Исследование распределения расстояний между графами с упорядоченными вершинами

А.А. Рогов, Н.Д. Москин, Р.В. Воронов, К.А. Кулаков

Петрозаводский государственный университет, г. Петрозаводск, Россия

Аннотация. В статье развивается подход, основанный на вероятностной модели генерации объектов, между которыми находится расстояние. Рассматривается распределение расстояний между графами с упорядоченными вершинами на основе максимального общего подграфа. Одним из возможных вариантов применения подобных расстояний является задача стилистической диагностики текстов. Представлено два алгоритма подсчета расстояния на множестве графов. Один из них заключается в генерации и полном переборе всех пар деревьев, второй – эвристический. Это приближенный метод сбора статистики, где перебирается заданное число пар псевдослучайных деревьев, так как полный перебор может занимать много времени. С помощью этих алгоритмов были найдены матрицы расстояний между деревьями с малым и большим числом вершин n. Результаты экспериментов показали, что при малых n значение метрики не превосходит 0,5. При больших n среднее значение метрики слабо растет и стабилизируется в точке 0,587. Гипотеза о соответствии распределения нормальному закону при n=100 была отвергнута с помощью критерия Пирсона на уровне значимости 0,1.

Ключевые слова: *граф, сравнение, метрика, максимальный общий подграф, упорядочение вершин.*

DOI: 10.14357/20790279240307 **EDN: NCISUO**

Введение

Графовые модели в настоящее время используются очень активно при моделировании различных объектов. При этом возникает задача сравнения двух графовых моделей. Существуют различные метрики для определения расстояния (сходства) между ними [1]. Однако получив значения расстояния, невозможно его оценить по критерию «далеко-близко». Для этого требуется оценка частоты его встречаемости между парами графовых моделей. Решение данной проблемы позволит ответить на вопрос о том, является ли отличие между графовыми моделями существенным или возникшим случайным образом. Одним из методов получения такой оценки является поиск всевозможных числовых значений расстояний между двумя графовыми моделями. При этом большим расстоянием будем называть такое, равное которому и больше которого встречаются редко. С другой стороны, графовые модели можно называть похожими, если расстояние, меньшее расстоянию между этими моделями, встречается редко. В данной статье развивается подход, основанный на вероятностной модели генерации объектов, между которыми находится расстояние, ранее описанный в работах [2–4].

Одним из возможных вариантов применения подобных расстояний является задача стилистической диагностики (подробнее об этом в [1]). Например, для представления синтаксической структуры текста используются так называемые деревья зависимостей (деревья подчинения) и деревья составляющих, сравнивая которые мы можем определить сходство или различие сюжетов. Это требуется, например, для решения задачи атрибуции, в частности, плагиата. Другим важным вариантом применения анализа расстояний является сравнение классификации текстов, полученных с разным набором признаков. Это позволяет выявить наиболее информативные признаки.

Остановимся более подробно на деревьях зависимостей в терминах работы [5]. Если P – конечное линейно упорядоченное множество, то всякое дерево D, для которого P служит множеством узлов, называется деревом синтаксического подчинения на P. Если P – это множество точек не-

которой цепочки x, говорят, что D – дерево синтаксического подчинения для x. Дерево зависимостей $< P; \rightarrow >$ для цепочки x называется проективным, если для любых трех его узлов α, β, γ цепочки x из того, что $\alpha \rightarrow \beta$ и γ лежит между α и β, следует, что γ зависит от α. Дерево зависимостей $< P; \rightarrow >$ называется слабо проективным, если для любых его четырех узлов $\alpha, \beta, \gamma, \delta$ цепочки x из того, что $\alpha \rightarrow \beta$ и $\gamma \rightarrow \delta$ следует, что пары α, β и γ, δ не разделяют друг друга. При изображении деревьев зависимостей слабая проективность означает возможность провести ребра так, чтобы никакие два из них не пересекались.

