 The Open Cybernetics & Systemics Journal, 2015, 9, 110-114

Matching Similar Splits between Unrooted Leaf-labeled Trees

Li Shuguang[1, 2,*] and Xin Xiao[3]

[1]*Key Laboratory of Intelligent Information Processing in Universities of Shandong (Shandong Institute of Business and Technology), Yantai, 264005, China;* [2]*College of Computer Science and Technology, Shandong Institute of Business and Technology, Yantai, 264005, China;* [3]*College of Foreign Studies, Shandong Institute of Business and Technology, Yantai, 264005, China*

Abstract: Tree comparison is ubiquitous in many areas. The simplest way for tree comparison is to define a pairwise distance measure. In a more refined comparison, one can establish a mapping between similar parts in two trees according to certain similarity measure. The best match problem for rooted leaf-labeled trees has been studied in the literature. However, no result has been found for the best match problem for unrooted leaf-labeled trees. The problem of mapping similar splits between unrooted leaf-labeled trees is considered in this paper. Based on a new similarity measure obtained from the classical Jaccard coefficient, the mapping can be computed in quadratic time.

Keywords: Jaccard coefficient, leaf-labeled trees, similarity measure, splits, tree comparison.

1. INTRODUCTION

Trees are suitable structures for representing data for which hierarchical relations can be defined. They have been utilized in many areas, such as bioinformatics [1], image processing [2], natural language processing [3, 4], document analysis [5], to name just a few. A leaf-labeled tree is a tree labeled only at the leaves. More precisely, each leaf of a leaf-labeled tree is assigned a distinct label. Such trees arise in the areas such as music comparison and retrieval [6, 7], classfication [8-10], phylogenetics [11-13], etc.

Comparison of trees is a recurrent task in many areas mentioned above. The most popular method for tree comparison is to define a pairwise distance measure. Many such distance measures for leaf-labeled trees have been proposed in the literature [11, 14-19]. However, in many situations, a single distance value is not adequate. It is better to establish a mapping between similar parts in two trees according to certain similarity measure. Such a mapping is useful in determining corresponding parts in compared trees, especially in the analysis of large trees [20-23].

In a rooted leaf-labeled tree, each vertex associates with a cluster, *i.e.*, the set of leaves under it. Hence, it is quite natural to establish the mapping between similar clusters in two rooted trees. The consensus tree method [24, 25] only computes the mapping between perfectly matching vertices, the pairs of vertices with identical clusters. For vertices having no perfectly matching vertices, the mapping is undefined. The *s*-consensus tree method [26, 27] computes the mapping between best matching vertices instead of perfectly matching vertices. The measure used in [26, 27] is the classical Jaccard coefficient [28, 29]: the similarity $J(A,B)$ between two sets A, B is defined to be the size of their intersection divided by the size of their union. One nice property of Jaccard coefficient is that $1-J(A,B)$ is a metric [30, 31]. L Zhang [32] studied the following best match problem for rooted leaf-labeled trees: computes for every vertex in a tree the best matching vertex in another tree according to Jaccard coefficient. He presented two algorithms whose worst case time complexities are $O(n^2 \log n)$ and $O((n \log n)^{1.5})$ respectively, where n is the number of leaves in a tree.

In an unrooted leaf-labeled tree, each edge associates with a split (a bipartition of the label set). Hence, it is natural to establish the mapping between similar splits in two unrooted trees rather than between vertices or clusters in rooted trees. We introduce the following best match problem for unrooted leaf-labeled trees: computes for every split in a tree the best matching split in another tree according to certain similarity measure. To the best of our knowledge, this problem has not been studied so far. In this paper, we define a new similarity measure for comparing pairs of splits. The proposed measure can be regarded as an extension of Jaccard coefficient, and also has the nice property that the function defined by one minus it is a metric. We propose two algorithms to solve the best match problem for unrooted leaf-labeled trees according to this measure, with worst case time complexity $O(n^3)$ and $O(n^2)$ respectively, where n is the number of leaves in a tree.

