A Similarity Function with Pruning Strategy for Tree Structured Data

SIMILA.K
Research Scholar
School of IT & Science, Dr.GRDCS
Simi.grd@gmail.com

SRIVIDHYA.R
Assistant Professor
School of IT & Science, Dr.GRDCS
srividhya.r@grd.edu.in

Abstract—Although several distance or similarity functions for trees have been introduced, their performance is not always satisfactory in different applications. In the base paper the Extended Sub tree (EST) function, where a new sub tree mapping is proposed. This similarity function is to compare tree structured data by defining a new set of mapping rules where sub trees are mapped rather than nodes. To reduce the time complexity as well as computational complexity of the system, efficient pruning algorithm is proposed. In the proposed system the unnecessary computation is reduced in the tree structured data by using the lossless pruning strategy. This paper provides major advancement in efficiency. This pruning strategy is ignoring the node or sub tree which has greater value than the ignoring probability. By using this technique, we can reduce the extra computation complexity.

KEYWORDS: TREE DISTANCE, TREE STRUCTURED DATA, PRUNING STRATEGY, EST (Extended Sub Tree).

1. INTRODUCTION

In today's information technology, the extensive application of tree structured data is obvious. Trees can be in the form of XML and HTML. A tree comparison is required for many applications involving tree structured data. This tree comparison is performed by tree distance and similarity functions. These applications includes document clustering, natural language clustering and automatic web testing etc., Different approaches like edit base distances, isolated sub tree distances, multi set distances, path distance, Entropy distance are used , to find out the tree distance functions. The existing work, a new similarity function for trees, namely Extended Sub tree (EST), where a new sub tree mapping is proposed. EST generalizes the edit base distances by providing new rules for sub tree mapping. This similarity function is to compare tree structured data by defining a new set of mapping rules where sub trees are mapped rather than nodes. But in the case of high dimensional data, this system has more computation complexity. Due to this characteristic, the efficiency of the system is reduced. This strategy is lossless in the sense that no duplicate objects are lost. Only object pairs incapable of reaching a given ignoring probability threshold are discarded.

2. TREE BASICS AND DEFINITIONS:

A tree data structure can be defined recursively (locally) as a collection of nodes (starting at a root node), where each node is a data structure consisting of a value, together with a list of references to nodes (the "children"), with the constraints that no reference is duplicated, and none points to the root.

Applications of trees

• Class hierarchy in Java

• file system

• storing hierarchies in organizations.

Trees are referring to rooted, ordered and labeled trees. A tree is orders if right left ordered amongst sibling nodes in the tree are important. Finally a labeled tree represents a tree where each node has an assigned label.

A tree is denoted as T and | T| indicates the size of a tree in terms of the nodes/vertices. Multiple trees are indicated by T_p and T_q. T_i represents the i th node of T numbered in a post-order format. In this paper v (T) defines the vertices or nodes and the depth of the tree is calculated using depth (T).

3. CURRENT APPROACHES:

Different types of tree distance approaches are used they are as follows,

1) EDIT BASED DISTANCES
   - I. Tree Edit Distance (TED)
   - II. Isolated Sub tree Distance (IST)
2) MULTISETS DISTANCE
3) PATH DISTANCE
4) ENTROPY DISTANCE

1) EDIT BASED DISTANCES:

This edit base distance has three operations they are delete, update and insert, its associated cost as ($W_{delete,}$ W_{insert}, and W_{update}). For example a sample tree with edit base distance is mapped. A mapping is a set of ordered integers such as (i_p, i_q) where i_p and $_{iq}$ are the index of the node from tree t_p and t_q.This means node t_p is mapped to t_q. There are some conditions to be satisfied

- ✓ One node cannot be mapped into two nodes ($i_p = i_q$)
- ✓ Sibling order preservation condition($i_p > i_q$)
- ✓ Ancestor order preservation

TED is familiar edit base distance function it measures the minimum cost between two trees. Many algorithms have been introduced to find out the optimal tree edit distance between two trees.

IST is another type of edit base distance approach. It maps between two disjoint sub trees (t_p, t_q). In these two approaches mapping are done under the restriction of structure preserving mapping.

2) MULTISETS DISTANCE:

Multi set allow repeated elements, where t_p, t_q are converted into multisets, m_p and m_q. m_p and m_q consists of all the complete sub trees of the corresponding trees. A complete sub tree is defined as a sub tree that: if t_i is a node in a complete sub tree, all of t_i's children are in the sub tree.

3) PATH DISTANCE:

In path distance approach, it considers path as a tree's building blocks. So each tree is converted into a multiset of paths such as "/a/c/d" which describes a path in t_p. One possible way is that all paths start from a root node t_i. Another approach is that any node to any possible node where a path to t_i can start from any ancestor of t_i.

4) ENTROPY DISTANCE:

Entropy distance approach is to calculate a bounded, between zero and one, distance function between two trees. This type is similar to path distance metric, the m_p and m_q multisets are generated which contain all possible paths in t_p and t_q. These are used to calculate the tree edit distances.