В качестве примера рассмотрим дерево зависимостей, построенное на основе текстового фрагмента из творчества Ф. М. Достоевского (роман «Идиот»): «Мы уже сказали сейчас что сам генерал хотя был человек и не очень образованный, а, напротив, как он сам выражался о себе, «человек самоучный», но был, однакоже, опытным супругом и ловким отцом» [6]. При этом нумерация вершин дерева соответствует порядку появления соответствующих слов в тексте (рис. 1): первая вершина – первому слову, вторая вершина – второму слову и т. д. В [6] показано, как с помощью деревьев зависимостей и вычисленных на их основе диагностических параметров можно распознать индивидуальный стиль таких известных писателей, как Ф. М. Достоевский, А. П. Чехов, М. А. Шолохов, Л. Н. Толстой и др. Говоря о перспективах подобных исследований И. П. Севбо отмечает, что с помощью теоретико-графовых моделей важно исследовать «многообразие деревьев зависимостей в однородном тексте (перечень типов структур)», «закономерности чередования структур, следующих друг за другом в связном тексте» и пр. В этом случае можно поставить задачу исследования расстояний между деревьями с упорядоченными вершинами.

В работе предлагается способ построения количественных оценок для понятий «редко» и «часто» на основе вероятностного распределения значений мер близости.

1. Метрики на множестве графов

Рассмотрим расстояния на множестве графов, которые позволяют оценить насколько те или иные структуры «похожи» друг на друга. Как правило, эта мера выражает степень неточностей, которые возникают при нахождении изоморфизма графов или подграфов. Одним из способов оценки возможных ошибок сравнения является максимальный общий подграф $mcs(G_i, G_j)$. Очевидно, что чем больше похожи G_i и G_j, тем больше их максимальный общий подграф.

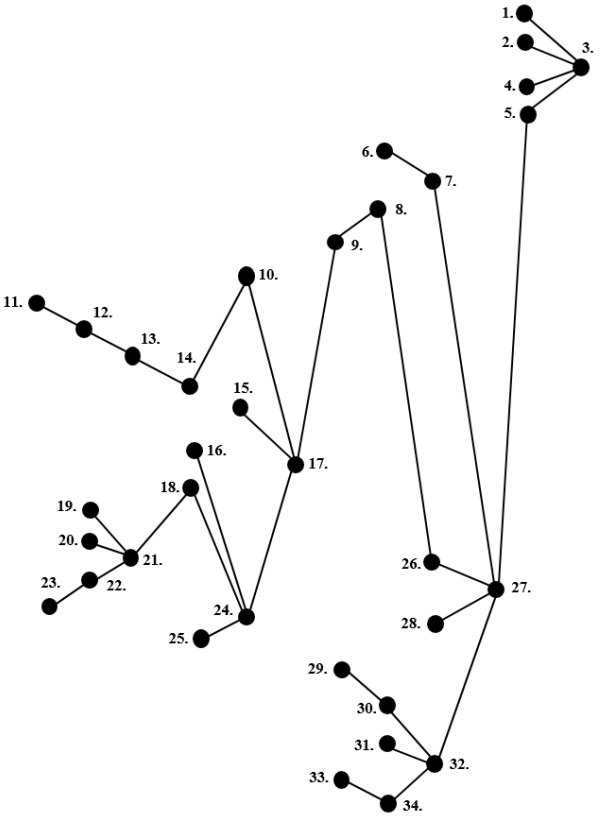

Рис. 1. Дерево зависимостей (Ф. М. Достоевский, фрагмент романа «Идиот»)

В табл. 1 представлен ряд мер (см. [1] и [7]), которых объединяет понятие максимального общего подграфа (здесь $|G|$ – число вершин графа G). Рассмотрим подробнее расстояние:

$$d(G,F) = 1 - \min_{i=1,\ldots,k} \left(\frac{\left| mcs'(g_{\min(i,m)}, f_i) \right|}{i} \right).$$

Здесь для двух помеченных графов $G = (V_1, E_1)$, $F = (V_2, E_2)$ максимальным общим индуцированным помеченным подграфом $mcs'(G,F)$ (MCILS, maximum common induced labeled subgraph) графов G и F назовем граф $mcs'(G,F) = (V,E)$ с максимальным числом вершин, такой, что $V \subset V_1$, $V \subset V_2$ и $mcs'(G,F)$ изоморфен подграфам графов G и F, индуцированным подмножеством вершин V.

