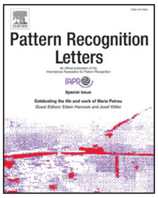
Contents lists available at ScienceDirect
Pattern Recognition Letters
journal homepage: www.elsevier.com/locate/patrec



A generalization of the most common subgraph distance and its application to graph editing

Michael Hecht*

Bioinformatics group, Department of Computer Science, University Leipzig, Härtelstr. 16-18, D-04107, Leipzig, Germany

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Graph editing
Most common subgraph problem
Tree editing
Sequence alignment
Edit distance

ABSTRACT

We relate the *graph editing distance* to a generalized weighted version of the *most common subgraph distance*. To do so, we introduce the new concepts of *isotonic shifts* and *vector weighted graphs*. As a consequence we can give a weak but sufficient condition on cost models to result in an *edit metric*, ensuring the richness of the class of these functions. Moreover, for arbitrary instances we are able to determine a within cubic time computable upper bound on the edit distance, which equals the minimized distance for infinitely many instances.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Summarizing, the concept of *graph editing* is to transform one graph into another graph by deleting, inserting and substituting as less vertices and edges as possible. This problem occurs in various fields of discrete mathematics and computer sciences. For instance the famous algorithms of Needleman [9] and Wunsch [15] are used to compute optimal graph editings in the case of linear ordered strings. The concept is also used for more complex graphs as for instance the representations of RNA molecules [10,11,16]. In *Big Data* sets often graphs are used to represent the structures of data, e.g., the relations in a social network. The question of how similar a given graph (relation) is to another can be modeled as an instance of the graph editing problem. Hence, important applications despite the classical origin in molecular chemistry have recently developed. In this context two questions are of high interest. Firstly: can the edit distance be computed effectively? Secondly: how can an effective solutions of the first question be used to effectively compute a clustering of graphs with respect to their edit distance ? Since already in the case of unordered strings the graph editing problem turns out to be NP-complete, see [14], question one has to be denied in general, unless $P = NP$. Thus, if it would be possible to effectively compute a close a priori upper bound on the edit distance the search space of possible edit operations can enormously be decreased. Consequently, the exact distance might be computed effectively a posteriori or at least the estimation can effectively be improved to be sufficient close. In regard of the second question, it would be enormous helpfull if we can ensure that the edit distance is a (pseudo) metric. In this case, the information that a set of graphs $\{G_1, \ldots, G_n\}$ is close to a graph G_0 implies that the diameter of the set $\{G_0, \ldots, G_n\}$ is small, i.e., if the edit distance of G_i and G_0 is bounded by d for all $i = 1, \ldots, n$ then the distance of G_i to G_j is bounded by $2d$ for all $1 \leq i, j \leq n$ due to the triangle inequality, which is still close for small d. Thus, instead of comparing all pairs of graphs, it might suffice to compare all graphs with respect to some preference set of graphs yielding a speed up of the performance by degree 1. Due to the high importance of the graph editing problem many approaches were already made. However, dealing with formal mathematical questions is often omitted or only roughly sketched. Therefore, in addition to generate algorithmical improvements this article is written with the aim to provide the mathematical foundations, which are crucial in this context.

1.1. Outline of the article

First we formally introduce the concept of graph editing by using the classical notion of *isotonic mappings* in Section 2. Furthermore, we state the question whether the corresponding edit distance with respect to given edit costs results in a metric. Though, for instance in [13] an answer to this question could be given in the special case of ordered trees, in general a satisfacting answer to this problem was not given yet. Since the complexity class of the graph editing problem often depends on the vertex orderings of the given instances, every consideration is done in regard of partially ordered graphs. In particular in Section 3 we introduce the new concepts of *vector weighted graphs* and *isotonic shifts operators*, which enable us to show that a generalized version of the *most common subgraph distance* is a (pseudo)-metric. Moreover, we establish a 1: 1 correspondence between the notion of isotonic

[*] Corresponding author. Tel.: +4917670027983; Fax: +4917670027983.
E-mail address: hecht@mpi-cbg.de

http://dx.doi.org/10.1016/j.patrec.2016.09.008
0167-8655/© 2016 Elsevier B.V. All rights reserved.

shifts and isotonic mappings. As a consequence, the translation of the most common subgraph distance results in a sufficient condition on cost functions to induce an edit (pseudo) metric. In fact, the condition allows a much wider class of cost functions than the special type considered in [2]. In Section 4 we use these results to give a *non-heuristical upper bound* on the edit distance, which can be computed in cubic time. Finally, the most relevant applications, which are the problem of *clustering graphs* and improvements of the performances of *graph editing solvers* are discussed.

2. Graph editing

To introduce the concept of graph editing we reformulate the notion of Bunke [2] in a more formal version.