The remainder of this paper is organized as follows. In Section 2, after reviewing terminology needed, we define a new similarity measure, and prove that the function defined

1874-110X/15 2015 Bentham Open

by one minus this measure is a metric. In Section 3, we present two algorithms to solve the best match problem for unrooted leaf-labeled trees according to this similarity measure. We conclude this paper in Section 4.

2. PRELIMINARIES

A *tree* is a connected undirected graph with no cycles. A *leaf-labeled tree* is a tree whose leaves are labeled bijectively by a set L and each non-leaf vertex is unlabeled and has degree at least 3. Let $|L| = n$. Denote by T_n the set of leaf-labeled trees over L.

Let T be a leaf-labeled tree over L. Any two vertices of T are connected by a unique path. Cutting an edge from T induces a *split* (*bipartition*), *i.e.*, a partition of L into two non-empty sets. Denote the split whose blocks are A and B by $A \mid B$. Since the position of A and B is arbitrary, we make no distinction between the splits $A \mid B$ and $B \mid A$. If $\min\{|A|,|B|\} = 1$, then $A \mid B$ is *trivial*, otherwise it is *non-trivial*. Clearly, each pedant edge associates with a trivial split which must be present in every tree, while each internal edge associates with a nontrivial split. Denote by $\Sigma(T)$ the collection of the splits induced by the edges of T. There exist efficient algorithms for reconstructing T from $\Sigma(T)$ [33, 34].

Similarity measure is commonly used in clustering and similarity searching of large structure files. One of the most popular similarity measures is Jaccard coefficient, which is defined on two sets A and B as $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ [28, 29]. Similarity is somewhat opposite to the concept of distance between structures. The Jaccard coefficient can be used to define a distance $1 - J(A,B)$ (the so-called Soergel distance [35, 36]), which is indeed a metric.

Definition 1. [37] A metric on a set S is a function $d : S \times S \to R^{\geq 0}$ such that, for all $x, y, z \in S$, the following hold:

(i) $d(x,y) = 0$ if and only if $x = y$ (definiteness);

(ii) $d(x,y) = d(y,x)$ (symmetry);

(iii) $d(x,z) \leq d(x,y) + d(y,z)$ (triangle inequality).

The pair (S,d) is called a metric space.

The triangle inequality is a desirable mathematical property. It ensures that any two structures having low dissimilarity to a third structure will have low dissimilarity to each other.

Lemma 1. [30, 31] $1 - J(A,B)$ is a metric.

We are now ready to define a similarity measure for comparing splits between unrooted leaf-labeled trees.

Let $\Sigma_n = \bigcup_{T \in T_n} \Sigma(T)$. The *split similarity* of two splits $A_1 \mid B_1$ and $A_2 \mid B_2$ in Σ_n, is defined as follows:

$$SSi(A_1 \mid B_1, A_2 \mid B_2) = \frac{1}{2}\max\{\frac{|A_1 \cap A_2|}{|A_1 \cup A_2|} + \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|}, \frac{|A_1 \cap B_2|}{|A_1 \cup B_2|} + \frac{|B_1 \cap A_2|}{|B_1 \cup A_2|}\}. \quad (1)$$

Theorem 1. $1 - SSi$ is a metric on Σ_n.

Proof. The first two properties of Definition 1 (definiteness and symmetry) are trivially true. Presented here is a proof for the triangle inequality.

Let $A_1 \mid B_1$, $A_2 \mid B_2$ and $A_3 \mid B_3$ be three arbitrary splits in Σ_n. For possible combinations of $SSi(A_1 \mid B_1, A_2 \mid B_2)$ and $SSi(A_2 \mid B_2, A_3 \mid B_3)$, we distinguish between the following four cases:

(i) $SSi(A_1 \mid B_1, A_2 \mid B_2) = \frac{1}{2}(\frac{|A_1 \cap A_2|}{|A_1 \cup A_2|} + \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|})$,

$SSi(A_2 \mid B_2, A_3 \mid B_3) = \frac{1}{2}(\frac{|A_2 \cap A_3|}{|A_2 \cup A_3|} + \frac{|B_2 \cap B_3|}{|B_2 \cup B_3|})$.

