УДК 519.17

МЕТРИКА ДЛЯ СРАВНЕНИЯ ГРАФОВ С УПОРЯДОЧЕННЫМИ ВЕРШИНАМИ НА ОСНОВЕ МАКСИМАЛЬНОГО ОБЩЕГО ПОДГРАФА

Н. Д. Москин

Петрозаводский государственный университет, г. Петрозаводск, Россия

Работа посвящена методам сравнения и классификации графов. Данное направление известно под названием «graph matching». Приводится обзор метрик для сравнения графов, основанных на максимальном общем подграфе. Предложена модификация расстояния на основе максимального общего подграфа, которое учитывает упорядоченность вершин. Показано, что эта функция удовлетворяет всем свойствам метрики (неотрицательность, тождественность, симметричность, неравенство треугольника).

Ключевые слова: *граф, сравнение, метрика, максимальный общий подграф, graph matching.*

DOI 10.17223/20710410/52/7

METRIC FOR COMPARING GRAPHS WITH ORDERED VERTICES BASED ON THE MAXIMUM COMMON SUBGRAPH

N. D. Moskin

Petrozavodsk State University, Petrozavodsk, Russia

E-mail: moskin@petrsu.ru

The paper is devoted to the methods of comparison and classification of graphs. This direction is known as graph matching. An overview of metrics for comparing graphs based on a maximum common subgraph is given. A graph $\mathrm{mcs}(G, F)$ is a maximal common subgraph of graphs G and F if it is isomorphic to $G' \subseteq G$ and $F' \subseteq F$ and contains the maximum number of vertices. In some tasks (for example, comparing texts), it is important to take into account one more factor: vertex numbering. A modification of the distance based on the maximum common subgraph is proposed, taking into account this factor (each vertex has its own unique number). We determine a function of graphs G and F as follows: $d(G, F) = 1 - \min\limits_{i=1,\ldots,k}(|\mathrm{mcs}(g_{\min(i,m)}, f_i)|/i)$. Here $|G|$ denotes the number of vertices in G, $|G| = m$, $|F| = k$, $m \leqslant k$; and g_i is the subgraph of G containing vertices with numbers from 1 to i and all edges of G incident to these vertices (the graphs f_i are defined similarly). It is shown that this function satisfies all the properties of the metric (nonnegativity, identity, symmetry, triangle inequality). This metric can be used to solve various problems of image recognition (for example, to establish the authorship of texts).

Keywords: *graph, comparison, metric, maximum common subgraph, graph matching.*

Введение

В настоящее время графы используются в различных областях науки и могут быть построены по разным принципам. Одной из задач, которая возникает при построении подобных моделей, является задача сравнения и классификации [1]. Методы, известные в рамках направления graph matching, нашли своё применение в обработке изображений [2], молекулярной биологии [3], дактилоскопии [4], распознавании почерка [5], исследовании социальных сетей [6], при анализе документов [7] и т. д. На множестве графов задаётся расстояние, которое позволяет оценить, насколько те или иные структуры похожи друг на друга. Как правило, эта функция выражает степень неточностей, которые возникают при нахождении изоморфизма графов или подграфов. При этом в некоторых задачах (например, при сравнении текстов) важно учитывать ещё один фактор: нумерацию вершин.

Одной из таких задач является атрибуция текстов. Например, данная проблема возникает при анализе коллекции текстов из дореволюционных журналов «Время» (1861–1863), «Эпоха» (1864–1865) и еженедельника «Гражданин» (1873–1874). Известно, что Ф. М. Достоевский (вместе со своим братом М. М. Достоевским) редактировал и возглавлял эти журналы, поэтому уже давно ведутся исследования на предмет принадлежности его перу данных произведений. Большое количество этих статей опубликовано анонимно, т. е. либо без подписи, либо под псевдонимами. Впрочем, это относится и к статьям, которые исследователи давно приписывали Достоевскому, более или менее основываясь на документальных данных.

Для решения задачи атрибуции текстов могут быть использованы «графы сильных связей» [8]. Множество вершин подобного графа — это множество грамматических форм, которые встречаются в текстах, а рёбра отражают «сильные связи» между вершинами. Две вершины v_i и v_j связаны ребром, если частота встречаемости пары грамматических классов больше или равна заданному в исследовании пороговому значению α. Далее грамматические формы и соответствующие им вершины могут быть отсортированы в порядке убывания по степени их встречаемости в определённых текстах (или группе текстов). Кроме того, если вершина не имеет инцидентных рёбер в результате отбрасывания «слабых» связей, она рассматривается как изолированная вершина. С точки зрения филологии при распознавании авторского стиля писателя важно не только наличие в корпусе тех или иных грамматических конструкций, но и порядок их встречаемости в текстах.