Definition 1. Let $G_1 = (V_1, E_1), G_2 = (V_2, E_2)$ be two simple graphs $\le_{V_1}$ and $\le_{V_2}$ be partial orders on V_1, V_2 respectively. Let furthermore $V_1' \subseteq V_1$ and $V_2' \subseteq V_2$ be such that there is a isotonic bijection $f : V_1' \longrightarrow V_2'$, i.e., if $v_1 \le_{V_1} v_2$ then $f(v_1) \le_{V_2} f(v_2)$. We set $E_1' = E_1 \cap V_1' \times V_1'$, $E_2' := E_2 \cap V_2' \times V_2'$ and furthermore

$$E_{1,f} := \left\{ (u,v) \in E_1' \;\middle|\; \big(f(u), f(v)\big) \in E_2' \right\},$$
$$E_{2,f} := \left\{ (x,y) \in E_2' \;\middle|\; \big(f^{-1}(x), f^{-1}(y)\big) \in E_1' \right\}.$$

Now we define the bijection

$$F_f : E_{1,f} \longrightarrow E_{2,f}\,, \quad \text{by} \quad F\big((u,v)\big) = \big(f(u), f(v)\big)\,.$$

Thus, $(f, F_f) : (V_1', E_{1,f}) \longrightarrow (V_2', E_{2,f})$ is a graph isomorphism, which we call a *mapping* between G_1 and G_2 and shortly denote with *F*. Furthermore, $M(G_1, G_2)$ shall denote the set of all isotonic mappings between G_1 and G_2.

Note that the partial orders can also be trivial, i.e., $v \le_{V_i} u \Longleftrightarrow v = u,\ \ i = 1, 2$. Thus, the case of unordered graphs is included in our considerations.

Definition 2. Let G_1 and G_2 be two given graphs. We consider two types of *objective* or *cost functions* $C_a, C_b : M(G_1, G_2) \longrightarrow \mathbb{R}^+$ given by

$$\begin{aligned} C_a(F) = & \sum_{v \in V_1 \setminus V_1'} c_{del}(v) + \sum_{v \in V_2 \setminus V_2'} c_{ins}(v) + \sum_{v \in V_1'} c_{sub}(v) \\ & + \sum_{e \in E_1 \setminus E_{1,f}} d_{del}(e) + \sum_{e \in E_2 \setminus E_{2,f}} d_{ins}(e) \\ & + \sum_{e \in E_{1,f}} d_{sub}(e)\,, \\ C_b(F) = & \sum_{v \in V_1'} c_{sub}(v) + \sum_{e \in E_{1,f}} d_{sub}(e) \\ & + \max \Bigg\{ \sum_{v \in V_1 \setminus V_1'} c_{del}(v) + \sum_{e \in E_1 \setminus E_{1,f}} d_{del}(e), \\ & \sum_{v \in V_2 \setminus V_2'} c_{ins}(v) + \sum_{e \in E_2 \setminus E_{2,f}} d_{ins}(e) \Bigg\}\,, \end{aligned}$$

where we assume that $c_{del} : V_1 \longrightarrow \mathbb{R}^+$, $c_{ins} : V_2 \longrightarrow \mathbb{R}^+$, $d_{del} : E_1 \longrightarrow \mathbb{R}^+$, $d_{ins} : E_2 \longrightarrow \mathbb{R}^+$ are arbitrary functions, $c_{sub} : V_1 \longrightarrow \mathbb{R}^+$ is defined with respect to some function

$$D_V : V_1 \times V_2 \longrightarrow \mathbb{R}^+ \quad \text{via} \quad c_{sub}(v) = D_V(v, f(v))$$

and $d_{sub} : E_1 \longrightarrow \mathbb{R}^+$ is defined with respect to some function

$$D_E : E_1 \times E_2 \longrightarrow \mathbb{R}^+ \quad \text{via} \quad d_{sub}(e) = D_E(e, F(e))\,.$$

We formulate the following problem.

Problem 1. Let G_1, G_2 be finite simple graphs and $\le_{V_1}$ and $\le_{V_2}$ be partial orders on V_1, V_2 respectively. Let $C = \alpha C_a + \beta C_b : M(G_1, G_2) \longrightarrow \mathbb{R}^+$, α, $\beta \ge 0$, $\alpha + \beta = 1$ be a cost function of merged type. Then compute a mapping F in the set of all optimal isotonic mappings

$$F_{\mathrm{opt}}(C) := \left\{ F \in M(G_1, G_2) \;\middle|\; C(F) = \min_{F' \in M(G_1, G_2)} C(F') \right\}$$

with respect to C. We denote with $\Omega(G_1, G_2, C) := C(F)$, $F \in F_{\mathrm{opt}}(C)$ the optimal score.

Remark 1. Assume that $F \in F_{\mathrm{opt}}(C)$ is an optimal mapping between G_1 and G_2 with respect to C then F defines a set of edit operations, i.e., F decodes how to delete and substitute as less vertices and edges as possible with respect to their costs to transform G_1 into G_2.