(ii) $SSi(A_1 \mid B_1, A_2 \mid B_2) = \frac{1}{2}(\frac{|A_1 \cap A_2|}{|A_1 \cup A_2|} + \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|})$,

$SSi(A_2 \mid B_2, A_3 \mid B_3) = \frac{1}{2}(\frac{|A_2 \cap B_3|}{|A_2 \cup B_3|} + \frac{|B_2 \cap A_3|}{|B_2 \cup A_3|})$.

(iii) $SSi(A_1 \mid B_1, A_2 \mid B_2) = \frac{1}{2}(\frac{|A_1 \cap B_2|}{|A_1 \cup B_2|} + \frac{|B_1 \cap A_2|}{|B_1 \cup A_2|})$,

$SSi(A_2 \mid B_2, A_3 \mid B_3) = \frac{1}{2}(\frac{|A_2 \cap A_3|}{|A_2 \cup A_3|} + \frac{|B_2 \cap B_3|}{|B_2 \cup B_3|})$.

(iv) $SSi(A_1 \mid B_1, A_2 \mid B_2) = \frac{1}{2}(\frac{|A_1 \cap B_2|}{|A_1 \cup B_2|} + \frac{|B_1 \cap A_2|}{|B_1 \cup A_2|})$,

$SSi(A_2 \mid B_2, A_3 \mid B_3) = \frac{1}{2}(\frac{|A_2 \cap B_3|}{|A_2 \cup B_3|} + \frac{|B_2 \cap A_3|}{|B_2 \cup A_3|})$.

We choose to prove the triangle inequality for the second case. The other three cases can be proved similarly. We need to show that

$$1 - SSi(A_1 \mid B_1, A_3 \mid B_3) \leq 1 - SSi(A_1 \mid B_1, A_2 \mid B_2) + 1 - SSi(A_2 \mid B_2, A_3 \mid B_3). \quad (2)$$

By the definition of the split dissimilarity, we have:

$$1 - SSi(A_1 \mid B_1, A_3 \mid B_3) \leq 1 - \frac{1}{2}\left(\frac{|A_1 \cap B_3|}{|A_1 \cup B_3|} + \frac{|B_1 \cap A_3|}{|B_1 \cup A_3|}\right) = \frac{1}{2}\left(1 - \frac{|A_1 \cap B_3|}{|A_1 \cup B_3|} + 1 - \frac{|B_1 \cap A_3|}{|B_1 \cup A_3|}\right).$$

Since we are proving the second case, we have:

$$1 - SSi(A_1 \mid B_1, A_2 \mid B_2) = 1 - \frac{1}{2}\left(\frac{|A_1 \cap A_2|}{|A_1 \cup A_2|} + \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|}\right) = \frac{1}{2}\left(1 - \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|} + 1 - \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|}\right).$$

$$1 - SSi(A_2 \mid B_2, A_3 \mid B_3) = 1 - \frac{1}{2}\left(\frac{|A_2 \cap B_3|}{|A_2 \cup B_3|} + \frac{|B_2 \cap A_3|}{|B_2 \cup A_3|}\right) = \frac{1}{2}\left(1 - \frac{|A_2 \cap B_3|}{|A_2 \cup B_3|} + 1 - \frac{|B_2 \cap A_3|}{|B_2 \cup A_3|}\right).$$

By Lemma 1, we have:

$$1 - \frac{|A_1 \cap B_3|}{|A_1 \cup B_3|} \leq 1 - \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|} + 1 - \frac{|A_2 \cap B_3|}{|A_2 \cup B_3|},$$

$$1 - \frac{|B_1 \cap A_3|}{|B_1 \cup A_3|} \leq 1 - \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|} + 1 - \frac{|B_2 \cap A_3|}{|B_2 \cup A_3|}.$$

Combining the above inequalities, we get the desired inequality (2).