В п. 1 данной работы описаны несколько известных метрик, основанных на максимальном общем подграфе. В п. 2 предложена модификация расстояния на основе максимального общего подграфа, которая учитывает упорядоченность вершин (т. е. каждой вершине ставится в соответствие её уникальный порядковый номер). Показано, что эта функция удовлетворяет всем свойствам метрики (неотрицательность, тождественность, симметричность, неравенство треугольника).

1. Метрики на множестве графов, основанные на максимальном общем подграфе

Одним из способов сравнения графов является расстояние на основе максимального общего подграфа. Максимальным общим подграфом графов G_1 и G_2 будем называть граф $\mathrm{mcs}(G_1, G_2)$, который изоморфен $G_1' \subseteq G_1$, $G_2' \subseteq G_2$ и содержит максимальное число вершин.

Обозначим через $|G|$ число вершин в графе G. Расстояние между непустыми графами G_1 и G_2 можно вычислить следующим образом [9]:

$$d_1(G_1, G_2) = 1 - \frac{|\mathrm{mcs}(G_1, G_2)|}{\max(|G_1|, |G_2|)}. \quad (1)$$

Расстояние $d_1(G_1, G_2)$ принимает значения от 0 до 1 включительно. Если графы изоморфны, то $d_1(G_1, G_2) = 0$. Заметим, что, исходя из определения, максимальный общий подграф $\mathrm{mcs}(G_1, G_2)$ не обязательно уникален.

Вторая функция на основе максимального общего подграфа предложена в [10]:

$$d_2(G_1, G_2) = 1 - \frac{|\mathrm{mcs}(G_1, G_2)|}{|G_1| + |G_2| - |\mathrm{mcs}(G_1, G_2)|}.$$

Значения $d_2(G_1, G_2)$, как и $d_1(G_1, G_2)$, находятся в пределах от 0 до 1. Ещё одно похожее расстояние предложено в [11], но оно не нормализовано для отрезка [0, 1]:

$$d_3(G_1, G_2) = |G_1| + |G_2| - 2|\mathrm{mcs}(G_1, G_2)|.$$

При задании функции на множестве графов желательно, чтобы $d(G_i, G_j)$ удовлетворяла следующим свойствам метрики (для произвольных i, j, k):

1) $d(G_i, G_j) \geqslant 0$ (неотрицательность);
2) $d(G_i, G_j) = 0 \Leftrightarrow G_i = G_j$ (тождественность);
3) $d(G_i, G_j) = d(G_j, G_i)$ (симметричность);
4) $d(G_i, G_j) \leqslant d(G_i, G_k) + d(G_k, G_j)$ (неравенство треугольника).

В [9] показано, что функция (1) удовлетворяет всем свойствам метрики.

2. Метрика на основе максимального общего подграфа для сравнения графов с упорядоченными вершинами

При сравнении графов с упорядоченными вершинами важно учитывать нумерацию вершин. Представим граф с упорядоченными вершинами в виде цепочки порождающих его подграфов $g_1, g_2, g_3, \ldots, g_m$, как показано на рис. 1 (m — число вершин графа G). Здесь граф g_i является подграфом G, который содержит вершины с номерами от 1 до i включительно и все рёбра G, инцидентные этим вершинам. Подграф g_m совпадает с графом G.

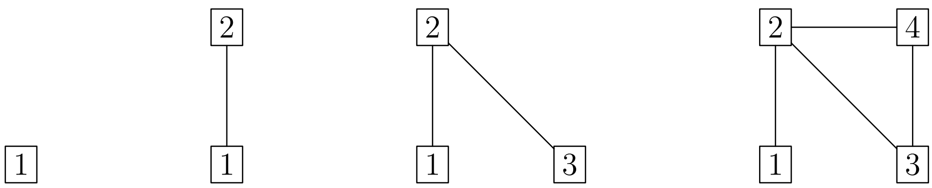

Рис. 1. Цепочка порождающих графов для графа G

Цепочку порождающих графов можно рассматривать как частный случай динамических графов [12]. Динамический граф представляет собой последовательность конечных невзвешенных (не всегда связных) графов $g_1, g_2, g_3, \ldots, g_l, \ldots$, в которой переход к последующему графу g_{l+1} осуществляется применением операции $\varphi(g_l) = g_{l+1}$.