При подсчете d учитывается упорядоченность (нумерация) вершин (пусть для определенности число вершин в графах m меньше k, где $|G|=m$, $|F|=k$). Здесь граф g_i является подграфом G, который содержит вершины с номерами от 1 до i и все ребра G, инцидентные этим вершинам (аналогично определяются графы f_i). Эта функция

Табл. 1

Меры на основе максимального общего подграфа.

№	Расстояние между графами	Нормализовано для отрезка [0, 1]	Рассмотрена в работах
1	$d_1(G_1,G_2)=\|G_1\|+\|G_2\|-2\|mcs(G_1,G_2)\|$	Нет	H. Bunke
2	$d_2(G_1,G_2)=1-\frac{\|mcs(G_1,G_2)\|}{\max(\|G_1\|,\|G_2\|)}$	Да	H. Bunke, K. Shearer
3	$d_3(G_1,G_2)=1-\frac{\|mcs(G_1,G_2)\|}{\|G_1\|+\|G_2\|-\|mcs(G_1,G_2)\|}$	Да	W. Wallis, P. Shoubridge и др.
4	$d_4(G,F)=1-\min_{i=1,\dots,k}(\frac{\|mcs'(g_{\min(i,m)},f_i)\|}{i})$	Да	Н. Д. Москин

удовлетворяет всем свойствам метрики (неотрицательность, тождественность, симметричность, неравенство треугольника).

В качестве примера рассмотрим три дерева, изображенные на рис. 2. Для графов G_1 и G_2 максимальный общий подграф изоморфен первому графу G_1 (аналогично для пары G_1 и G_3). Для пары G_2 и G_3 максимальный общий подграф также будет изоморфен графу G_1. Однако меры принимают различные значения (табл. 2).

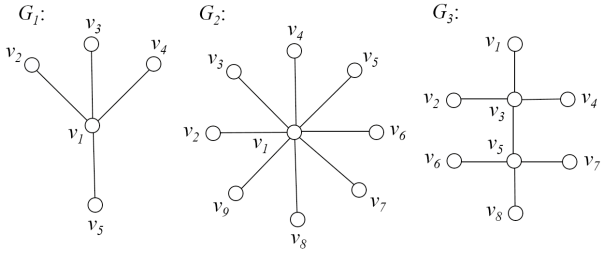

Рис. 2. Три дерева

Для того чтобы в табл. 2 подсчитать расстояния d_4 требуется поставить в соответствие каждой вершине ее порядковый номер. Предположим, номер вершины v_i совпадает с i. Тогда построим цепочки порождающих графов для G_i (в качестве примера на рис. 3 показана такая цепочка для G_1) и найдем число вершин в соответствующих максимальных общих подграфах. Затем подставив в формулу для подсчета d_4, найдем значения метрики.

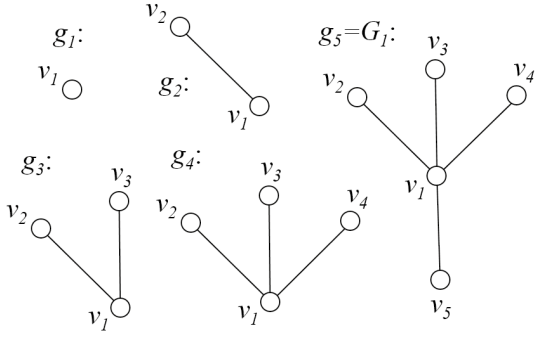

Рис. 3. Подграфы g_i графа G_1

Табл. 2

Меры d_1-d_4 на основе максимального общего подграфа

	G_1 и G_2	G_2 и G_3	G_1 и G_3
d_1	$d_1(G_1,G_2)=4$	$d_1(G_2,G_3)=5$	$d_1(G_1,G_3)=3$
d_2	$d_2(G_1,G_2)=\frac{4}{9}\approx 0{,}444$	$d_2(G_2,G_3)=\frac{1}{3}\approx 0{,}333$	$d_2(G_1,G_3)=\frac{3}{8}=0{,}375$
d_3	$d_3(G_1,G_2)=\frac{4}{9}\approx 0{,}444$	$d_3(G_2,G_3)=\frac{5}{11}\approx 0{,}455$	$d_3(G_1,G_3)=\frac{3}{8}=0{,}375$
d_4	$d_4(G_1,G_2)=\frac{4}{9}\approx 0{,}444$	$d_4(G_2,G_3)=0{,}5$	$d_4(G_1,G_3)=\frac{5}{8}=0{,}625$

