INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

PRUNING BASED DUPLICATE DETECTION IN HIERARCHICAL DATA

Ms.M.Lakshmipriya (M.E/CSE), Mr.G.Loganathan,M.E(AP/IT)
Muthayammal Engineering College, India
lpmlakshmipriya@gmail.com, logu.mscme@gmail.com

Abstract – Duplicate detection is the process of detecting multiple representations of a same real-world object and that for every object represented in a data source. In existing system a novel method of duplicate detection called XMLDup , which uses Bayesian network and Network pruning algorithm to detect duplicate and non-duplicate XML data. The proposed system compares XML object with different structures and apply support vector machine algorithm to derive conditional probability, which is the easiest way to detect the duplicate data quickly and efficiently.

Keywords – XML, Duplicate detection, Bayesian Network, SVM

I. INTRODUCTION

Duplicate detection is the process of detecting different entries in the data source representing same real world entries DogmatiX[10], which compares XML elements based not only on their direct data values, but also on the similarity of their parents, children, structure, etc. The most prominent application area for duplicate detection is customer relationship management (CRM), where multiple entries of the same customer can result in multiple mailings to the same person, incorrect aggregation of sales to a certain customer, etc. Other application areas include bioinformatics, catalog integration, and in general any domain where independently collected data is integrated.

Duplicate detection has been studied extensively for relational data stored in a single table[1]. Algorithms performing duplicate detection in a single table generally compare tuples (each of which represents an object) based on attribute values. The existing system uses a novel method for duplicate detection in XML data called XMLDup. Detecting duplicates in XML is more challenging than detecting duplicates in relational data because there is no schematic distinction between object types among which duplicates are detected and attribute types describing objects.The XMLDup uses Bayesian network model to compute the probability of any two XML objects, represented by XML elements, being duplicates. It considers the hierarchical structure of XML elements by considering probabilities for descendant XML elements. The model is built automatically based on the structure of the objects being compared. Since the structure contains no cycles, the duplicate probability can be determined efficiently.

Although the existing algorithm gain high precision and recall scores. The system proposes machine learning algorithm such as Support vector machine(SVM) to derive the conditional probability directly. Support vector machine are supervised learning models with associated learning algorithms that analyze and recognize patterns, used for classification and regression analysis.

II. METHODS

The method for detecting duplicate data is to construct the Bayesian network model then construct the similarity between XML objects. Using this similarity we can classify two XML objects as duplicates if it falls above a threshold.

A.BAYESIAN NETWORK CONSTRUCTION

Basic assumption for XML duplicate detection, the fact that two XML nodes are duplicates depends only on the fact that their values are duplicates and that their children nodes are duplicates. Then, two XML trees are duplicates if their root nodes are duplicates. An XML tree is defined as a triple $U = (t, V, C)$ Where, **t** is a root tag label, e.g., for tree U.**V** is a set of (attribute, value) pairs. If the node itself has a value, we can consider it as a special (attribute, value) pair. **C** is a set of XML trees, i.e., the sub-trees of U. These sub trees are again each described by a triple.

B.DERIVING CONDITIONAL PROBABILITY

The conditional probability can be derived based on CP1,CP2,CP3,CP4.

CONDITIONAL PROBABILITY 1:

The probability of the values of the nodes being duplicates, given that each individual pair of values contains duplicates.

CONDITIONAL PROBABILITY 2:

The probability of the children nodes being duplicates, given that each individual pair of children are duplicates.

CONDITIONAL PROBABILITY 3:

The probability of two nodes being duplicates, given that their values and their children are duplicates.

CONDITIONAL PROBABILITY 4:

The probability of a set of nodes of the same type being duplicates given that each pair of individual nodes in the set are duplicates.