Операция, осуществляющая переход, может быть как простой (удаление/добавление ребра, удаление/добавление вершины), так и сложной (это операция, которую можно описать чередованием простых операций). Последовательность графов, составляющих динамический граф, называют траекторией динамического графа.

Определим функцию на графах G и F следующим образом (пусть для определённости $|G| = m$, $|F| = k$, $m \leqslant k$):

$$d(G, F) = 1 - \min_{i=1,\ldots,k}(|\mathrm{mcs}(g_{\min(i,m)}, f_i)|/i). \tag{2}$$

Теорема 1. Величина d удовлетворяет всем свойствам метрики.

Доказательство.

а) Докажем первое свойство. Так как количество вершин в графах $g_{\min(i,m)}$ и f_i не превосходит i для $i = 1, \ldots, k$, то $|\mathrm{mcs}(g_{\min(i,m)}, f_i)| \leqslant i$, т. е.

$$|\mathrm{mcs}(g_{\min(i,m)}, f_i)|/i \leqslant 1.$$

Следовательно, справедиво неравенство

$$\min_{i=1,\ldots,k}(|\mathrm{mcs}(g_{\min(i,m)}, f_i)|/i) \leqslant 1.$$

Тогда получаем

$$d(G, F) = 1 - \min_{i=1,\ldots,k}(|\mathrm{mcs}(g_{\min(i,m)}, f_i)|/i) \geqslant 0.$$

б) Докажем второе свойство.

$\Rightarrow$ Пусть $d(G, F) = 0$. Тогда по формуле (2)

$$\min_{i=1,\ldots,k}(|\mathrm{mcs}(g_{\min(i,m)}, f_i)|/i) = 1,$$

поэтому для $i = 1, \ldots, k$

$$|\mathrm{mcs}(g_{\min(i,m)}, f_i)|/i \geqslant 1,$$

т. е. $|\mathrm{mcs}(g_{\min(i,m)}, f_i)| \geqslant i$.

С другой стороны, $|\mathrm{mcs}(g_{\min(i,m)}, f_i)| \leqslant i$. Поэтому $|\mathrm{mcs}(g_{\min(i,m)}, f_i)| = i$. При $i \geqslant m$ выполняется равенство $|\mathrm{mcs}(g_m, f_i)| = i$, что возможно только если $i = m$. Следовательно, $|\mathrm{mcs}(g_m, f_m)| = m$, т. е. графы G и F изоморфны.

$\Leftarrow$ Пусть $G = F$. Тогда $m = k$ и $\min(i, m) = i$ для $i = 1, \ldots, k$; $g_i = f_i$ для $i = 1, \ldots, k$, следовательно, максимальный общий подграф имеет i вершин, т. е. $|\mathrm{mcs}(g_i, f_i)| = i$ для $i = 1, \ldots, k$. Отсюда

$$\min_{i=1,\ldots,k}(|\mathrm{mcs}(g_{\min(i,m)}, f_i)|/i) = 1,$$

поэтому $d(G, F) = 0$.

в) Третье свойство справедливо, поскольку $|\mathrm{mcs}(g_{\min(i,m)}, f_i)| = |\mathrm{mcs}(f_i, g_{\min(i,m)})|$ для $i = 1, \ldots, k$.

г) Докажем четвёртое свойство. Пусть H — произвольный граф с t вершинами. Необходимо показать, что $d(G, F) \leqslant d(G, H) + d(H, F)$. Рассмотрим три случая (по условию $m \leqslant k$, поэтому возможны три варианта расположения t относительно m и k: $t < m \leqslant k$, $m \leqslant t \leqslant k$ и $m \leqslant k < t$).

