .If the final probability is below the normalized value fit between 0 and 1, then it as set duplicates otherwise classified as non duplicates.

5. Experimental Results

Performance evaluation through statistical measures involves precision and recalls during this stage the effective and efficiency of the system is measured by comparing the XMLDude with XMLDup as shown in the fig 5.1

Table 5.1 Performance time and No of Comparisons

XMLDude		XMLDup	
CORA Dataset			
Time	Comparison	Time	Comparison
00:01:00	41623415	00:02:42	71888193
CD Dataset			
00:00:43	43543216	00:01:03	7228014
IMBD dataset			
00:30:05	4056765402	00:47:08	603480020

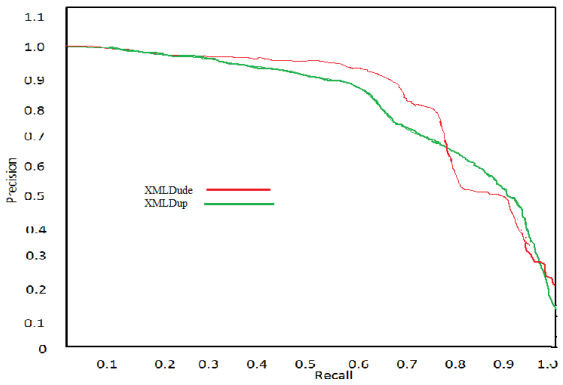

Fig 5.2Comparing XmlDude with XmlDup

6. Conclusion

In this paper the problem of identifying the duplicates in hierarchical data is proposed. It is necessary to achieve effective and efficient way of identifying the duplicates by avoiding the unnecessary comparisons in finding the similarity between the two candidate pairs. In order to avoid unnecessary comparison, a new similarity function with respect to tree structured data, namely Extended Sub tree (EST) is introduced. The new similarity function avoids these problems by preserving the structure of the trees. That is mapping sub trees rather than nodes are utilized by new mapping rules. The motivation of proposing EST is to enhance the edit base mappings, by generalizing the one-to-one and order preserving mapping rules. Consequently, EST introduces new rules for sub tree mapping. This new approach seeks to resolve the problems and limitations of edit based approaches.

7. Future Work

To extend the BN model construction algorithm to compare XML objects with different structures and apply machine learning methods to derive the conditional probabilities and network structure based on the existing data.

REFERENCES

[1] M. Weis and F. Naumann, "Dogmatix Tracks Down Duplicate in XML," Proc. ACM SIGMOD Conf. Management of Data, pp. 431-442, 2005.

[2] L. Leita˜o, P. Calado, and M. Weis, "Structure-Based Inference of XML Similarity for Fuzzy Duplicate Detection," Proc. 16th ACM Int'l Conf. Information and Knowledge Management, pp. 293-302, 2007.

[3] A.M. Kade and C.A. Heuser, "Matching XML Documents in Highly Dynamic Applications," Proc. ACM Symp Document Eng. (DocEng), pp. 191-198, 2008.

[4] D. Milano, M. Scannapieco, and T.Catarci, "Structure Aware XML Object Identification," Proc. VLDB Workshop Clean Databases (CleanDB), 2006.

[5] P. Calado, M. Herschel, and L. Leita˜o, "An Overview of XML Duplicate Detection Algorithms," Soft Computing in XML Data Management, Studies in Fuzziness and Soft Computing, vol. 255, pp. 193-224, 2010.

[6] S. Puhlmann, M. Weis, and F. Naumann, "XML Duplicate Detection Using Sorted Neighborhoods," Proc. Conf.Extending Database Technology (EDBT), pp. 773-791, 2006.

[7] S. Guha, H.V. Jagadish, N. Koudas,D.Srivastava, and T. Yu,"Approximate XML Joins," Proc. ACM SIGMOD Conf. Management of Data, 2002.

[8] L.Leito, Pavel Calado and Melanie Herschel, "Effective and efficient duplicate detection in Hierarchical Data", IEEE Transaction on knowledge and Data Engineering, 2013

[9] E. Rahm and H.H. Do, "Data Cleaning: Problems and Current Approaches," IEEE Data Eng. Bull., vol. 23, no. 4, pp. 3-13, Dec.2000.

[10] R. Ananthakrishna, S. Chaudhuri, and V. Ganti,"Eliminating Fuzzy Duplicates in Data Warehouses," Proc.Conf. Very Large Databases (VLDB), pp. 586-597, 2002.

[11] J.C.P. Carvalho and A.S. da Silva, "Finding Similar Identities among Objects from Multiple Web Sources,"Proc. CIKM Workshop Web Information and Data Management (WIDM), pp. 90-93, 2003.

[12] R.A. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., 1999.

[13] M.A. Herna´ndez and S.J. Stolfo, "TheMerge/ Perge Problem for Large Databases," Proc. ACM SIGMOD Conf. Management of Data,pp. 127-138, 1995.

[14] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, second ed. Morgan Kaufmann Publishers, 1988.

[15] L. Leita˜o and P. Calado, "Duplicate Detection through Structure Optimization," Proc. 20th ACM Int'l Conf. Information and Knowledge Management, pp. 443-452, 2011.

AUTHORS PROFILE

G.Bharathi Mohan is a Research Scholar, Department of Computer Science, Anna University, chennai, Tamilnadu. He got his B.E. Degree in Electrical and Electronics Engineering from Madras University and M.Tech Degree in Information Technology from Anna University. He has about 9 years of teaching experience. He has He has published 3 articles in international journal and many national journals.

Dr.T.Ravi is recognized Research Supervisor in Anna University, Chennai. He has published more than 25 papers in National and international journal. He has presented more than 50 papers in International conferences. He is currently working as Principal in Anna University affiliated college and has more than 21 years of teaching experience

J.Lin Eby Chandra working as an Asst.professor in Department of Computer Science Engineering, Jaya Engineering College. He got his B.E Degree in Computer Science Engineering from Anna University and M.Tech Degree in Computer Science Engineering from Anna University. He has about more than 8 years of teaching experience and published 3 international article and many national papers.

M.A Mukunthan is working as an Associate Professor,Department of Computer Science, Jaya Engineering College, Tamil nadu, India. He received his B.E.Degre in Computer science and Engineering from Madras University and M.E Degree in Computer science and Engineering from Anna University. He has about 10years of teaching experience and 2 years Industrial experience. He has published four articles in international journal and many national journals.