C. NETWORK PRUNING FOR BN

The network pruning strategy is lossless in the sense that no duplicate objects are lost. Only object pair incapable of reaching a given duplicate probability threshold are discarded. The idea behind our pruning proposal lies in avoiding the calculation of prior probabilities, unless they are strictly necessary. The strategy is that before comparing two objects 1)All the similarities are assumed to be 1. 2) At every step of the process, maintain an upper bound on the final probability value. 3) At each step, whenever a new similarity is computed, the final probability is estimated taking into consideration the already known similarities and the unknown similarities that assume to be 1. 4) When verify that the network root node probability can no longer achieve a score higher than the defined duplicate threshold, the object pair is discarded and, thus, the remaining calculations are avoided.

D. PRUNING FACTOR ALLOCATION

Before evaluation, every node is assumed to have a duplicate probability of 1. We call this assumed probability the **pruning factor**. Having a pruning factor equal to 1 guarantees that the duplicate probability estimated for a given node is always above the true node probability.

Therefore, no duplicate pair of objects is ever lost. By lowering the pruning factor, we lose this guarantee. Thus, a pair of objects may be prematurely discarded, even if they are true duplicates. By lower pruning factor, all probability estimates will be lower, this will cause the defined duplicate threshold to be reached sooner and the network evaluation to stop sooner. Thus, fewer similarity calculations will be performed.

E. SUPPORT VECTOR MACHINE

The SVM is a supervised learning model that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training data, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new data in to one category, making it a non-probabilistic binary linear classifier. Given some training data D, a set of n points of the form,

$$D=\{(x_i,y_i) \mid x_i \in R^p , y_i \in \{-1,1\}\}_{i=1}^{n}$$

where the y_i is either 1 or −1, indicating the class to which the point $\mathbf{X}_i$ belongs. Each $\mathbf{X}_i$ is a p-dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having $y_i=1$ from

those having y_i=-1. Any hyperplane can be written as the set of points $\mathbf{X}$ satisfying,

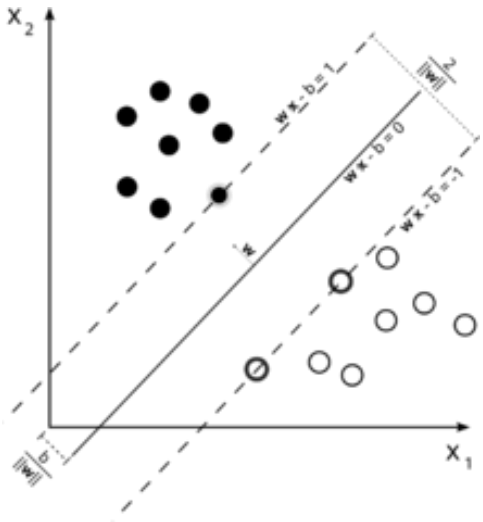

Fig 1: Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.

W.X-B=C

For Test data, **W.Y-B=?**

Where, W= Support vector, X= Input, B= Boundary, C=Class Label. The advantage of SVM is that it will compute the duplicate and non-duplicate data by directly deriving conditional probability using the train data, so the time taken to produce output is less compared to existing system.

III EXPERIMENTS AND RESULTS

Several datasets have been taken and experimented for detecting the duplicate data.

