A Fragment-Based Similarity Measure for Concept Hierarchies and Ontologies

Hui Yang
Department of Computer Science
Georgetown University
Washington DC, 20057, USA
huiyang@cs.georgetown.edu

ABSTRACT

Despite the popularity of concept hierarchies and ontologies, such as Yahoo! Directory, a similarity measure that considers both hierarchy content and topology and is highly efficient has not yet been reached. A commonly used metric, Tree Edit Distance, exhibits extreme inefficiency when measuring similarities between unordered hierarchies. In this paper, we propose a novel and feasible solution, Fragment-based Similarity (FBS), to serve as an efficient and effective measurement for hierarchy similarity evaluation. Experimental results and a user study show that FBS not only well-approximates but also is more efficient than Tree Edit Distance.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Experimentation

Keywords: Hierarchy Evaluation, Fragment-based Similarity

1. INTRODUCTION

From UNIX file systems to Web services such as Yahoo! Directory and Open Directory Project (ODP), concept hierarchies are widely used in many areas. Despite the popularity of categorical hierarchies, measuring similarity between hierarchies and ontologies remains a challenging problem.

Oftentimes categorical hierarchies are expressed as unordered trees, i.e., ordering among sibling nodes is ignored. Two hierarchies are considered similar if they are similar in both content and topology. However, most hierarchy similarity measures used in Information Retrieval (IR) fail to account for hierarchy topology. For example, set-based metrics, such as word overlaps [2], and precision and recall of parent-child node pairs [1], rely solely on content of the hierarchies. On the other hand, Tree Edit Distance (TED) is popular for its effectiveness in measuring the distance between hierarchies, and accounting for both content and topology. However, computing TED between unordered hierarchies has proven to be NP-complete and seriously limits its practical uses [3]. Therefore, it is necessary to explore a new metric that not only evaluates both hierarchy content and topology, but also is highly efficient.

We observe that when comparing two hierarchies, it is fragments of similar concepts that first capture our eyes. As such, we are inspired to design a novel approach, Fragment-based Similarity (FBS), which represents hierarchies in vector space model and compares hierarchies fragment by fragment. Experiment results show that the proposed method not only achieves a much higher efficiency than TED, but also well-approximates TED to effectively measure both content and topology similarity between categorical hierarchies.

2. METHODOLOGY

The gist of our method is to express a hierarchy with a set of vectors, each representing a non-leaf node by the presence and absence of its descendant subtree. With this vector space representation, the ordering among the nodes within a subtree is ignored; hence the unordered nature of hierarchies is preserved. Each non-leaf node and its subtree are considered as a *fragment*.

Specifically, in a hierarchy, each vector corresponds to a fragment. The vector consists of 1s and 0s indicating a word's presence and absences in the fragment, and its length equals the size of the vocabulary (number of unique words in the entre hierarchy). Figure 1 presents two hierarchies $\boldsymbol{H_a}$ and $\boldsymbol{H_b}$.. Figure 2 demonstrates their respective vector representation.

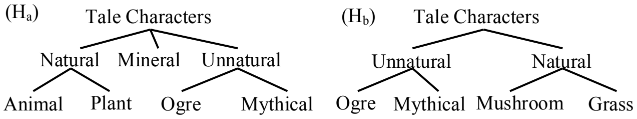

Figure 1. Hierarchy examples $\boldsymbol{H_a}$ and $\boldsymbol{H_b}$

(H_a)	TC	Natual	Animal	Plant	Unnatural	Ogre	Mythical	Mineral	Mushroom	Grass
TC	0	1	1	1	1	1	1	1	0	0
Natural	0	0	1	1	0	0	0	0	0	0
Unnatural	0	0	0	0	0	1	1	0	0	0
(H_b)	TC	Natual	Animal	Plant	Unnatural	Ogre	Mythical	Mineral	Mushroom	Grass
TC	0	1	0	0	1	1	1	0	1	1
Natural	0	0	0	0	0	0	0	0	1	1
Unnatural	0	0	0	0	0	1	1	0	0	0

Figure 2. Vector representation of hierarchy $\boldsymbol{H_a}$ and $\boldsymbol{H_b}$

Once hierarchies are represented as vectors, the calculation of FBS between them can be carried out in two steps: identifying matching fragments and aggregating the similarity scores.

2.1 Identifying Matching Fragments

To find out matching fragments within two hierarchies T_i and T_j, we exhaustively calculate the pairwise similarity between each fragment pair in them. Particularly, we calculate the cosine similarity between two fragments $t_p \subseteq T_i \, t_q \subseteq T_j$ as follows:

$$sim_{\cos}(t_p, t_q) = (\vec{t_p} \bullet \vec{t_q}) / (\|\vec{t_p}\| \|\vec{t_q}\|) \tag{1}$$

We can align matching fragments based on their cosine similarities. We investigate three matching criteria: *All Pairs*, all pairs of fragments whose cosine similarity is above 0.1; *Maximum Matched Subtree*, the fragments containing the greatest number of nodes and whose cosine similarity is above 0.1; and *Highest Valued Subtree*, the fragments with the highest and above 0.1 cosine similarity. Suppose Z is the number of words in the vocabulary, *M* is the number of nodes in T_i, and *N* is the number of nodes in T_j. the time complexity of calculating matching fragments is $O(MNZ) = O(N^3)$. We compare the above three matching criteria in Section 3.1.

2.2 Aggregating Similarity Scores

After identifying the matching fragments, we aggregate their similarity scores to obtain the final FBS score. The following shows how the overall FBS similarity is calculated.

$$FBS(T_i,T_j)=\frac{1}{D}\sum_{p=1}^{m} sim_{\cos}(t_{ip},t_{jp}) \qquad (2)$$

where D is the denominator and m is the number of matched fragment pairs. There are also three choices for denominator D: (a) the number of matched fragment pairs m; (b) the max hierarchy size of T_i and T_j; and (c) the max number of non-leaf nodes of T_i and T_j. We compare the choices in Section 3.1.