Далее исследуем данную метрику более подробно на конечных деревьях с не более чем заданным числом вершин.

2. Алгоритмы подсчета расстояния на множестве графов

Опишем два алгоритма, которые позволяют рассчитать функцию распределения значений метрики на конечных деревьях с не более чем заданным числом вершин.

Первый алгоритм является точным, в нем перебираются все пары деревьев с не более чем N вершинами. Для каждой пары рассчитается значение метрики и строится функция распределения. Второй алгоритм – приближенный, в нем рассчитывается эмпирическая функция распределения путем перебора пар псевдослучайных деревьев и применением эвристического метода для расчета метрики.

Первый алгоритм (полный перебор всех пар деревьев). Деревья представляются при помощи кода Прюфера [8], который взаимно однозначно сопоставляет деревьям с n помеченными вершинами размещения с повторениями из n по (n-2). Известен метод, который для каждого кода дерева строит множество его ребер.

Таким образом, для генерации всех деревьев с не более чем N вершинами необходимо для каждого n = 1,...,N перебирать все размещения с повторениями из n по (n–2) и для каждого размещения получать множество ребер дерева. Обозначим такой алгоритм генерации через Generate (N). Каждый вызов такого алгоритма возвращает множество ребер очередного дерева. Рассмотрим алгоритм, в котором сгенерированные множества ребер всех деревьев запоминаются в массиве *tree*:

Generate_Trees(N)
Вход: число вершин в дереве N
Выход: множество деревьев *tree* и их число M
M = 0
foreach E in *Generate* (N)
  $M = M + 1$
  tree[M] = E
return *tree*, M

Далее для каждой пары различных деревьев ищутся максимальные общие индуцированные помеченные подграфы порождающих подграфов. Пусть $T = (V, E)$ - помеченное дерево, V={1,...,N}. Порождающим подграфом $T(i)$ назовем подграф дерева T, индуцированный подмножеством вершин {1,...,i}.

Пусть $G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$ – два помеченных графа. Назовем модулярным пересечением двух помеченных графов $G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$ граф $G = (V, E)$ такой, что

$$V = V_1 \cap V_2 \text{ и } (v_1, v_2) \in E \Leftrightarrow (v_1, v_2) \in E_1 \text{ и } (v_1, v_2) \in E_2 \text{ или } (v_1, v_2) \notin E_1 \text{ и } (v_1, v_2) \notin E_2.$$

Для поиска MCILS при помощи алгоритма Брона – Кербоша [9] ищется максимальная клика в модулярном пересечении графов. Обозначим алгоритм поиска максимальной клики через *Bron_Kerbosh* (G). На входе граф G, на выходе – число вершин в максимальной клике.

Приведем псевдокод алгоритма построения модулярного пересечения подграфов из двух цепочек порождающих подграфов.

MPG (E_1, E_2, n)
Вход: множества ребер первого и второго дерева E_1, E_2, максимальный используемый номер вершин n
Выход: множество ребер модулярного произведения E
E = 0
for v_1 = 1 to n
  for v_2 = 1 to n
    if (v_1, v_2) in E_1 = (v_1, v_2) in E_2 then
      add (E, (v_1, v_2))

Приведем псевдокод алгоритма расчета метрики для двух деревьев.