3. THE ALGORITHMS

Given two unrooted leaf-labeled trees T_1 and T_2 in T_n. For a split $\sigma_1 \in \Sigma(T_1)$, the *best match* $M(\sigma_1) \in \Sigma(T_2)$ of σ_1 is the split that maximizes the split similarity between σ_1 and any split in $\Sigma(T_2)$, *i.e.*, $M(\sigma_1) = \arg\max_{\sigma_2 \in \Sigma(T_2)} SSi(\sigma_1, \sigma_2)$. The best match problem for unrooted leaf-labeled trees is to compute for every split in T_1 the best match in T_2 according to the split similarity measure.

In this section we will present two algorithms to solve this problem. Since each trivial split of T_1 is also present in T_2, we need only find the best matches for the nontrivial splits of T_1. However, it is possible that the best match for a nontrivial split of T_1 is a trivial split of T_2. Hence, when we compute for every nontrivial split of T_1 the best match in T_2, we have to compare it with every split of T_2.

Without loss of generality, we assume that the label set $L = \{1, 2, \ldots, n\}$. Since there is a one-one correspondence between the leaf set and the label set, we also denote the leaf set of a tree by $L = \{1, 2, \ldots, n\}$.

Given an unrooted leaf-labeled tree $T \in T_n$. We root T at the leaf n to get a rooted leaf-labeled tree T'. Clearly, each vertex v of T' except the root corresponds to a split $L(v) \mid (L - L(v))$ of T, where $L(v)$ denotes the cluster associated with v.

Algorithm 1:

Step 1: Root T_1 and T_2 at the leaf n to get rooted leaf-labeled trees T_1' and T_2'.

Step 2: Traverse T_1' and T_2' in post-order respectively. During the traversal, compute and store nontrivial splits of T_1 and all splits of T_2.

Step 3: For each nontrivial split $\sigma_1 \in \Sigma(T_1)$, compute the best match of σ_1, $M(\sigma_1) = \arg\max_{\sigma_2 \in \Sigma(T_2)} SSi(\sigma_1, \sigma_2)$.

We then get the following theorem.

Theorem 1. Algorithm 1 solves the best match problem for unrooted leaf-labeled trees T_1 and T_2 in $O(n^3)$ time, where n is the number of leaves in T_1 and T_2.

Proof. Steps 1, 2 and 3 can be executed in $O(n)$, $O(n^2)$ and $O(n^3)$ time, respectively. Hence the running time of Algorithm 1 is $O(n^3)$.

We next modify Algorithm 1 such that the best match of each nontrivial split $\sigma_1 \in \Sigma(T_1)$ can be computed in linear time and the time complexity can be reduced to $O(n^2)$.

Fix a nontrivial split $\sigma_1 = A_1 \mid B_1 \in \Sigma(T_1)$. Suppose that $A_1 \mid B_1$ corresponds to vertex v_1' of T_1', where $A_1 = L(v_1')$. Traverse T_2' in post-order. Suppose that the current vertex being checked in T_2' is v_2', and v_2' corresponds to the split $\sigma_2 = A_2 \mid B_2 \in \Sigma(T_2)$, where $A_2 = L(v_2')$. Let $p = |A_1|$, $q = |A_1 \cap A_2|$, $r = |B_1 \cap A_2|$. Store the values of p, q, r. We can use p, q, r to calculate the following values, and then get $SSi(A_1 \mid B_1, A_2 \mid B_2)$ by (1).

$$\frac{|A_1 \cap A_2|}{|A_1 \cup A_2|} = \frac{q}{p+r}, \quad \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|} = \frac{n-p-r}{n-q},$$

$$\frac{|A_1 \cap B_2|}{|A_1 \cup B_2|} = \frac{p-q}{n-r}, \quad \frac{|B_1 \cap A_2|}{|B_1 \cup A_2|} = \frac{r}{n-p+q}.$$

For the leaves of T_2', the values of q and r can be obtained in $O(1)$ time. For any interior vertex of T_2', the values of q and r can be computed by adding respectively the values of q and r of all the children. Hence for each nontrivial split $\sigma_1 \in \Sigma(T_1)$, the best match can be computed in $O(n)$ time.

