A graph matching method based on probe assignments

Romain Raveaux, Jean-Christophe Burie, Jean-Marc Ogier

A graph matching method based on probe assignments

Romain Raveaux, Jean-christophe Burie and Jean-Marc Ogier.

L3I, University of La Rochelle, av M. Crépeau, 17042 La Rochelle Cedex 1, France, E-mail: {romain.raveaux01}@univ-lr.fr

Abstract. In this paper, a graph matching method and a distance between attributed graphs are defined. Both approaches are based on graph probes. Probes can be seen as features exctracted from a given graph. They represent a local information. According two graphs G_1,G_2, the univalent mapping can be expresssed as the minimum-weight probe matching between G_1 and G_2 with respect to the cost function c.

1 Probe Matching and Probe Matching Distance

1.1 Probes of graph

Let L_V and L_E denote the set of node and edge labels, respectively. A labeled, undirected graph G is a 4-tuple $G = (V, E, \mu, \xi)$, where

- V is the set of nodes,
- $E \subseteq V \times V$ is the set of edges
- $\mu : V \rightarrow L_V$ is a function assigning labels to the nodes, and
- $\xi : E \rightarrow L_E$ is a function assigning labels to the edges.

From this definition of graph, probes of graph for the matching problem can be expressed as follow:

Let G be an attributed graphs with edges labeled from the finite set $\{l1, l2, ..., a\}$. Let P be a set of probes extracted from G. There is a probe p for each vertex of the graph G. A probe (p) is defined as a pair $< V_i, H_i >$ where H_i is an edge structure for a given vertex (V_i), H_i is a $2a$-tuple of non-negative integers $\{x_1, x_2, ..., x_a, y_1, y_2, ..., y_a\}$ such that the vertex has exactly x_i incoming edges labeled l_i, and y_j outgoing edges labeled l_j.

1.2 Probe Matching

Let $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ be two attributed graphs. Without loss of generality, we assume that $\mid P1 \mid \geq \mid P2 \mid$. The complete bipartite graph $G_{em}(V_{em} = P1 \cup P2 \cup \triangle, P1 \times (P2 \cup \triangle))$, where $\triangle$ represents an empty dummy probe, is

called the probe matching graph of $G1$ and $G2$. A probe matching between $G1$ and $G2$ is defined as a maximal matching in G_{em}. Let there be a non-negative metric cost function $c : P1 \times P2 \rightarrow \Re_0^+$. We define the matching distance between $G1$ and $G2$, denoted by $d_{match}(G1, G2)$, as the cost of the minimum-weight probe matching between $G1$ and $G2$ with respect to the cost function c.

1.3 Cost function for probe matching

Let p_1 and p_2 be two probes. The cost function can be expressed as a distance between two probes : $c(p_1, p_2) =\mid V_1 - V_2 \mid + \mid H_1 - H_2 \mid$

1.4 Time complexity analysis

The matching distance can be calculated in $O(n^3)$ time in the worst case. To calculate the matching distance between two attributed graphs $G1$ and $G2$, a minimum-weight probe matching between the two graphs has to be determined. This is equivalent to determining a minimum-weight maximal matching in the probe matching graph of $G1$ and $G2$. To achieve this, the method of Kuhn [1] and Munkres [2] can be used. This algorithm, also known as the Hungarian method, has a worst case complexity of $O(n^3)$, where n is the number of probes in the larger one of the two graphs [3].

1.5 The probe matching distance for attributed graphs is a metric.

Proof. To show that the probe matching distance is a metric, we have to prove the three metric properties for this similarity measure.

- $d_{match}(G_1, G_2) >= 0$
  The probe matching distance between two graphs is the sum of the cost for each probe matching. As the cost function is non-negative, any sum of cost values is also non-negative.
- $d_{match}(G_1, G_2) = d_{match}(G_2, G_1)$
  The minimum-weight maximal matching in a bipartite graph is symmetric, if the edges in the bipartite graph are undirected. This is equivalent to the cost function being symmetric. As the cost function is a metric, the cost for matching two probes is symmetric. Therefore, the probe matching distance is symmetric.
- $d_{match}(G_1, G_2) <= d_{match}(G_1, G_2) + d_{match}(G_2, G_3)$
  As the cost function is a metric, the triangle inequality holds for each triple of probes in G1, G2 and G3 and for those probes that are mapped to an

empty probe. The probe matching distance is the sum of the cost of the matching of individual probes. Therefore, the triangle inequality also holds for the probe matching distance.