Случай 1: $t < m \leqslant k$. Согласно (2), надо доказать, что

$$1 - \min_{i=1,\ldots,k}(|\mathrm{mcs}(g_{\min(i,m)}, f_i)|/i) \leqslant 1 - \min_{i=1,\ldots,m}(|\mathrm{mcs}(g_i, h_{\min(i,t)})|/i) + \\ +1 - \min_{i=1,\ldots,k}(|\mathrm{mcs}(h_{\min(i,t)}, f_i)|/i).$$

Так как графы g_i, f_i, h_i содержат i вершин, т. е. $|g_i| = |f_i| = |h_i| = i$, выполним следующую замену:

$$1 - \min_{i=1,\ldots,k}\left(\frac{|\mathrm{mcs}(g_{\min(i,m)}, f_i)|}{\max(|g_{\min(i,m)}|, |f_i|)}\right) \leqslant 1 - \min_{i=1,\ldots,m}\left(\frac{|\mathrm{mcs}(g_i, h_{\min(i,t)})|}{\max(|g_i|, |h_{\min(i,t)}|)}\right) + \\ +1 - \min_{i=1,\ldots,k}\left(\frac{|\mathrm{mcs}(h_{\min(i,t)}, f_i)|}{\max(|h_{\min(i,t)}|, |f_i|)}\right).$$

Перейдя от поиска минимума к максимуму, получим

$$\max_{i=1,\ldots,k}\left(1 - \frac{|\mathrm{mcs}(g_{\min(i,m)}, f_i)|}{\max(|g_{\min(i,m)}|, |f_i|)}\right) \leqslant \max_{i=1,\ldots,m}\left(1 - \frac{|\mathrm{mcs}(g_i, h_{\min(i,t)})|}{\max(|g_i|, |h_{\min(i,t)}|)}\right) + \\ + \max_{i=1,\ldots,k}\left(1 - \frac{|\mathrm{mcs}(h_{\min(i,t)}, f_i)|}{\max(|h_{\min(i,t)}|, |f_i|)}\right).$$

Согласно формуле (1), перепишем неравенство:

$$\max_{i=1,\ldots,k} d_1(g_{\min(i,m)}, f_i) \leqslant \max_{i=1,\ldots,m} d_1(g_i, h_{\min(i,t)}) + \max_{i=1,\ldots,k} d_1(h_{\min(i,t)}, f_i).$$

1.1. Предположим, что $\max\limits_{i=1,\ldots,k} d_1(g_{\min(i,m)}, f_i)$ достигается при некотором индексе i^*, который находится на интервале $1 \leqslant i^* \leqslant t$, тогда

$$\max_{i=1,\ldots,k} d_1(g_{\min(i,m)}, f_i) = d_1(g_{i^*}, f_{i^*}) \leqslant d_1(g_{i^*}, h_{i^*}) + d_1(h_{i^*}, f_{i^*}) \leqslant \\ \leqslant \max_{i=1,\ldots,t} d_1(g_i, h_i) + \max_{i=1,\ldots,t} d_1(h_i, f_i) \leqslant \max_{i=1,\ldots,m} d_1(g_i, h_{\min(i,t)}) + \max_{i=1,\ldots,k} d_1(h_{\min(i,t)}, f_i).$$

Неравенство выполняется, так как величина d_1 удовлетворяет всем свойствам метрики, а граф h_{i^*} существует при $i^* \leqslant t$.

1.2. Предположим, что $\max\limits_{i=1,\ldots,k} d_1(g_{\min(i,m)}, f_i)$ достигается при некотором индексе i^*, который находится на интервале $t < i^* \leqslant m$, тогда

$$\max_{i=1,\ldots,k} d_1(g_{\min(i,m)}, f_i) = d_1(g_{i^*}, f_{i^*}) \leqslant d_1(g_{i^*}, h_t) + d_1(h_t, f_{i^*}) \leqslant \\ \leqslant \max_{i=t+1,\ldots,m} d_1(g_i, h_t) + \max_{i=t+1,\ldots,m} d_1(h_t, f_i) \leqslant \max_{i=1,\ldots,m} d_1(g_i, h_{\min(i,t)}) + \max_{i=1,\ldots,k} d_1(h_{\min(i,t)}, f_i).$$

Неравенство выполняется, так как величина d_1 удовлетворяет всем свойствам метрики.

1.3. Предположим, что $\max\limits_{i=1,\ldots,k} d_1(g_{\min(i,m)}, f_i)$ достигается при некотором индексе i^*, который находится на интервале $m < i^* \leqslant k$, тогда

$$\max_{i=1,\ldots,k} d_1(g_{\min(i,m)}, f_i) = d_1(g_m, f_{i^*}) \leqslant d_1(g_m, h_t) + d_1(h_t, f_{i^*}) \leqslant d_1(g_m, h_t) + \\ + \max_{i=m+1,\ldots,k} d_1(h_t, f_i) \leqslant \max_{i=1,\ldots,m} d_1(g_i, h_{\min(i,t)}) + \max_{i=1,\ldots,k} d_1(h_{\min(i,t)}, f_i).$$

Неравенство выполняется, так как величина d_1 удовлетворяет всем свойствам метрики.