Metric (E_1, E_2)
Вход: множества ребер первого и второго деревьев E_1, E_2
Выход: значение метрики
m_1 = число вершин первого графа
m_2 = число вершин второго графа
k = max(m_1, m_2)
d = 1
for i = 1 to k
  G = MPG (E_1, E_2, i)
  c = Bron_Kerbosh (G)
  d = min(d, c / i)
return 1-d

Подсчет статистики для значений метрики осуществляется следующим образом (M – число всех деревьев, полученных процедурой *Generate_Trees*):

```
for i = 1 to M - 1
    for j = i + 1 to M
        d = Metric (tree[i], tree[j])
        count[d] = count[d] + 1
```

Здесь count[d] равно числу пар деревьев, значение метрики между которыми равно d.

Второй алгоритм (эвристический). Так как полный перебор работает очень долго, предлагается следующий приближенный метод сбора статистики с помощью имитационного моделирования. Перебирается заданное число пар псевдослучайных деревьев. Каждое дерево получается из псевдослучайно полученного кода Прюфера. При этом строится псевдослучайное размещение с повторениями (массив из n–2 элементов заполнятся псевдослучайными целыми числами из интервала от 1 до n). Для полученной пары деревьев применяется приближенный алгоритм расчета метрики. Здесь при поиске максимальной клики в модулярном пересечении применяется следующий эвристический алгоритм. На первом шаге ищется вершина u с максимальной степенью и формируется множество $U=\{u\}$. Далее на каждом шаге среди вершин $V \setminus U$ ищется вершина с максимальной степенью, смежная со всеми вершинами множества U. Если такой вершины нет, то алгоритм завершает работу. Если такая вершина найдена, то она добавляется в множество U и делается переход на следующий шаг. Отметим, что задача о клике относится к классу NP-полных задач. Как и для других NP-полных задач, эффективного алгоритма для поиска клики достаточно большого размера на данный момент не найдено.

Табл. 3

Распределение значений метрики для пар деревьев при малых значениях числа вершин n

Значение метрики *d*	**Доля графов пар графов, со значением метрики не больше, чем *d***
n = 3	
0	0,333
0,5	1
n = 4	
0	0,125
0,25	0,208
0,33	0,467
0,5	1
n = 5	
0	0,036
0,2	0,068
0,25	0,214
0,33	0,485
0,4	0,5
0,5	1
n = 6	
0	0,007
0,167	0,018
0,2	0,064
0,25	0,211
0,33	0,517
0,4	0,532
0,5	1

3. Анализ статистики расстояний

Рассмотрим возможные значения расстояний для деревьев с заданным числом вершин n. В табл. 3,4 приведены значения расстояний между возможными парами деревьев. Из табл. 3 видно, что при малых n значение метрики не превосходит 0,5. Однако при больших n ситуация меняется. Как видно из табл. 4, среднее значение метрики слабо растет и стабилизируется в точке 0,587. 90%-ный диапазон значений метрики сужается с ростом n и является более островершинным по сравнению с нормальным законом. Коэффициент асимметрии – отрицательный и приближается к нулю с ростом n. Как видно из рис. 4, распределение значений расстояний является мультимодальным. С ростом n количество мод растет. Попытка аппроксимации распределения расстояний с помощью нормального закона при n=100 закончилась неудачей. Соответствующая гипотеза была отвергнута с помощью критерия Пирсона на уровне значимости 0,1.

Последние 4 столбца табл. 4 позволяют оценить редкие расстояния. Значения, приведенные в этих столбцах, можно назвать критическими. Если будем считать, что близких объектов не более 10% среди всех, то используя столбец 5, получим при n=10, что маленьким расстоянием будем считать любое расстояние меньшее 0,453. Можно заметить, что критическое значение длины маленького расстояния растет с ростом n, от 0,453 (n=10) до 0,561 (n=100). Аналогично определяются аномально большие расстояния. Расстояния между критическими значениями (при n=10 от 0,453 до 0,625) можно считать часто встречающимися.