1.6 The probe matching distance is a lower bound for the edit distance.

Given a cost function for the edge matching which is always less than or equal to the cost for editing an probe, the matching distance between attributed graphs is a lower bound for the edit distance between attributed graphs:

$$\forall G_1, G_2 : d_{match}(G_1, G_2) <= d_{ED}(G_1, G_2) \tag{1}$$

Proof. The edit distance between two graphs is the number of edit operations which are necessary to make those graphs isomorphic. To be isomorphic, the two graphs have to have identical probe sets. As the cost function for the probe matching distance is always less than or equal to the cost to transform two probes into each other through an edit operation, the probe matching distance is a lower bound for the number of edit operations, which are necessary to make the two probe sets identical. It follows that the edge matching distance is a lower bound for the edit distance between attributed graphs.

2 Experiments

2.1 Protocol

In this paragraph, we assess the correlation concerning the responses to k-NN queries when using edit distance, graph probing or probe matching distance as dissimilarity measures. The setting is the following: in a graph dataset we select a number N of graphs, that are used to query by similarity the rest of the dataset. Top k responses to each query obtained in the first place using edit distance, graph probing and probe matching distance. These k responses are compared using Kendalland correlation coefficient while the k distance values are evaluated using Pearson correlation. We consider a null hypothesis of independence between the two responses and then, we compute by means of a two-sided statistical hypothesis test the probability (p-value) of getting a value of the statistic as extreme as or more extreme than that observed by chance alone, if H0 is true. Kendall's rank correlation measures the strength of monotonic association between the vectors x and y (x and y may represent ranks or ordered categorical variables). Kendall's rank correlation coefficient τ may be expressed as $\tau = \frac{S}{D}, where$

$$S = \sum_{i<j} (sign(x[i] - y[i]).sign(y[i] - x[i])) \tag{2}$$

$$D = \frac{k(k-1)}{2} \tag{3}$$

2.2 Data Set Description

The last database used in the experiments consists of graphs representing distorted letter drawings. In this experiment we consider the 15 capital letters of the Roman alphabet that consists of straight lines only (A, E, F, ...). For each class, a prototype line drawing is manually constructed. To obtain arbitrarily large sample sets of drawings with arbitrarily strong distortions, distortion operators are applied to the prototype line drawings. This results in randomly shifted, removed, and added lines. These drawings are then converted into graphs in a simple manner by representing lines by edges and ending points of lines by nodes. Each node is labeled with a two-dimensional attribute giving its position, since our approach only focuses on nominal attributes, a quantification is performed by the use of a bi-dimensional mesh Fig.1. More information concerning these data set is detailed on table 1.

Table 1. Characteristics of the data set used in our computational experiments

	Base D
Number of classes (N)	15
$\mid Training \mid$	3796
$\mid Test \mid$	1266
$\mid Validation \mid$	1688
Average number of nodes	4.7
Average number of edges	3.6
Average degree of nodes	1.3
Max number of nodes	9
Max number of edges	7

Using N = 400, K = 30, we present in Tab.3,Tab.4 and Fig.2, the results obtained in terms of τ and cor values. From the 400 tests (Tab. 2), only 45 have a p-value greater than 0.05, so we can say that the hypothesis H0 of independence is rejected in 88.75% cases, with a significance level of 0.05. The observed correlation obtained for k-NN queries, strengthen our decision to use a faster (and simpler) dissimilarity measure than edit distance in order to perform a graph classification. Moreover, the Probe Matching Distance outperfom the Graph Probing in terms of linear relation with the edit distance while keeping a reasonnable time complexity Tab.3.

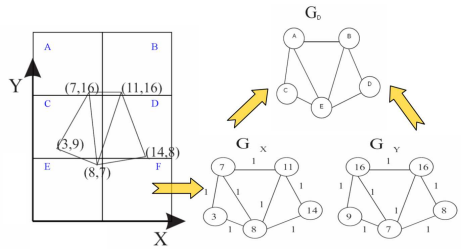


Fig. 1. From symbols to graphs using a 2D mesh

Table 2. Summary of P-values

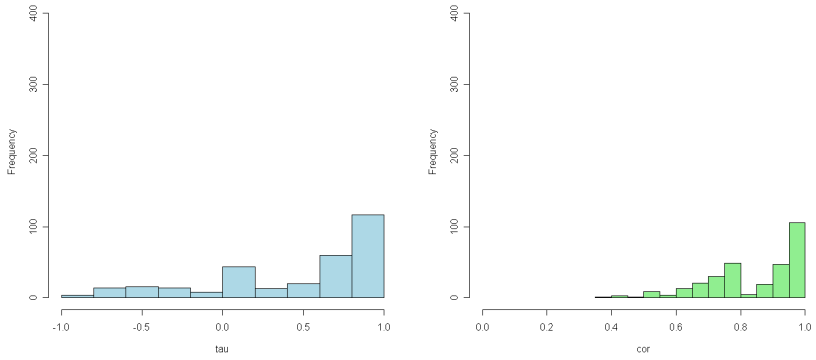
	Minimum	1st quantile	Median	Mean	3rd quantile	Maximum
$p - values$	0.000	0.000	2.682e-6	5.018e-5	1.455e-3	8.278e-1