Algorithm 2:

Step 1: Root T_1 and T_2 at the leaf n to get rooted leaf-labeled trees T_1' and T_2'.

Step 2: Traverse T_1' in post-order, and store nontrivial splits of T_1.

Step 3: For each nontrivial split $\sigma_1 \in \Sigma(T_1)$, traverse T_2' in post-order, and during the traversal compute $M(\sigma_1)$ in linear time using the method described above.

We then get the following theorem.

Theorem 2. Algorithm 2 solves the best match problem for unrooted leaf-labeled trees T_1 and T_2 in $O(n^2)$ time, where n is the number of leaves in T_1 and T_2.

CONCLUSION

We defined a similarity measure for comparing splits between unrooted leaf-labeled trees having the property that the function defined by one minus this measure is considered a metric. We studied the best match problem according to this measure for unrooted leaf-labeled trees, and presented two algorithms with cubic and quadratic time complexities respectively. Note that the best match problem for rooted leaf-labeled trees can be solved in sub-quadratic time in the worst case. It would be interesting to investigate whether the best match problem for unrooted leaf-labeled trees can be solved in sub-quadratic time.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (Nos. 61173173, 61272016, 61272430, 61373079 and 61432010), Key project of Chinese Ministry of Education (No. 212101), Shandong Provincial Natural Science Foundation of China (Nos. ZR2013FM015 and ZR2011FL004).

REFERENCES

[1] R. B. Russell, and G. J. Barton, "Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels," *Proteins: Structure, Function, and Bioinformatics*, vol. 14, pp. 309-323, 1992.

[2] R. A. Finkel, and J. L. Bentley, "Quad trees a data structure for retrieval on composite keys," *Acta informatica*, vol. 4, pp. 1-9, 1974.

[3] G. Sampson, "A proposal for improving the measurement of parse accuracy," *International Journal of Corpus Linguistics*, vol. 5, pp. 53-68, 2000.

[4] G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology*, vol. 37, pp. 51-89, 2003.

[5] M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger, "The Penn Treebank: annotating predicate argument structure," In: *Proceedings of the workshop on Human Language Technology*, pp. 114-119, 1994.

[6] A. Habrard, J. M. Inesta, D. Rizo, and M. Sebban, "Melody recognition with learned edit distances," In: *Structural, Syntactic, and Statistical Pattern Recognition,* Springer: US, pp. 86-96, 2008.

[7] D. Rizo, K. Lemström, and J. M. Inesta, "Tree representation in combined polyphonic music comparison," In: *Computer Music Modeling and Retrieval. Genesis of Meaning in Sound and Music,* Springer: US, pp. 177-195, 2009.

[8] E. N. Adams, "Consensus techniques and the comparison of taxonomic trees," *Systematic Biology*, vol. 21, pp. 390-397, 1972.

[9] W. H. Day, "Optimal algorithms for comparing trees with labeled leaves," *Journal of Classification*, vol. 2, pp. 7-28, 1985.

[10] T. N. Phyu, "Survey of classification techniques in data mining," In: *Proceedings of the International MultiConference of Engineers and Computer Scientists*, pp. 18-20, 2009.

[11] D. Robinson, and L. R. Foulds, "Comparison of phylogenetic trees," *Mathematical Biosciences*, vol. 53, pp. 131-147, 1981.

[12] B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp, and L. Zhang, "On distances between phylogenetic trees," In: *Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 427-436, 1997.

[13] C.-M. Lee, L.-J. Hung, M.-S. Chang, and C.-Y. Tang, "An improved algorithm for the maximum agreement subtree problem", In: *Proceedings of 4th IEEE Symposium on Bioinformatics and Bioengineering BIBE* 2004, pp. 533-536.