4. PROPOSED PRUNING STRATEGY:

The proposed distance function's performance is evaluated against TED (Tree Edit Distance), IST (Isolated Sub Tree), Entropy, Multisets, Path Distances. In this section, we propose a new similarity named pruning strategy along the EST approach. In this research, a new similarity function with respect to tree structured data is proposed, namely Extended Sub tree (EST). 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, 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. To evaluate the performance of the proposed similarity function against previous approaches, an extensive experimental study is performed. In the proposed system, the computational complexity is reduced by using the pruning strategy technique. This pruning strategy is ignoring the node or sub tree which has greater value than the ignoring probability. By using this technique, extra computation can be reduced, in other words; reduce the unnecessary computations in the system. The strategy follows the premise that, before comparing two objects, all the similarities are assumed to be 1 (i.e., the maximum possible score). The idea is to, at every step of the process; maintain an upper bound on the final probability value. 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 we assume to be 1. When we verify that the network root node probability can no longer achieve a score higher than the defined ignoring threshold, the object pair is discarded and, thus, the remaining calculations are avoided. To compute the similarity score, there are different steps

- Identify all the mappings

- Identify each node's largest mapping
- Compute the weight of each sub tree
- Calculate similarity values

In step 1, find all the possible mappings, valid or invalid, and store two lists of nodes for each mapping, one for each sub tree. T_p and T_q are the inputs to this step and V_p and V_q are the outputs.

In step 2, Let us assume P as tree1 & q as tree 2 and two arrays namely LS_p, LSq (LargestSubtree) respectively. $LS_p[i]$ indicates the largest sub tree that tp_i belongs to by the indexes of root nodes of the mapping denoted by LSp[i] and LSp[i].

In step 3, calculate W (tp_i, j) and W (tq_i, i) for all the sub trees in the mappings. Using the formula

$$W(T^{px}) = \sum_{t_i^{px} \in T^{px}} W(t_i^{px})$$

In step 4 finally compute S (Tp, Tq) based on all the possible valid mappings as:

$$S(T^p, T^q) = \alpha \sqrt{\sum_{m_k \varepsilon M} \beta_k \times W(m_k)^{\alpha}},$$

Where α; $\alpha \geq 1$, is a coefficient to adjust the relation among different sizes of mappings. It amplifies the importance of large sub trees compared to small sub trees or single nodes in accordance with the discussion in the previous section. $\alpha = 1$ does not amplify the importance of large sub trees compared to small sub trees. As α grows larger, more emphasis is placed on larger sub trees. Further, βk is a geometrical parameter which reflects the importance of the mapping with respect to the position of Tp_k and Tq_k in Tp and Tq, respectively. β_k is the unit scalar.

5. CONCLUSION:

In this existing research, the novel 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. But in the case of high dimensional data, this system has more computation complexity. Due to this characteristic, the efficiency of the system is reduced. To reduce the time complexity as well as computational complexity of the system, we proposed one efficient pruning algorithm. Further studies are required to validate the use of this tree structured data. Future research is to improve the quality, investigating the distance functions on real world applications.

REFERENCES:

[1] M.J. Zaki, "Efficiently Mining Frequent Trees in a Forest: Algorithms and Applications," IEEE Transactions. Knowledge and Data Eng., vol. 17, no. 8, pp. 1021-1035, Aug. 2005.

[2] J. Punin, M. Krishnamoorthy, and M. Zaki, "LOGML: Log Markup Language for Web Usage Mining," Process. Revised Papers from the Third Int'l Workshop Mining Web Log Data across All Customers Touch Points (WEBKDD '01), pp. 273-294, 2002.

[3] M.J. Zaki and C.C. Aggarwal, "XRules: An Effective Structural Classifier for XML Data," Proc. Ninth ACM SIGKDD Int'l Conference Knowledge Discovery and Data Mining, pp. 316-325, 2003.

[4] W. Lian, D. W. -l. Cheung, N. Mamoulis, and S.-M. Yiu, "An Efficient and Scalable Algorithm for Clustering XML Documents by Structure," IEEE Transactions. Knowledge and Data Eng., vol. 16, no. 1, pp. 82-96, Jan. 2004.

[5] M. Kouylekov and B. Magnini, "Recognizing Textual Entailment with Tree Edit Distance Algorithms," Process. First ChallengeWorkshop Recognising Textual Entailment, pp. 17-20, 2005.

[6] A. Mesbah and M.R. Prasad, "Automated Cross-Browser Compatibility Testing," Proc. 33rd Int'l Conf. Software Eng. (ICSE), pp. 561-570, 2011.

[7] A. Mesbah, A. van Deursen, and D. Roest, "Invariant-Based Automatic Testing of Modern Web Applications," IEEE Transactions Software Eng., vol. 38, no. 1, pp. 35-53, Jan./Feb. 2012.

[8] R. Connor, F. Simeoni, M. Iakovos, and R. Moss, "A Bounded Distance Metric for Comparing Tree Structure," Information Systems, vol. 36, no. 4, pp. 748-764, 2011.

[9] S. Flesca, G. Manco, E. Masciari, L. Pontieri, and A. Pugliese, "Fast Detection of XML Structural Similarity," IEEE Transaction Knowledge and Data Eng., vol. 17, no. 2, pp. 160-175, Feb. 2005.

[10] G. Valiente, "An Efficient Bottom-Up Distance between Trees," Proc. Eighth Int'l Symp. String Processing and Information Retrieval (SPIRE '01), pp. 212-219, 2001.

[11] K. Zhang, "Algorithms for the Constrained Editing Distance between Ordered Labeled Trees and Related Problems," Pattern Recognition, vol. 28, no. 3, pp. 463-474, 1995.

[12] T. Jiang, L. Wang, and K. Zhang, "Alignment of Trees—An Alternative to Tree Edit," Theoretical Computer Science, vol. 143, no. 1, pp. 137-148, 1995.