Časopis pro pěstování matematiky

Vladimír Baláž; Jaroslav Koča; Vladimír Kvasnička; Milan Sekanina

A metric for graphs

Časopis pro pěstování matematiky, Vol. 111 (1986), No. 4, 431--433

Persistent URL: http://dml.cz/dmlcz/118290

Časopis pro pěstování matematiky, roč. 111 (1986), Praha

A METRIC FOR GRAPHS*)

VLADIMÍR BALÁŽ, Bratislava, JAROSLAV KOČA, Brno, VLADIMÍR KVASNIČKA, Bratislava, MILAN SEKANINA, Brno

(Received June 20, 1984)

1. The purpose of this communication is to use the concept of maximal common subgraph for defining a distance (with properties of the metric) between graphs. The approach presented here was initially stimulated by our recent studies on the mathematical model of organic chemistry [1], where we have dealt very often with "measuring similarity" of graphs. A special kind of the distance has been reported in this journal by Zelinka [2]; his approach is entirely based on the number of vertices in both compated graphs and does not reflect in an explicit way their edges.

2. A graph $G = (V, E)$ consists of a non-empty finite vertex set V and an edge set E. The graphs considered in the present paper may be, in general, directed or undirected, multiple edges and loops are permitted. A subgraph G' of the graph G is a graph obtained from G by deleting subsets of its vertices and edges, $G' \subseteq G$. Two graphs G_1 and G_2 are isomorphic, $G_1 \sim G_2$, if there exists a $1-1$ correspondence between the vertices of one and the vertices of the other such that the adjacent pairs of vertices in one graph are mapped only to adjacent pairs in the other. A common subgraph of two graphs G_1 and G_2 consists of a subgraph $G_1' \subseteq G_1$ and a subgraph $G_2' \subseteq G_2$ such that $G_1' \sim G_2'$. A maximal common subgraph (MCS) of two graphs is the common subgraph which contains the largest possible number of edges. Recently, McGregor [3] has suggested a back-track searching algorithm for the construction of MCS of two graphs.

3. Let us consider two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, and let $G_1' = (V_1^{(1,2)}, E_1^{(1,2)}) \subseteq G_1$ and $G_2' = (V_2^{(1,2)}, E_2^{(1,2)}) \subseteq G_2$ be the MCS. Since the subgraphs G_1' and G_2' are isomorphic, $G_1' \sim G_2'$, the cardinalities of the vertex and the edge subsets, respectively, must be the same, $|V_1^{(1,2)}| = |V_2^{(1,2)}|$ and $|E_1^{(1,2)}| = |E_2^{(1,2)}| = \max$. The subset $E_1^{(1,2)}$ is mapped onto the subset $E_2^{(1,2)}$ by a $1-1$ function (induced by the isomorphism) $f_{12}, f_{12}: E_1^{(1,2)} \to E_2^{(1,2)}$ or $f_{12}(E_1^{(1,2)}) = E_2^{(1,2)}$. A distance between graphs G_1 and G_2 is determined by

*) Part VIII in the series Mathematical Model of Organic Chemistry.

$$(1) \qquad d(G_1, G_2) = |E_1 - E_1^{(1,2)}| + |E_2 - E_2^{(1,2)}| + ||V_1| - |V_2|| = \\ = |E_1| + |E_2| - 2|E_1^{(1,2)}| + ||V_1| - |V_2||\,.$$

In particular, for graphs with the same number of edges, this distance corresponds to the number of edges that cannot be matched in the construction of MCS of two graphs G_1 and G_4.

Theorem. *The distance $d(G_1, G_2)$ is a metric, the following three properties are satisfied:*

(i) *Positive semidefiniteness*

$$d(G_1, G_2) \geqq 0 \quad (= 0 \textit{ only for } G_1 \sim G_2)\,.$$

(ii) *Symmetry*

$$d(G_1 G_2) = d(G_2, G_1)\,.$$

(iii) *Triangular inequality*

$$d(G_1, G_2) + d(G_2, G_3) \geqq d(G_1, G_3)\,.$$

Proof. The first two properties of the distance (1) are obvious. Using (1) and its analogues for the two pairs of graphs G_2, G_3 and G_1, G_3 we get

$$(2) \qquad d(G_1, G_2) + d(G_2, G_3) - d(G_1, G_3) = \\ = 2(|E_2| + |E_1^{(1,3)}| - |E_2^{(1,2)}| - |E_2^{(2,3)}|) + \\ + ||V_1| - |V_2|| + ||V_2| - |V_2|| - ||V_1| - |V_3|| = \\ = 2(|E_2 - E_2^{(1,2)} \cup E_2^{(2,3)}| + |E_1^{(1,3)}| - |E_2^{(1,2)} \cap E_2^{(2,3)}|) + \\ + ||V_1| - |V_2|| + ||V_2| - |V_3|| - ||V_1| - |V_3||\,,$$