С л у ч а й 2: $m \leqslant t \leqslant k$. Согласно (2), надо доказать, что

$$1 - \min_{i=1,\dots,k}(|\mathrm{mcs}(g_{\min(i,m)}, f_i)|/i) \leqslant 1 - \min_{i=1,\dots,t}(|\mathrm{mcs}(g_{\min(i,m)}, h_i)|/i)+$$
$$+1 - \min_{i=1,\dots,k}(|\mathrm{mcs}(h_{\min(i,t)}, f_i)|/i).$$

Так как $|g_i| = |f_i| = |h_i| = i$, выполним следующую замену:

$$1 - \min_{i=1,\dots,k}\left(\frac{|\mathrm{mcs}(g_{\min(i,m)}, f_i)|}{\max(|g_{\min(i,m)}|, |f_i|)}\right) \leqslant 1 - \min_{i=1,\dots,t}\left(\frac{|\mathrm{mcs}(g_{\min(i,m)}, h_i)|}{\max(|g_{\min(i,m)}|, |h_i|)}\right)+$$
$$+1 - \min_{i=1,\dots,k}\left(\frac{|\mathrm{mcs}(h_{\min(i,t)}, f_i)|}{\max(|h_{\min(i,t)}|, |f_i|)}\right).$$

Перейдя от поиска минимума к максимуму, получим

$$\max_{i=1,\dots,k}\left(1 - \frac{|\mathrm{mcs}(g_{\min(i,m)}, f_i)|}{\max(|g_{\min(i,m)}|, |f_i|)}\right) \leqslant \max_{i=1,\dots,t}\left(1 - \frac{|\mathrm{mcs}(g_{\min(i,m)}, h_i)|}{\max(|g_{\min(i,m)}|, |h_i|)}\right)+$$
$$+\max_{i=1,\dots,k}\left(1 - \frac{|\mathrm{mcs}(h_{\min(i,t)}, f_i)|}{\max(|h_{\min(i,t)}|, |f_i|)}\right).$$

Согласно формуле (1), перепишем неравенство так:

$$\max_{i=1,\dots,k} d_1(g_{\min(i,m)}, f_i) \leqslant \max_{i=1,\dots,t} d_1(g_{\min(i,m)}, h_i) + \max_{i=1,\dots,k} d_1(h_{\min(i,t)}, f_i).$$

2.1. Предположим, что $\max\limits_{i=1,\dots,k} d_1(g_{\min(i,m)}, f_i)$ достигается при некотором индексе i^*, который находится на интервале $1 \leqslant i^* \leqslant m$, тогда

$$\max_{i=1,\dots,k} d_1(g_{\min(i,m)}, f_i) = d_1(g_{i^*}, f_{i^*}) \leqslant d_1(g_{i^*}, h_{i^*}) + d_1(h_{i^*}, f_{i^*}) \leqslant \max_{i=1,\dots,m} d_1(g_i, h_i)+$$
$$+\max_{i=1,\dots,m} d_1(h_i, f_i) \leqslant \max_{i=1,\dots,t} d_1(g_{\min(i,m)}, h_i) + \max_{i=1,\dots,k} d_1(h_{\min(i,t)}, f_i).$$

Неравенство выполняется, так как величина d_1 удовлетворяет всем свойствам метрики, а граф h_{i^*} существует при $i^* \leqslant m \leqslant t$.

2.2. Предположим, что $\max\limits_{i=1,\dots,k} d_1(g_{\min(i,m)}, f_i)$ достигается при некотором индексе i^*, который находится на интервале $m < i^* \leqslant t$, тогда

$$\max_{i=1,\dots,k} d_1(g_{\min(i,m)}, f_i) = d_1(g_m, f_{i^*}) \leqslant d_1(g_m, h_{i^*}) + d_1(h_{i^*}, f_{i^*}) \leqslant$$
$$\leqslant \max_{i=m+1,\dots,t} d_1(g_m, h_i) + \max_{i=m+1,\dots,t} d_1(h_i, f_i) \leqslant \max_{i=1,\dots,t} d_1(g_{\min(i,m)}, h_i) + \max_{i=1,\dots,k} d_1(h_{\min(i,t)}, f_i).$$

Неравенство выполняется, так как величина d_1 удовлетворяет всем свойствам метрики, а граф h_{i^*} существует при $i^* \leqslant t$.