Table 3. Summary of correlations for the Graph Probing Dissimilarity Measure

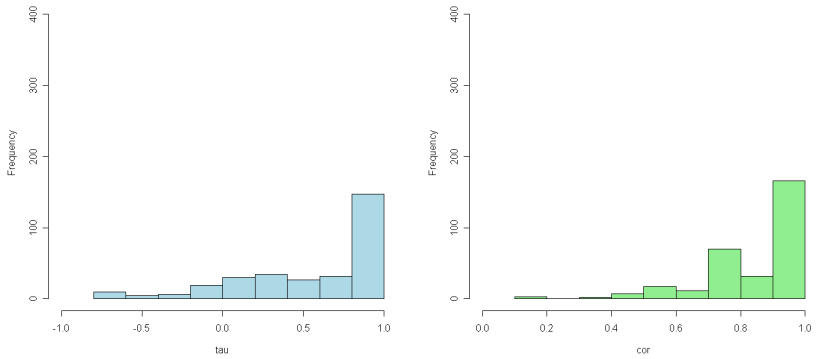
	Minimum	1st quantile	Median	Mean	3rd quantile	Maximum
τ	-0.93170	0.05542	0.73330	0.48350	0.99900	1.00000
cor	0.3541	0.7384	0.8951	0.8484	1.0000	1.0000

Table 4. Summary of correlations for the Probe Matching Distance

	Minimum	1st quantile	Median	Mean	3rd quantile	Maximum
τ	-0.7216	0.2853	0.7429	0.5704	1.0000	1.0000
cor	0.1571	0.7358	0.9785	0.8619	1.0000	1.0000

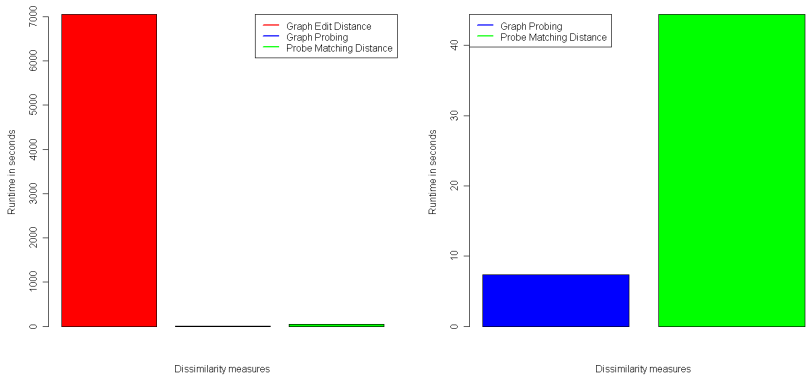


(a) tau GP (b) cor GP



(c) tau PMD (d) cor PMD

Fig. 2. Histogram of correlations



(a) (b)

Fig. 3. Time complexity

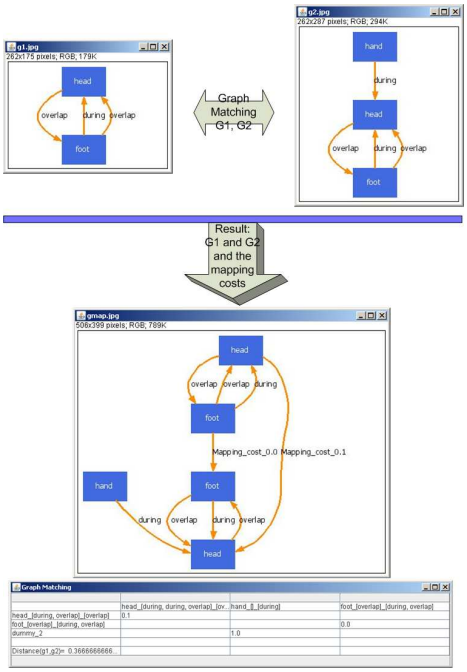


Fig. 4. Graph matching between two graphs g1, and g2. Gmap represents g1 and g2 in a single graph with the addition of ”cost” edges to map node of g1 in g2

References

1. H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.
2. J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society of Industrial and Applied Mathematics*, 5(1):32–38, March 1957.
3. Hans-Peter Kriegel and Stefan Schonauer. Similarity search in structured data.