Табл. 4

Распределение значений метрики для пар деревьев при разных значениях числа вершин *n*

1. К-во вершин	2. Среднее значение	3. Коэффициент асимметрии	4. Длина малых расстояний с общим количеством меньшим 5 % меньше	5. Длина малых расстояний с общим количеством меньшим 10 % меньше	6. Длина больших расстояний с общим количеством меньшим 10 % больше	7. Длина больших расстояний с общим количеством меньшим 5 % больше
10	0,562	−0,317	0,440	0,453	0,625	0,652
15	0,575	−0,255	0,469	0,495	0,638	0,649
20	0,579	−0,230	0,486	0,509	0,630	0,648
30	0,583	−0,187	0,530	0,533	0,632	0,633
40	0,585	−0,139	0,531	0,549	0,623	0,627
50	0,586	−0,133	0,540	0,553	0,619	0,624
75	0,587	−0,103	0,548	0,560	0,613	0,621
100	0,587	−0,085	0,556	0,561	0,610	0,616

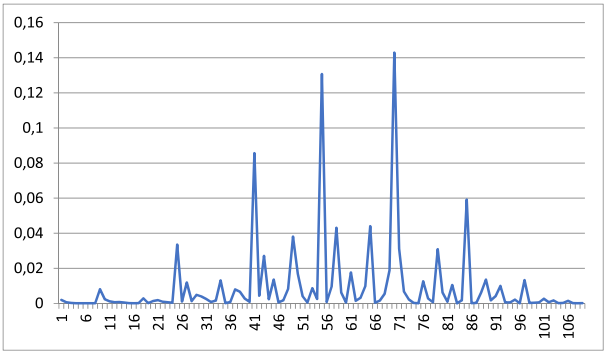

Рис. 4. Распределение частот расстояний для *n*=50

В качестве примера в табл. 5 представлена матрица расстояний между графами фрагментов фольклорных и литературных текстов (описание в [5 и 6]). Первый фрагмент Ф.М. Достоевского (см. рисунок во введении, число вершин *n*=34), автор второго фрагмента М.Ю.Лермонтов (*n*=19), третьего фрагмента – А. С. Пушкин (*n*=21), четвертый фрагмент взят из Голубиной книги (*n*=29), пятый – из былины «Илья Муромец» (*n*=20). Шестой и седьмой фрагменты принадлежат Ф. М. Достоевскому (*n*=22, 27, 41 соответственно). Анализируя табл. 5, можно заметить, что в ней самое маленькое расстояние 0,429, а самое большое 0,732. Это согласуется с проведенными экспериментами. Для более детального объяснения различий между графами требуется помощь специалистов-филологов.

Заключение

В основе предложенной математической модели лежит идея генерации графов с упорядоченными вершинами, между которыми находится расстояние. Были реализованы два алгоритма: эвристический и полный перебор с заданными условиями (при этом деревья представлялись при помощи кода Прюфера). При малом числе вершин *n* значение метрики не превосходит 0,5. При больших *n* среднее значение метрики слабо растет и стабилизируется в точке 0,587. Гипотеза о соответствии распределения нормальному закону при *n*=100 была отвергнута на уровне значимости 0,1. В дальнейшем данный подход можно применить при исследовании других метрик и подмножеств графов.

Табл. 5

Матрица расстояний между графами фрагментов фольклорных и литературных текстов.

	1	2	3	4	5	6	7	8
1	0	0,676	0,667	0,588	0,706	0,667	0,618	0,585
2	0,676	0	0,667	0,552	0,5	0,667	0,63	0,732
3	0,667	0,667	0	0,552	0,429	0,5	0,556	0,683
4	0,588	0,552	0,552	0	0,655	0,552	0,667	0,61
5	0,706	0,5	0,429	0,655	0	0,6	0,571	0,707
6	0,667	0,667	0,5	0,552	0,6	0	0,556	0,732
7	0,618	0,63	0,556	0,667	0,571	0,556	0	0,667
8	0,585	0,732	0,683	0,61	0,707	0,732	0,667	0