where we have used $|A| - |B_1| - |B_2| = |A - B_1 \cup B_2| - |B_1 \cap B_2|$ for $B_1, B_2 \subseteq \subseteq A$. The term $||V_1| - |V_2|| + ||V_2| - |V_3|| - ||V_1| - |V_3||$ is automatically positive semidefinite, which follows immediately from the well-known inequality $|a - b| + |b - c| \geqq |a - c|$. Hence, in order to prove the triangular inequality it is sufficient to verify only the positive semidefiniteness of the term $|E_2 - E_2^{(1,2)} \cup \cup E_2^{(2,3)}| + |E_1^{(1,3)}| - |E_2^{(1,2)} \cap E_2^{(2,3)}|$. If the subsets $E_2^{(1,2)}$ and $E_2^{(2,3)}$ are disjoint, then the triangular inequality is fulfilled. Let us assume that the intersection $E_2^{(1,2)} \cap \cap E_2^{(2,3)}$ is a non-empty subset of E_2. Using the mappings f_{12}^{-1} and f_{23} we can form two subsets $f_{12}^{-1}(E_2^{(1,2)} \cap E_2^{(2,3)}) = \tilde{E}_2 \subseteq E_2$ and $f_{23}(E_2^{(1,2)} \cap E_2^{(2,3)}) = \tilde{E}_3 \subseteq E_3$ with the same cardinality, $|\tilde{E}_2| = |\tilde{E}_3|$. The subset $\tilde{E}_1$ can be mapped onto the subset $\tilde{E}_3$ by a $1-1$ function $\tilde{f}_{13} = f_{12} \circ f_{23}$, $\tilde{E}_3 = \tilde{f}_{13}(\tilde{E}_1) = f_{23}[f_{12}(\tilde{E}_1)]$. This means that we have formed a common subgraph of the graphs G_1 and G_3, composed of the edge subsets $\tilde{E}_1$ and $\tilde{E}_3$, respectively. Since the MCS of the graphs G_1 and G_3 contains $|E_1^{(1,3)}| = |E_3^{(1,3)}| =$ max edges, we have

$$(3) \qquad |E_1^{(1,3)}| \geqq |\tilde{E}_1| = |\tilde{E}_3| = |E_2^{(1,2)} \cap E_2^{(2,3)}|\,.$$

Inserting this relation into (2) we arrive at the triangular inequality, which was to be demonstrated.

4. Let us consider a pair of graphs G_1 and G_2 composed of the same number of vertices, $|V_1| = |V_2| = n$. Following McGregor [3], the construction of MCS is carried out in such a way that (using a back-track searching algorithm) we look for a $1-1$ mapping of V_1 onto V_2 such that the induced common subgraph is composed of the largest possible number of edges. Let $\mathbf{A}_1$ and $\mathbf{A}_2$ be the adjacency matrices of G_1 and G_2, respectively. The above mentioned $1-1$ mapping may be simply realized by a permutation P of n objects $(1, 2, \ldots, n)$. This directly implies that the second alternative definition of distance for a pair of graphs (with the same number of vertices) is

$$d(G_1, G_2) = \min_{\mathbf{P}} |\mathbf{A}_1 - \mathbf{P}^T \mathbf{A}_2 \mathbf{P}| , \tag{4}$$

where $|\mathbf{A}| = \sum_{i \leq j} |a_{ij}|$ is the Hamming (linear) norm of a symmetric matrix $\mathbf{A}$. The relation (4) is nothing elese than our determination of the so-called chemical distance [1] between two graphs representing molecular structure formulas.

Acknowledgment. We wish to express our thanks to Dr. Michal Šabo for his critical reading the manuscript and pointing-out some inconsistencies.

References

[1] *V. Kvasnička, M. Kratochvíl, J. Koča:* Reaction Graphs. Collect. Czech. Chem. Commun. *48* (1983) 2284.

[2] *B. Zelinka:* On a Certain Distance Between Isomorphism Classes of Graphs, Čas. pěst. mat. *100* (1975) 371.

[3] *J. J. McGregor:* Backtrack Search Algorithm and the Maximal Common Subgraph Problem. Software Pract. Exper. *12* (1982) 23.

Authors' addresses: V. Kvasnička, V. Baláž, 812 37 Bratislava, Jánská 1 (Katedra matematiky, Chemickotechnologická fakulta SVŠT), M. Sekanina, J. Koča, 602 37 Brno, Janáčkovo nám. 5 (Katedra matematiky, Přírodovědecká fakulta UJEP).