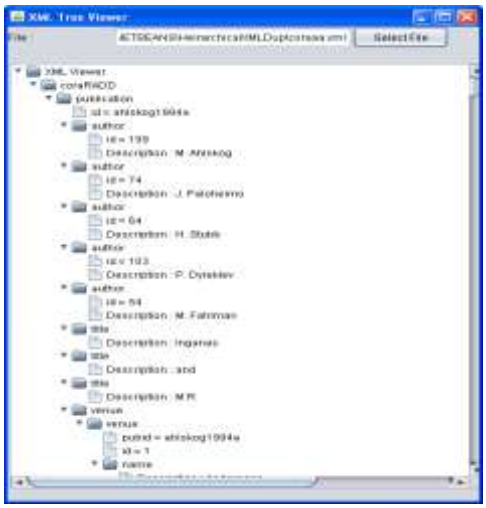

Fig 2: XML TREE VIEWER

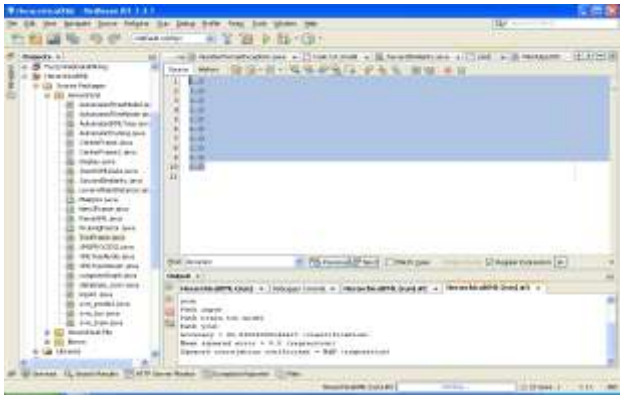

Fig 3: The result shows that 1.0 represent duplicate and 2.0 represent Non-duplicate

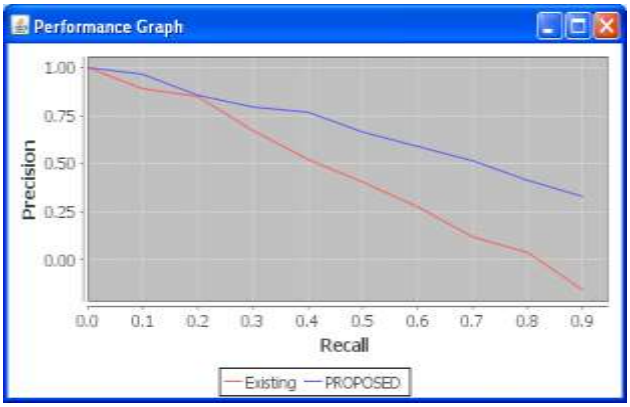

Fig 4: PERFORMANCE EVALUATION

IV CONCLUSION

The novel method of duplicate detection called XMLDup uses Bayesian Network model to determine the probability of two XML objects being duplicate. The system uses Network pruning algorithm to improve the BN evaluation time. Although the existing system gives efficient result. The proposed system applies Support vector machine algorithm to compute the conditional probability directly based on training data which has class label. Then, it will match each of the values and create the model ,predict the class label. If the predicted value falls between given threshold then the corresponding value is duplicate. Compare to the existing system SVM produce result within a fraction of time and then gain high recall and precision scores.

V REFERENCES

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

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

[3] Chen .L, Zhang .L, Jing F, Deng K.F, and Ma W.Y, "Ranking Web Objects from Multiple Communities,"

Proc. 15th ACM Int'l Conf. Information and Knowledge Management, pp. 377-386, 2006.

[4] Kalashnikov D.V and Mehrotra .S, "Domain-Independent Data Cleaning via Analysis of Entity-Relationship Graph." ACM Trans.Database Systems, vol. 31, no. 2, pp. 716-767, 2006.

[5] Kirkpatrick .S, Gelatt C.D, and Vecchi M.P, "Optimization by Simulated Annealing," Science, vol. 220, pp. 671-680, 1983.

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

[7] Melanie Weis and Felix Naumann," Relationship-BasedDuplicate Detection" Institut fur Informatik, Humboldt-Universit¨at zu Berlin Unter den Linden 6, D-10099 Berlin, Germany.

[8] Melanie Weis, Felix Naumann," Detecting Duplicates in Complex XML Data" Humboldt-University at zu Berlin Unter den Linden 6, 10099 Berlin.

[9] Nie .Z, Zhang .Y, Wen J.R, and Ma W.Y,"Object-Level Ranking:Bringing Order to Web Objects," Proc. Int'l Conf. World Wide Web (WWW), pp. 567-574, 2005.

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