[14] G. F. Estabrook, F. McMorris, and C. A. Meacham, "Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units," *Systematic Biology*, vol. 34, pp. 193-200, 1985.

[15] M. A. Steel and D. Penny, "Distributions of tree comparison metrics-some new results," *Systematic Biology*, vol. 42, pp. 126-141, 1993.

[16] D. Bogdanowicz and K. Giaro, "Matching split distance for unrooted binary phylogenetic trees," In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, pp. 150-160, 2012.

[17] Y. Lin, V. Rajan, and B. M. Moret, "A metric for phylogenetic trees based on matching," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, pp. 1014-1022, 2012.

[18] D. Bogdanowicz, and K. Giaro, "On a matching distance between rooted phylogenetic trees," *International Journal of Applied Mathematics and Computer Science*, vol. 23, pp. 669-684, 2013.

[19] S. Böcker, S. Canzar, and G. W. Klau, "The generalized robinson-foulds metric," *Algorithms in Bioinformatics,* Springer: US, pp. 156-169, 2013.

[20] N. Shivakumar and H. G. Molina, "*SCAM: A copy detection mechanism for digital documents,"* In: *2nd International Conference in Theory and Practice of Digital Libraries*, Austin, Taxas,1995.

[21] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic clustering of the web," *Computer Networks and ISDN Systems*, vol. 29, pp. 1157-1166, 1997.

[22] A. Z. Broder, "On the resemblance and containment of documents," In: *Proceedings Compression and Complexity of Sequences* 1997, pp. 21-29, 1997.

[23] T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou, "TreeJuxtaposer: scalable tree comparison using Focus+ Context with guaranteed visibility," In: *ACM Transactions on Graphics (TOG*), pp. 453-462, 2003.

[24] T. Margush, and F. R. McMorris, "Consensusn-trees," *Bulletin of Mathematical Biology*, vol. 43, pp. 239-244, 1981.

[25] E. O. Wiley, D. S. Causey, D. R. Brooks, and V. Funk, *The Compleat Cladist: A Primer of Phylogenetic Procedures, vol. 19: Museum of Natural History*. University of Kansas Lawrence: Kansas, 1991.

[26] R. Stinebrickner, "s-Consensus trees and indices," *Bulletin of Mathematical Biology*, vol. 46, pp. 923-935, 1984.

[27] R. Stinebrickner, "s-Consensus index method: An additional axiom," *Journal of Classification*, vol. 3, pp. 319-327, 1986.
[28] P. Jaccard, "Nouvelles recherches sur la distribution florale," *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 44, pp. 223-270, 1908.
[29] M. M. Deza, and E. Deza, *Encyclopedia of Distances*. Springer: US, 2009.
[30] H. Spath, *Cluster Analysis Algorithms for Data Reduction and Classification of Objects: Ellis Horwood, Ltd*. Chichester, England, 1980.
[31] A. H. Lipkus, "A proof of the triangle inequality for the Tanimoto distance," *Journal of Mathematical Chemistry*, vol. 26, pp. 263-265, 1999.
[32] L. Zhang, "*On matching nodes between trees*," Tech. Rep. 2003-67, HP Labs 2003.
[33] C. A. Meacham, "Theoretical and computational considerations of the compatibility of qualitative taxonomic characters," In: *Numerical taxonomy,* Springer: US, pp. 304-314, 1983.
[34] D. Gusfield, "Efficient algorithms for inferring evolutionary trees," *Networks*, vol. 21, pp. 19-28, 1991.
[35] J. C. Gower, "Measures of similarity, dissimilarity, and distance," *Encyclopedia of statistical sciences*, vol. 5, p. 3, 1985.
[36] P. Willett, J. M. Barnard, and G. M. Downs, "Chemical similarity searching," *Journal of chemical information and computer sciences*, vol. 38, pp. 983-996, 1998.
[37] C. Semple and M. A. Steel, *Phylogenetics*. Oxford University Press: England, 2003.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014