ELSEVIER

Pattern Recognition Letters 22 (2001) 701–704

Pattern Recognition Letters

www.elsevier.nl/locate/patrec

Graph distances using graph union

W.D. Wallis [a], P. Shoubridge [b,*], M. Kraetz [b], D. Ray [c]

[a] *Southern Illinois University, USA*

[b] *Communication Division, Defence Science and Technology Organisation, PO Box. 1500, Salisbury, SA 5108, Australia*

[c] *Department of Defense, USA*

Received 23 December 1999; received in revised form 8 January 2001

Abstract

An existing graph distance metric based on maximum common subgraph has been extended by a proposal to define the problem size with the union of the two graphs being measured, rather than the larger of the two graphs used in the existing metric. For some applications the graph distance measure is more appropriate if the graph union approach is used. This graph distance measure is shown to be a metric. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Subgraph isomorphism; Graph distance; Maximum common subgraph

1. Introduction

Various problems exist where a measure of difference between two graphs is required (Sanil et al., 1995; Sarkar and Boyer, 1998; Shoubridge et al., 1999a,b). Graph and subgraph isomorphism algorithms determine whether or not two graphs are the same, or whether one graph is an isomorphic subgraph of the other (Bondy and Murty, 1976; Corneil and Gotlieb, 1970; Myaeng and Lopez-Lopez, 1992; Ullman, 1976). Techniques that identify the maximum common subgraph of two graphs provide a representation of the commonality between graphs (Horaud and Skordas, 1989; Levinson, 1992). Error tolerant graph matching, typically using methods such as error correcting graph matching, can be used as a graph distance measure indicating the difference between graphs (Bunke and Messmer, 1997). A distance metric has also been defined based on the determination of the maximal common subgraph (Bunke and Shearer, 1998) and shown to be related to error correcting graph matching (Bunke, 1997). In order to make numerical comparisons between differences in different situations, it is desirable that measures of graph difference have metric properties.

2. Distances using maximum common subgraph

The distance metric based on maximum common subgraph is of the form

$$d(G_1, G_2) = 1 - \frac{m(G_1, G_2)}{M(G_1, G_2)}, \tag{1}$$

where $m(G_1, G_2)$ is a measure of similarity between G_1 and G_2 and $M(G_1, G_2)$ is a measure of the size

* Corresponding author. Tel.: +61-8-8259-6841; fax: +61-8-8259-7110.

E-mail address: peter.shoubridge@dsto.defence.gov.au (P. Shoubridge).

PII: S0167-8655(01)00022-8

of the problem. For example, Bunke and Shearer (1998) use

$$m(G_1, G_2) = |G_{12}|, \quad M(G_1, G_2) = \max\{|G_1|, |G_2|\}, \tag{2}$$

where G_{12} is a maximal common subgraph of G_1 and G_2, and $|G|$ denotes the number of vertices of the graph G. The size of the problem may also be defined as the number of vertices in the union of the two graphs. This distance measure is also a metric.

If $d(G_1, G_2)$ is to be a metric, four properties are required.

1. $0 \leqslant m(G_1, G_2) \leqslant M(G_1, G_2) \neq 0$ (actually if both were negative and $M(G_1, G_2) \leqslant m(G_1, G_2)$ it would suffice, but the only practical reason for allowing negative measures is if m and M are both directed, e.g., $m(G_1,G_2) = -m(G_2,G_1)$, and this is easily handled by taking absolute values);
2. $m(G_1, G_2) = M(G_1, G_2)$ if and only if G_1 and G_2 are isomorphic;
3. $(m(G_1,G_2)/M(G_1,G_2)) = (m(G_2,G_1)/M(G_2,G_1))$ for all G_1 and G_2 (this is usually handled by requiring both m and M to be symmetric);
4. $d(G_1, G_3) \leqslant d(G_1, G_2) + d(G_2, G_3)$ for all G_1, G_2 and G_3.

Of the four conditions, only the last seems to present any difficulty. In our experience, the reasonable candidate functions so far considered (such as spectral measures (Umeyama, 1988), error correcting graph matching (Sanfeliu and Fu, 1983), weight-difference based measures (Parkes and Wallis, 1978; Umeyama, 1988), graph edit distance (Bunke, 1997), maximal common subgraph distance (Bunke and Shearer, 1998) – most of these are also described in the survey (Bunke and Messmer, 1997)) satisfy 1, 2 and 3. Bunke and Shearer (1998) proved that (1) is a metric if (2) is used. This clearly remains true if $|G|$ is replaced by $e(G)$, the number of edges of G, or if (additive) edge weights are allowed.

3. Using the union

We propose the following measure: $m(G_1, G_2)$ is as before, and $M(G_1, G_2) = |G_1| + |G_2| - |G_{12}|$.