2.3. Предположим, что $\max\limits_{i=1,\dots,k} d_1(g_{\min(i,m)}, f_i)$ достигается при некотором индексе i^*, который находится на интервале $t < i^* \leqslant k$, тогда

$$\max_{i=1,\dots,k} d_1(g_{\min(i,m)}, f_i) = d_1(g_m, f_{i^*}) \leqslant d_1(g_m, h_t) + d_1(h_t, f_{i^*}) \leqslant d_1(g_m, h_t)+$$
$$+\max_{i=t+1,\dots,k} d_1(h_t, f_i) \leqslant \max_{i=1,\dots,t} d_1(g_{\min(i,m)}, h_i) + \max_{i=1,\dots,k} d_1(h_{\min(i,t)}, f_i).$$

Неравенство выполняется, так как величина d_1 удовлетворяет всем свойствам метрики.

С л у ч а й 3: $m \leqslant k < t$. Согласно (2), надо доказать, что

$$1 - \min_{i=1,\ldots,k}(|\mathrm{mcs}(g_{\min(i,m)}, f_i)|/i) \leqslant 1 - \min_{i=1,\ldots,t}(|\mathrm{mcs}(g_{\min(i,m)}, h_i)|/i) +$$
$$+1 - \min_{i=1,\ldots,t}(|\mathrm{mcs}(h_i, f_{\min(i,k)})|i).$$

Так как $|g_i| = |f_i| = |h_i| = i$, выполним следующую замену:

$$1 - \min_{i=1,\ldots,k}\left(\frac{|\mathrm{mcs}(g_{\min(i,m)}, f_i)|}{\max(|g_{\min(i,m)}|, |f_i|)}\right) \leqslant 1 - \min_{i=1,\ldots,t}\left(\frac{|\mathrm{mcs}(g_{\min(i,m)}, h_i)|}{\max(|g_{\min(i,m)}|, |h_i|)}\right) +$$
$$+1 - \min_{i=1,\ldots,t}\left(\frac{|\mathrm{mcs}(h_i, f_{\min(i,k)})|}{\max(|h_i|, |f_{\min(i,k)}|)}\right).$$

Перейдя от поиска минимума к максимуму, получим

$$\max_{i=1,\ldots,k}\left(1 - \frac{|\mathrm{mcs}(g_{\min(i,m)}, f_i)|}{\max(|g_{\min(i,m)}|, |f_i|)}\right) \leqslant \max_{i=1,\ldots,t}\left(1 - \frac{|\mathrm{mcs}(g_{\min(i,m)}, h_i)|}{\max(|g_{\min(i,m)}|, |h_i|)}\right) +$$
$$+ \max_{i=1,\ldots,t}\left(1 - \frac{|\mathrm{mcs}(h_i, f_{\min(i,k)})|}{\max(|h_i|, |f_{\min(i,k)}|)}\right).$$

Согласно формуле (1), перепишем неравенство так:

$$\max_{i=1,\ldots,k} d_1(g_{\min(i,m)}, f_i) \leqslant \max_{i=1,\ldots,t} d_1(g_{\min(i,m)}, h_i) + \max_{i=1,\ldots,t} d_1(h_i, f_{\min(i,k)}).$$

3.1. Предположим, что $\max\limits_{i=1,\ldots,k} d_1(g_{\min(i,m)}, f_i)$ достигается при некотором индексе i^*, который находится на интервале $1 \leqslant i^* \leqslant m$, тогда

$$\max_{i=1,\ldots,k} d_1(g_{\min(i,m)}, f_i) = d_1(g_{i^*}, f_{i^*}) \leqslant d_1(g_{i^*}, h_{i^*}) + d_1(h_{i^*}, f_{i^*}) \leqslant \max_{i=1,\ldots,m} d_1(g_i, h_i) +$$
$$+ \max_{i=1,\ldots,m} d_1(h_i, f_i) \leqslant \max_{i=1,\ldots,t} d_1(g_{\min(i,m)}, h_i) + \max_{i=1,\ldots,t} d_1(h_i, f_{\min(i,k)}).$$

Неравенство выполняется, так как величина d_1 удовлетворяет всем свойствам метрики, а граф h_{i^*} существует при $i^* \leqslant m < t$.