Литература

1. *Москин Н.Д.* Теоретико-графовые модели, методы и программные средства интеллектуального анализа текстовой информации на примере фольклорных и литературных произведений. Дис. … докт. техн. наук. Петрозаводск. 2022. 370 с.
2. *Варфоломеев А.Г., Кириков П.В., Рогов А.А.* Вероятностный подход к сравнению расстояний между подмножествами конечного множества // Ученые записки Петрозаводского государственного университета. 2010. №8(113). С. 83-88.
3. *Сидоров Ю.В., Кириков П.В., Рогов А.А.* Сравнение дендрограмм с равным числом вершин // Ученые записки Петрозаводского государственного университета. Серия: Естественные и технические науки. 2011. № 8 (121). С. 108-110.
4. *Rogov A.A., Varfolomeyev A.G., Timonin A.O., Proenca K.A.* A probabilistic approach to comparing the distances between partitions of a set // Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes. 2018. Vol. 14, № 1. P. 14-19. DOI: 10.21638/11701/spbu10.2018.102.
5. *Гладкий А.В.* Синтаксические структуры естественного языка. М.: ЛКИ. 2007. 152 с.
6. *Севбо И.П.* Графическое представление синтаксических структур и стилистическая диагностика. Киев: Наукова Думка. 1981. 192 с.
7. *Москин Н.Д.* Метрика для сравнения графов с упорядоченными вершинами на основе максимального общего подграфа // Прикладная дискретная математика. 2021. № 52. С. 105-113. DOI: 10.17223/20710410/52/7.
8. *Prüfer H.* Neuer Beweis eines Satzes über Permutationen (нем.) // Archiv für Mathematik und Physik. 1918. Bd. 27. P. 742–744.
9. *Bron C., Kerbosh J.* Algorithm 457 – Finding all cliques of an undirected graph // Communications of the ACM. 1973. Vol. 16. P. 575–577.

Рогов Александр Александрович. Петрозаводский государственный университет, г. Петрозаводск, Россия. Заведующий кафедрой. Доктор технических наук, профессор. Область научных интересов: математическое моделирование, прикладная статистика, математические методы распознавания образов, математические методы анализа литературных текстов. E-mail: rogov@petrsu.ru

Москин Николай Дмитриевич. Петрозаводский государственный университет, г. Петрозаводск, Россия. Профессор. Доктор технических наук, доцент. Область научных интересов: цифровые гуманитарные науки, теоретико-графовые модели, интеллектуальный анализ данных, компьютерная лингвистика, мультимедиа-технологии, компьютерная графика. E-mail: moskin@petrsu.ru (ответственный за переписку)

Воронов Роман Владимирович. Петрозаводский государственный университет, г. Петрозаводск, Россия. Профессор. Доктор технических наук, доцент. Область научных интересов: математическое моделирование, задачи оптимизации, комбинаторные задачи на графах, математические методы и модели систем локального позиционирования мобильных объектов. E-mail: rvoronov@petrsu.ru

Кулаков Кирилл Александрович. Петрозаводский государственный университет, г. Петрозаводск, Россия. Доцент. Кандидат физико-математических наук, доцент. Область научных интересов: интеллектуальные пространства, сетевые технологии, мобильные приложения, электронный туризм, мониторинг промышленного оборудования, робототехника, анализ естественных языков. E-mail: kulakov@cs.karelia.ru

Research of the distribution of distances between graphs with ordered vertices

A.A. Rogov, N.D. Moskin, R.V. Voronov, K.A. Kulakov
Petrozavodsk State University, Petrozavodsk, Russia

Abstract. The article develops an approach based on a probabilistic model of generating objects with a distance between them. The distribution of distances between graphs with ordered vertices based on the maximum common is considered. One of the possible applications of such distances is the task of stylistic diagnostics of texts. Two algorithms for calculating distances on a set of graphs are presented. One of them consists of generating and exhaustively enumerating all pairs of trees, the second is heuristic. This is an approximate method of collecting statistics, where a given number of pairs of pseudo-random trees are iterated, since a complete search can take a long time. Using these algorithms, distance matrices between trees with a small and large number of vertices n were found. The experimental results showed that for small n the metric value does not exceed 0,5. For large n the average value of the metric grows slightly and stabilizes at the point 0,587. The hypothesis that the distribution corresponds to the normal law for n=100 was rejected using the Pearson test at a significance level of 0,1.