The reason for this is as follows. In some applications it seems natural to view the size of the problem as the size of the union of two graphs being compared. Using the union rather than the larger of the two graphs distinguishes variations in the size of the smaller of the two graphs. If only the size of the larger graph is used to represent problem size, the distance between graphs will remain unchanged even if the smaller graph changes its size, assuming that the size of the maximum common subgraph remains constant. Therefore, for some applications representing the problem size by the union of the two graphs may provide a more appropriate measure of relative graph difference.

As an example, consider three graphs G_a, G_b and G_c that represent the information transactions in a communications network on three different days, and we are interested in detecting the difference in information transactions between those days. Each graph shares a common set X of 20 vertices, while G_a has 20 further vertices, G_b has 10 and and G_c has 5 (these 35 vertices are all different). Say the size of a graph, in this particular application, is interpreted as the number of vertices. If we use $M(G_1, G_2) = \max\{|G_1|, |G_2|\}$, we find $d(G_a, G_b) = d(G_a, G_c) = 0.50$, while $M(G_1, G_2) = |G_1| + |G_2| - |G_{12}|$ gives $d(G_a, G_b) = 0.60$, $d(G_a, G_c) = 0.56$. For the purposes of this application, we are interested in the fact that G_a and G_b differ in 30 vertices, while G_a and G_c differ in only 25. As a result, the latter measure using graph union is more appropriate in detecting the differences in communications.

It is not clear how "union" should be interpreted when graphs which are isomorphic, but not equal, are treated as equivalent. There are isomorphisms that preserve G_{12} but identify none, some or all of the vertices of G_1 outside the maximal common subgraph with vertices of G_2. For example, consider a set of seven vertices $\{a, b, c, d, e, f, g\}$ with links ab, bc, ca. If G_1 is the induced subgraph with vertices $\{a, b, c, d, e\}$ and G_2 is the induced subgraph with vertices $\{a, b, c, f, g\}$, the following maps α, β and γ are of the three types described:

$$\alpha(a) = a, \alpha(b) = b, \alpha(c) = c, \alpha(d) = d, \alpha(e) = e;$$
$$\beta(a) = a, \beta(b) = b, \beta(c) = c, \beta(d) = d, \beta(e) = f;$$
$$\gamma(a) = a, \gamma(b) = b, \gamma(c) = c, \gamma(d) = f, \gamma(e) = g.$$

In pattern-recognition applications, it is not clear when such identification is appropriate. The measure proposed here might be called the "largest-possible" union.

Consider any three (unlabelled) graphs G_1, G_2 and G_3. Suppose G_2 is labelled first. Without loss of generality, G_1 can be labelled so that the vertices of a maximal common subgraph G_{12} receive the same labels in G_1 and G_2, and similarly for G_2 and G_3. Moreover, the vertices of G_1 and G_3 not in the maximal subgraphs are all labelled differently from the remaining vertices of G_2 and (for the moment) from each other. Then $G_{12} = G_1 \cap G_2$ and $M(G_1, G_2) = |G_1 \cup G_2|$, and again similarly for G_{23}. The same holds in case of labelled G_1, G_2, G_3.

Notice that we can always choose a labelling so that $|G_1 \cap G_2 \cap G_3| = |G_1 \cap G_3|$, and that $G_1 \cap G_2 \cap G_3$ is a common subgraph of G_1 and G_3. It may not be maximal, but certainly one has $|G_1 \cap G_3| \leqslant |G_{13}|$ (with equality holding in case of labelled G_1 and G_3). Therefore,

$$\begin{aligned} M(G_1, G_3) &= |G_1| + |G_3| - |G_{13}| \\ &\leqslant |G_1| + |G_3| - |G_1 \cap G_3| = |G_1 \cup G_3|, \end{aligned}$$

so

$$\begin{aligned} d(G_1, G_3) &= 1 - \frac{|G_{13}|}{|G_1| + |G_3| - |G_{13}|} \\ &\leqslant 1 - \frac{|G_{13}|}{|G_1 \cup G_3|} \leqslant 1 - \frac{|G_1 \cap G_3|}{|G_1 \cup G_3|}. \end{aligned}$$

To show that d is a metric under this new interpretation, it is necessary to verify that

$$d(G_1, G_2) + d(G_2, G_3) \geqslant d(G_1, G_3),$$

for any graphs G_1, G_2 and G_3. Assuming the graphs have been labelled as above, it is sufficient to prove

$$\left(1 - \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|}\right) + \left(1 - \frac{|G_2 \cap G_3|}{|G_2 \cup G_3|}\right) \geqslant \left(1 - \frac{|G_1 \cap G_3|}{|G_1 \cup G_3|}\right). \tag{3}$$

Suppose the vertices are represented in a standard Venn diagram. The sizes of the sets will be denoted as follows:

- There are x_{123} vertices in $G_1 \cap G_2 \cap G_3$;
- There are x_{12} vertices in $G_1 \cap G_2$ but not in G_3;
- There are x_1 vertices in G_1 but not in $G_2 \cup G_3$,

and so on. It is convenient to write T for the total number of vertices:

$$T = x_1 + x_2 + x_3 + x_{12} + x_{13} + x_{23} + x_{123}.$$

Then, (3) becomes

$$\left(1 - \frac{x_{12} + x_{123}}{T - x_3}\right) + \left(1 - \frac{x_{23} + x_{123}}{T - x_1}\right) \geqslant \left(1 - \frac{x_{13} + x_{123}}{T - x_2}\right),$$

or

$$1 + \frac{x_{13} + x_{123}}{T - x_2} - \frac{x_{12} + x_{123}}{T - x_3} - \frac{x_{23} + x_{123}}{T - x_1} \geqslant 0.$$

Clearing the denominators, we have

$$\begin{aligned} &(T - x_1)(T - x_2)(T - x_3) \\ &\quad + (x_{13} + x_{123})(T - x_1)(T - x_3) \\ &\quad - (x_{12} + x_{123})(T - x_1)(T - x_2) \\ &\quad - (x_{23} + x_{123})(T - x_2)(T - x_3) \geqslant 0. \end{aligned} \tag{4}$$

The left-hand side can be written as a cubic polynomial in seven variables with all non-negative terms: for example,

$$\begin{aligned} &(2T - x_1 - x_3)(x_{13}T + x_2 x_{123}) + T[(x_2 + x_3)x_{23} \\ &\quad + (x_1 + x_2)x_{12} + x_1 x_3] + (T - x_{12} - x_3)x_1 x_2 \\ &\quad + (T - x_{23})x_2 x_3 + x_1 x_3 (x_{13} + x_{123}). \end{aligned}$$

So (4), and therefore (3), are always satisfied and the measure formed by using this m and M in (1) is a metric. Obviously, the calculation is the same if the number of edges is used instead of the number of vertices as a measure of magnitude.

4. Conclusion

In some applications a more appropriate graph distance measure, based on maximal common subgraph, can result by using the union of the two graphs as a representation of problem size. It has been shown in the paper that such a distance measure is a metric. If the number of edges is used

as a magnitude, instead of the number of vertices, the distance measure is still a metric.

References

Bondy, J.A., Murty, U.S.R., 1976. Graph Theory with Applications. Macmillan, UK.

Bunke, H., 1997. On a relation between graph edit distance and maximum common subgraph. Pattern Recognition Lett. 18 (8), 689–694.

Bunke, H., Messmer, B.T., 1997. Recent advances in graph matching. Internat. J. Pattern Recognition Artificial Intell. 11 (1), 169–203.

Bunke, H., Shearer, K., 1998. A graph distance metric based on the maximal common subgraph. Pattern Recognition Lett. 19, 255–259.

Corneil, D.G.C.C., Gotlieb, C.C., 1970. An efficient algorithm for graph isomorphism. J. ACM 17, 51–64.

Horaud, R., Skordas, T., 1989. Stereo correspondence through feature grouping and maximal cliques. IEEE Trans. Pattern Anal. Machine Intell. 11 (11), 1168–1180.

Levinson, R., 1992. Pattern associativity and the retrieval of semantic networks. Comput. Math. Appl. 23, 573–600.

Myaeng, S.H., Lopez-Lopez, A., 1992. Conceptual graph matching: a flexible algorithm and experiments. J. Experiment Theoret. Artificial Intell. 4, 107–126.

Parkes, D.N., Wallis, W.D., 1978. Graph theory and the study of activity structure. Timing Space and Spacing Time, volume 2: Human Activity and Time Geography. Edward Arnold, London, UK, pp. 75–99.

Sanfeliu, A., Fu, A.K.S., 1983. A distance measure between attributed relational graphs for pattern recognition. IEEE Trans. Systems Man. Cybernet. 13, 353–362.

Sanil, A., Banks, D., Carley, K., 1995. Models for evolving fixed node networks: model fitting and model testing. Social Networks 17, 65–81.

Sarkar, S., Boyer, K.L., 1998. Quantitative measures of change based on feature organization: eigenvalues and eigenvectors. Computer Vision Image Understanding 71 (1), 110–136.

Shoubridge, P., Kraetzl, M., Ray, D., 1999a. Detection of abnormal change in dynamic networks. In: IEEE Information, Decision and Control, IDC '99 Conference, Adelaide, Australia, pp. 557–562.

Shoubridge, P.J., Kraetzl, M., Bunke, H., Wallis, W.D., 1999b. Approaches to Measuring Network Change. In: Proceedings of the DASPII, LaSalle, IL.

Ullman, J.R., 1976. An algorithm for subgraph isomorphism. J. ACM 23 (1), 31–42.

Umeyama, S., 1988. An eigendecomposition approach to weighted graph matching problems. IEEE Trans. Pattern Anal. Machine Intell. 10, 695–703.