3.2. Предположим, что $\max\limits_{i=1,\ldots,k} d_1(g_{\min(i,m)}, f_i)$ достигается при некотором индексе i^*, который находится на интервале $m < i^* \leqslant k$, тогда

$$\max_{i=1,\ldots,k} d_1(g_{\min(i,m)}, f_i) = d_1(g_m, f_{i^*}) \leqslant d_1(g_m, h_{i^*}) + d_1(h_{i^*}, f_{i^*}) \leqslant$$
$$\leqslant \max_{i=m+1,\ldots,k} d_1(g_m, h_i) + \max_{i=m+1,\ldots,k} d_1(h_i, f_i) \leqslant \max_{i=1,\ldots,t} d_1(g_{\min(i,m)}, h_i) + \max_{i=1,\ldots,t} d_1(h_i, f_{\min(i,k)}).$$

Неравенство выполняется, так как величина d_1 удовлетворяет всем свойствам метрики, а граф h_{i^*} существует при $i^* \leqslant k < t$.

Таким образом, величина d удовлетворяет всем свойствам метрики. ■

На рис. 2 изображены три графа G_1, G_2, G_3 с разной нумерацией вершин. Расстояния между ними следующие: $d(G_1, G_2) = 1/3$, $d(G_1, G_3) = d(G_2, G_3) = 1/2$. Как видно на рисунке, отличия в первой паре графов возникают на уровне третьего порождающего графа, тогда как в двух остальных парах различия видны уже на втором уровне.

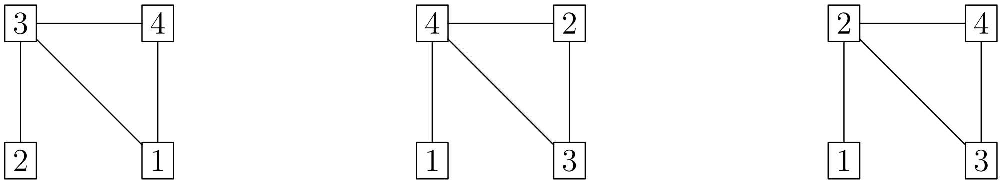

Рис. 2. Графы G_1, G_2, G_3 с различной нумерацией вершин

Заключение

В работе рассмотрена задача сравнения и классификации графов. Предложена функция для сравнения графов, которая основана на максимальном общем подграфе и учитывает упорядоченность вершин. Доказано, что эта величина удовлетворяет всем свойствам метрики. Классические алгоритмы для поиска максимального общего подграфа $\mathrm{mcs}(G_1, G_2)$ для двух графов основаны на алгоритме поиска с возвратом, предложенном МакГрегором, или на поиске максимальной клики, предложенном Леви [13]. Первый ищет все возможные общие подграфы и выбирает среди них максимальный, второй строит специальный граф соответствий и в нём ищет максимальную клику, которая определяет максимальный общий подграф. Заметим, что обе задачи являются NP-полными. Для проведения вычислительных экспериментов был реализован алгоритм первого типа, осуществляющий полный перебор возможных подграфов. При подсчёте расстояний между графами разной размерности, построенными случайным образом, свойства метрики выполняются.

Отметим, что данная метрика может быть использована в различных задачах теории распознавания образов (например, при сравнении текстов и изображений).

ЛИТЕРАТУРА

1. *Conte D., Foggia P., Sansone C., and Vento M.* Thirty years of graph matching in pattern recognition // Int. J. Pattern Recognit. Artif. Intell. 2004. V. 18. No. 3. P. 265–298.
2. *Sharma H., Pawar A., Chourasia C., and Khatri S.* Implementation of face recognition system based on elastic bunch graph matching // Intern. J. Engineering Sciences & Research Technology (IJESRT). 2016. V. 5. No. 3. P. 888–895.
3. *Schirmer S., Ponty Y., and Giegerich R.* Introduction to RNA secondary structure comparison // RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods. Methods in Molecular Biology. Totowa, NJ: Humana Press, 2014. V. 1097. P. 247–273.
4. *Pawar V. and Zaveri M.* K-Means graph database clustering and matching for fingerprint recognition // Intelligent Information Management. 2015. V. 7. No. 4. P. 242–251.
5. *Fischer A., Suen C., Frinken V., et al.* A fast matching algorithm for graph-based handwriting recognition // LNCS. 2013. V. 7877. P. 194–203.
6. *Ogaard K., Roy H., Kase S., et al.* Discovering patterns in social networks with graph matching algorithms // LNCS. 2013. V. 7812. P. 341–349.
7. *Stauffer M., Fischer A., and Riesen K.* Speeding-up graph-based keyword spotting in historical handwritten documents // LNCS. 2017. V. 10310. P. 83–93.
8. *Рогов А. А., Седов А. В., Сидоров Ю. В., Суровцова Т. Г.* Математические методы атрибуции текстов. Петрозаводск: Изд-во ПетрГУ, 2014. 96 с.
9. *Bunke H. and Shearer K.* A graph distance metric based on the maximal common subgraph // Pattern Recognit. Lett. 1998. V. 19. No. 3–4. P. 255–259.