Remark 2. If G_1 and G_2 are labelled with the same alphabet, i.e., there are maps

$$\sigma_{V_i} : V_i \longrightarrow \Sigma_V \qquad \sigma_{E_i} : E_i \longrightarrow \Sigma_E\,, \qquad i = 1, 2\,.$$

Then it suffices to define the maps $\overline{c}_{del} : \Sigma_V \longrightarrow \mathbb{R}^+$, $\overline{c}_{ins} : \Sigma_V \longrightarrow \mathbb{R}^+$, $\overline{d}_{del} : \Sigma_E \longrightarrow \mathbb{R}^+$, $\overline{d}_{ins} : \Sigma_E \longrightarrow \mathbb{R}^+$, $\overline{D}_V : \Sigma_V \times \Sigma_V \longrightarrow \mathbb{R}^+$ and $\overline{D}_E : \Sigma_E \times \Sigma_E \longrightarrow \mathbb{R}^+$ to determine a cost function C by setting

$$c_{del} := \overline{c}_{del} \circ \sigma_{V_1}\,, \quad c_{ins} := \overline{c}_{ins} \circ \sigma_{V_2}\,,$$
$$d_{del} := \overline{d}_{del} \circ \sigma_{E_1}\,, \quad d_{ins} := \overline{d}_{ins} \circ \sigma_{E_2}\,,$$

and

$$D_V := \overline{D}_V \circ (\sigma_{V_1}, \sigma_{V_2}) \quad D_E := \overline{D}_E \circ (\sigma_{E_1}, \sigma_{E_2})\,.$$

In the Examples 2 and 3 we will come back to this observation.

In the next section we give an answer to the following question : Let a collection of graphs $\mathcal{G} = \{G_l\}_{1 \le l \le q}$ and a collection of cost functions $\mathcal{C} = \{C_{ij}\}_{1 \le i,j \le q}$ be given, does

$$d_{\mathcal{C}}(G_i, G_j) := \Omega(G_i, G_j, C_{ij}) \tag{1}$$

become a (pseudo) metric on $\mathcal{G}$?

3. Vector weighted graphs

To answer the question of Section 2 we will generalize the concept of edge and vertex weights to *vectorial weights*. This concept enables us to relate the *most common subgraph distance* to the *edit distance* and results in a sufficient condition on cost functions to induce an edit (pseudo) metric.

Definition 3. Let $G = (V, E)$ be a graph and $\nu : V \longrightarrow \mathbb{R}^N$, $\varepsilon : E \longrightarrow \mathbb{R}^N$, $N \in \mathbb{N}$ be maps with non-negative entries, i.e, $(\nu(v))_k$, $(\varepsilon(e))_k \ge 0$ for all $v \in V$, $e \in E$, $1 \le k \le N$. Then we call (G, ν, ε) a *vector weighted graph*. Assume that V possesses a partial order $\le_V$, then we choose an isotonic embedding $\rho_V : V \hookrightarrow \mathbb{N}$, i.e.,

$$\text{if} \quad v_1 \le_V v_2 \Longrightarrow \rho(v_1) \le_{\mathbb{N}} \rho(v_2)\,.$$

We set $\rho_E([u, v]) = [\rho(u), \rho(v)]$ for $[u, v] \in E$ and denote with $\varrho(G) = \big(\rho_V(V), \rho_E(E)\big)$ the embedded graph. Furthermore, the maps $\bar{\nu} : \mathbb{N} \longrightarrow \mathbb{R}^N$ and $\bar{\varepsilon} : \mathbb{N} \times \mathbb{N} \longrightarrow \mathbb{R}^N$ defined by

$$\bar{\nu}_\rho(k) := \begin{cases} \nu\big(\rho^{-1}(k)\big), & \text{if } k \in \rho_V(V) \\ 0, & \text{else} \end{cases}$$

$$\bar{\varepsilon}_\rho(k, l) := \begin{cases} \varepsilon\big(\rho^{-1}(k), \rho^{-1}(l)\big), & \text{if } (k, l) \in \rho_E(E) \\ 0, & \text{else} \end{cases}$$

shall denote the natural extension of ν, ε to $\mathbb{N}$ and $\mathbb{N} \times \mathbb{N}$ with respect to the embedding ϱ. Moreover, we define the *vector weighted adjacency operator* or just *weighted adjacency operator* $A = A(\varrho(G), \bar{\nu}, \bar{\varepsilon}) \in \mathbb{R}^{n \times n} \times \mathbb{R}^N$ with $n := \max(\rho_V(V))$ by

$$A = (a_{ij})_{1 \le i,j \le n}\,, \quad \text{with } a_{ij} = \begin{cases} \bar{\nu}(i), & \text{if } i = j \\ \bar{\varepsilon}(i, j), & \text{else} \end{cases}\,.$$

Please cite this article as: M. Hecht, A generalization of the most common subgraph distance and its application to graph editing, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.09.008

Download English Version:

https://daneshyari.com/en/article/4970287

Download Persian Version:

https://daneshyari.com/article/4970287

Daneshyari.com