Since the aggregation only needs to calculate an average, its time complexity is *O(1)*. Thus, the overall time complexity of calculating FBS remains as $O(N^3)$. Compare to TED's time complexity as NP-complete, FBS is much more efficient.

3. EXPERIMENTS

In this section, we empirically evaluate FBS. We firstly explore the parameters of FBS and then exhibit its efficiency. We also conduct a user study and show that FBS is able to approximate TED by generating highly correlated similarity rankings.

3.1 Approximating TED and Parameter Selection

To show that FBS well-approximate TED, we show that the ranked lists generated for a group of hierarchies to one reference hierarchy by FBS and TED are highly correlated. We employ Spearman's correlation coefficient ρ [3] to compare these ranked lists. The bigger the correlation between the output rankings, the more equivalent the two methods are.

Table 1 displays the Spearman's ρ values for ranked lists generated by FBS and TED averaged over 50 random hierarchies at each representative size. The results indicate that FBS is able to approximate TED very well – many runs are above 0.9. Moreover, the best runs are highlighted in bold fonts. It indicates that the best matching criterion is Highest Valued Subtree and the best denominator is Hierarchy size.

3.2 Efficiency Comparison

Although theoretically we have known that FBS's time complexity is $O(N^3)$, which is much better than Ted's NP-complete, we compare their empirical efficiency for small-sized hierarchies in this experiment. Table 2 shows the mean running time of TED and FBS averaged over 500 random runs. The results are calculated on random hierarchies at different sizes, with Highest Value Subtree as the matching criterion, and Hierarchy size as the denominator. The results indicate that FBS requires statistically significant less running time than TED ($p<0.001$, t-test). As the tree size increases, the running time of TED rises at an exponential speed while the running time of FBS maintains a mild rising tendency.

3.3 User Study

To further test FBS' ability to approximate TED in measuring hierarchy similarity, a user study is conducted by recruiting participants from Amazon Mechanical Turk. We extracted 22 topics from the Open Project Directory (ODP) hierarchies and used them in the user study. Among the topics, ten are about clothing, six jewelries, and six household items.

In each task, the participants are presented with three hierarchies. One of them is used as the reference hierarchy (A) and the other two (B and C) are compared against it. FBS and TED calculate similarity scores for (A,B) and (A,C) and make their own decisions about whether B or C is more similar to A. The participants judged all 22 tasks and they were asked to judge with their intuition that which metric makes a better decision; and judge whether the decisions are correct if their decisions agree. There are several possible outcomes for the human assessments: (1) TED and FBS reach the same decisions; (2) FBS and TED reach different decisions; (3) When they reach different decisions, TED better matches with human intuition; and (4) When they reach different decisions, TED better matches with human intuition.

Table 1: Spearman's ρ of FBS and Tree Edit Distance.

Tree Size	*20*	*100*	*500*	*1000*
Denominator = # matched pairs				
Match=All pair	0.75	0.87	0.87	0.86
Match=Max.Subtree	0.79	0.90	0.91	0.93
Match=High. Value	0.80	0.90	0.92	0.94
Matching criterion = highest value				
D= #Matched	0.80	0.90	0.92	0.94
D= **#Non-leaf**	**0.80**	**0.92**	**0.96**	**0.97**
D= **Hier. size**	**0.80**	**0.92**	**0.96**	**0.97**

Table 2. Running time of TED and FBS

Tree Size	TED (msec)	FBS (msec)
10	52.08	1.28
15	247.24	2.36
20	3,117.36	4.69
25	19,741.84	8.39
30	211,446.41	13.74

Table 3. Majority Vote of Human Evaluation.

Both FBS and TED are true	12
Both FBS and TED are false	3
FBS is better	3
TED is better	4

Table 3 shows the majority vote from the responses in the user study. It shows that out of the 22 tasks, FBS and TED agree on 15 tasks; which gives an agreement of 68%. For the cases where FBS and TED do not agree, FBS is a better measure for three tasks, while TED for four. When TED and FBS agree, they match the participants' intuition for 12 tasks and contradict for three. In summary, both FBS and TED's results coincide with human intuition very well, and their decisions are rather comparable. It shows that FBS is a good approximation to TED.

4. CONCLUSION

Measuring similarity between concept hierarchies or ontologies is a challenging problem. This paper presents fragment-based similarity (FBS), a simple and feasible solution to this problem. Instead of comparing the entire hierarchy, our approach compares fragments between hierarchies and aggregates their similarity values as the final score. A comparison between the new metric and Tree Edit Distance (TED) shows that the proposed metric can generate similar rankings of similarities as tree edit distance, in but much more efficient than it.

ACKNOWLEDGMENT

This research was supported by NSF grant CNS-1223825.

5. REFERENCES

[1] Yifen Huang. "Mixed-Initiative Clustering". Ph.D. dissertation, Carnegie Mellon University, pp.79-80, 2010.

[2] Kummamuru, Krishna, et al. A hierarchical monothetic document clustering algorithm for summarizatison and browsing search results. In *Proceedings of the 13th conference on World Wide Web,* pages 658-665, 2004.

[3] D. D. Wackerly, W. Mendenhall, and R. L. Scheaffer. Mathematical Statistics with Applications. Duxbury advanced Series. 2002.

[4] Kaizhong Zhang and Tao Jiang. Some max snp-hard results concerning unordered labeled trees. In *Information Processing Letters,* 49:249-254, 1994.