Keywords: *graph, comparison, metric, maximum common subgraph, vertex ordering.*
DOI: 10.14357/20790279240307 **EDN: NCISUO**

References

1. *Moskin N.D.* Teoretiko-grafovye modeli, metody i programmnye sredstva intellektual'nogo analiza tekstovoj informacii na primere fol'klornyh i literaturnyh proizvedenij [Graph-theoretical models, methods and software for the intellectual analysis of textual information on the example of folklore and literary works]. D.Sc. Diss. Petrozavodsk. 2022. 370 p. (In Russ.)
2. *Varfolomeev A.G., Kirikov P.V., Rogov A.A.* Veroyatnostnyj podhod k sravneniyu rasstoyanij mezhdu podmnozhestvami konechnogo mnozhestva [A probabilistic approach to comparing distances between subsets of a finite set]. Uchenye zapiski Petrozavodskogo gosudarstvennogo universiteta [Scientific notes of Petrozavodsk State University]. 2010. 8(113). P. 83-88. (In Russ.)
3. *Sidorov Y.V., Kirikov P.V., Rogov A.A.* Sravnenie dendrogramm s ravnym chislom vershin [Comparison of dendrograms with an equal number of vertices]. Uchenye zapiski Petrozavodskogo gosudarstvennogo universiteta [Scientific notes of Petrozavodsk State University]. 2011. 8(121). P. 108-110. (In Russ.)
4. *Rogov A.A., Varfolomeyev A.G., Timonin A.O., Proenca K.A.* A probabilistic approach to comparing the distances between partitions of a set. Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes. 2018. 14(1): 14-19. DOI: 10.21638/11701/spbu10.2018.102.
5. *Gladkij A.V.* Sintaksicheskie struktury estestvennogo yazyka [Syntactic structures of natural language]. Moscow, LKI Publ. 2007. 152 p. (In Russ.)
6. *Sevbo I.P.* Graficheskoe predstavlenie sintaksicheskih struktur i stilisticheskaya diagnostika [Graphic representation of syntactic structures and stylistic diagnostics]. Kiev: Naukova Dumka. 1981. 192 p. (In Russ.)
7. *Moskin N.D.* Metrika dlya sravneniya grafov s uporyadochennymi vershinami na osnove maksimal'nogo obshchego podgrafa [Metric for comparing graphs with ordered vertices based on maximum common subgraph]. Prikladnaya diskretnaya matematika [Applied discrete mathematics]. 2021. 52. P. 105-113. DOI: 10.17223/20710410/52/7. (In Russ.)
8. *Prüfer H.* Neuer Beweis eines Satzes über Permutationen (нем.) // Archiv für Mathematik und Physik. 1918. 27: 742–744.
9. *Bron C., Kerbosh J.* Algorithm 457 – Finding all cliques of an undirected graph // Communications of the ACM. 1973. 16. P. 575–577.

Rogov Aleksandr A. Head of Department, Doctor of Technical Sciences, Professor, Petrozavodsk State University, 33 Lenin str., Petrozavodsk, 185910, Russia. Research interests: mathematical modeling, applied statistics, mathematical methods of pattern recognition, mathematical methods of analysis of literary texts. E-mail: rogov@petrsu.ru

Moskin Nikolai D. Professor, Doctor of Technical Sciences, Associate Professor, Petrozavodsk State University, 33 Lenin str., Petrozavodsk, 185910, Russia. Research interests: digital humanities, graph-theoretical models, data mining, computational linguistics, multimedia technologies, computer graphics. E-mail: moskin@petrsu.ru

Voronov Roman V. Professor, Doctor of Technical Sciences, Associate Professor, Petrozavodsk State University, 33 Lenin str., Petrozavodsk, 185910, Russia. Research interests: mathematical modeling, optimization problems, combinatorial graph problems, mathematical methods and models of local positioning systems for mobile objects. E-mail: rvoronov@petrsu.ru

Kulakov Kirill A. Associate Professor, PhD in Physics and Mathematics, Associate Professor, Petrozavodsk State University, 33 Lenin str., Petrozavodsk, 185910, Russia. Research interests: intelligent spaces, network technologies, mobile applications, electronic tourism, monitoring of industrial equipment, robotics, natural language analysis. E-mail: kulakov@cs.karelia.ru