10. *Wallis W., Shoubridge P., Kraetz M., and Ray D.* Graph distances using graph union // Pattern Recognit. Lett. 2001. V. 22. P. 701–704.
11. *Bunke H.* On a relation between graph edit distance and maximum common subgraph // Pattern Recognit. Lett. 1997. V. 18. P. 689–694.
12. *Кочкаров А. А., Сенникова Л. И.* Метрические характеристики динамических графов и их применение // Новые информационные технологии в автоматизированных системах. М.: Московский институт электроники и математики НИУ ВШЭ, 2015. № 18. С. 236–241.
13. *Bunke H., Foggia P., Guidobaldi C., et al.* A comparison of algorithms for maximum common subgraph on randomly connected graphs // LNCS. 2002. V. 2396. P. 123–132.

REFERENCES

1. *Conte D., Foggia P., Sansone C., and Vento M.* Thirty years of graph matching in pattern recognition. Int. J. Pattern Recognit. Artif. Intell., 2004, vol. 18, no. 3, pp. 265–298.
2. *Sharma H., Pawar A., Chourasia C., and Khatri S.* Implementation of face recognition system based on elastic bunch graph matching. Intern. J. Engineering Sciences & Research Technology (IJESRT), 2016, vol. 5, no. 3, pp. 888–895.
3. *Schirmer S., Ponty Y., and Giegerich R.* Introduction to RNA secondary structure comparison. RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods. Methods in molecular biology. Totowa, NJ, Humana Press, 2014, vol. 1097, pp. 247–273.
4. *Pawar V. and Zaveri M.* K-Means graph database clustering and matching for fingerprint recognition. Intelligent Information Management, 2015, vol. 7, no. 4, pp. 242–251.
5. *Fischer A., Suen C., Frinken V., et al.* A fast matching algorithm for graph-based handwriting recognition. LNCS, 2013, vol. 7877, pp. 194–203.
6. *Ogaard K., Roy H., Kase S., et al.* Discovering patterns in social networks with graph matching algorithms. LNCS, 2013, vol. 7812, pp. 341–349.
7. *Stauffer M., Fischer A., and Riesen K.* Speeding-up graph-based keyword spotting in historical handwritten documents. LNCS, 2017, vol. 10310, pp. 83–93.
8. *Rogov A. A., Sedov A. V., Sidorov Y. V., and Surovceva T. G.* Matematicheskie metody atribucii tekstov [Mathematical Methods for text Attribution]. Petrozavodsk, PetrSU Publ., 2014. 96 p. (in Russian)
9. *Bunke H. and Shearer K.* A graph distance metric based on the maximal common subgraph. Pattern Recognit. Lett., 1998, vol. 19, no. 3–4, pp. 255–259.
10. *Wallis W., Shoubridge P., Kraetz M., and Ray D.* Graph distances using graph union. Pattern Recognit. Lett., 2001, vol. 22, pp. 701–704.
11. *Bunke H.* On a relation between graph edit distance and maximum common subgraph. Pattern Recognit. Lett., 1997. vol. 18. pp. 689–694.
12. *Kochkarov A. A. and Sennikova L. I.* Metricheskiye kharakteristiki dinamicheskikh grafov i ikh primeneniye [Metric characteristics of dynamic graphs and their application]. Novyye Informatsionnyye Tekhnologii v Avtomatizirovannykh Sistemakh, 2015, no. 18, pp. 236–241. (in Russian)
13. *Bunke H., Foggia P., Guidobaldi C., et al.* A comparison of algorithms for maximum common subgraph on randomly connected graphs. LNCS, 2002, vol. 2396, pp. 123–132.