МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра математической кибернетики и компьютерных наук

РЕАЛИЗАЦИЯ И СРАВНЕНИЕ МЕТРИК СХОЖЕСТИ ГРАФОВ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 451 группы
направления 09.03.04 — Программная инженерия
факультета КНиИТ
Петрова Владимира Сергеевича

Научный руководитель
доцент, к. ф.-м. н. _____ С. В. Миронов

Заведующий кафедрой
к.ф.-м.н. _____ С. В. Миронов

Саратов 2019

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Актуальность темы

Графы являются одной из самых широко исследуемых областей в современных компьютерных науках. Очень многие задачи многих областей науки можно привести к графовому представлению, поэтому изучение графов и построение и анализ алгоритмов, применяемых к ним, являются очень востребованными областями исследований в последние годы.

Задача исследования графов на схожесть очень часто встречается, например, в экономической и новостной отраслях, так как при помощи графов специального вида можно отслеживать сильные изменения на торговых биржах или же какие-либо новостные всплески. Также схожесть графов используется в медицине или физике, например, для отслеживания молекул с похожей структурой.

Целью данной работы является изучение различных метрик оценки схожести графов и их сравнение друг с другом по критериям сложности реализации, времени работы и полученным результатам при анализе реальных данных в экономической среде.

В результате написания работы должны быть решены следующие задачи:

- нахождение реальных данных, к которым можно применить описываемые алгоритмы;
- структурирование больших объемов данных;
- реализация различных алгоритмов сравнения графов на схожесть (а именно, дистанции Хемминга, меры Жаккара, d-меры, δ-меры, d_1-меры и d_2-меры;
- применение вспомогательных алгоритмов для реализации некоторых мер (а именно, d_1-меры и d_2-меры);
- применение реализованных алгоритмов к полученным данным и анализ результатов;
- анализ сложности алгоритмов и их сравнение по различным критериям.

Цель бакалаврской работы — изучение различных метрик оценки схожести графов и их сравнение друг с другом по критериям сложности реализации, времени работы и полученным результатам при анализе реальных данных в экономической среде.

Поставленная цель определила **следующие задачи**:

1. найти реальные данные, к которым можно применить описываемые алгоритмы и структурировать их;
2. реализовать различные алгоритмы сравнения графов на схожесть;
3. применить реализованные алгоритмы к полученным данным и проанализировать результаты;
4. проанализировать сложности алгоритмов и сравнить их по различным критериям.

Методологические основы методов сравнения графов на схожесть представлены в работах Bunke, Papadimitriou, Broder, Kruskal, Sidorov и Aleskerov.

Практическая значимость бакалаврской работы

Данная работа является значимой, так как задача сравнения графов очень сильно помогает во многих областях науки, следовательно, требует плотного изучения для улучшения практических результатов и нахождения более оптимальных и точных алгоритмов ее решения. Также реализация, модификация и сравнение некоторых существующих алгоритмов позволит создать плацдарм для дальнейшего изучения данной темы, а также очень сильно поможет в сравнении различных метрик (уже существующих и новых) между собой.

Структура и объем работы

Бакалаврская работа состоит из введения, двух разделов, заключения, списка использованных источников и двух приложений. Общий объем работы — 49 страниц, из них 40 страниц — основное содержание, включая 7 рисунков, список использованных источников информации — 35 наименований.

1 Задача сравнения графов и методы ее решения

Этот раздел посвящен описанию проблемы сравнения двух графов на схожесть, обзору существующих подходов, их плюсов и минусов и обзору реализуемых в работе алгоритмов с их подробным объяснением.

В научных исследованиях графы показали себя как очень хороший способ представления информации о сетях или других похожих объектах (например, картах). Изучение графов — очень востребованная и очень сложная область в современных компьютерных науках. Сложные системы встречаются очень часто и им необходимо иметь какое-то удобное представление для работы. Обычно графы являются достаточно гибкой структурой представления сложных объектов, так как позволяют поддерживать сами компоненты системы и их иерархию или же влияние друг на друга. Также вследствие того, что системы могут меняться во времени, некоторые компоненты или связи между ними могут появляться или удаляться. Графы также хорошо могут отображать такие изменения [1].

Задача нахождения корректных метрик для оценки схожести графов является одной из очень плотно исследуемых в течение последних десятилетий. Численные алгоритмы, техники и метрики могут быть сгруппированы в несколько главных категорий: расстояние редактирования/изоморфизм графов [2, 3], общие подграфы, статистические методы (извлечение особенностей) [4] и итеративные методы. Другим простым методом оценить похожесть двух сетей является нахождение корреляции Пирсона между матрицами смежности соответствующих сетей. Тем не менее, оценка корреляции матриц смежности не является хорошим способом сравнения графов, потому что она не берет в учет топологию этих сетей и считает все ребра одинаково значимыми. Другими подходами являются, например, подсчет размера пересечений между ребрами и вершинами графов [5], подсчет расстояния редактирования между двумя графами [2], или подсчет максимального общего подграфа.

Численная оценка схожести и определение изоморфизма между графами являются фундаментальными открытыми задачами в компьютерных науках. Задача определения изоморфизма заключается в определении идентичности двух графов при сопоставлении каждой компоненты одного графа ровно одной компоненте другого графа. Эта задача заслуживает отдельного места в области оценки сложности, так как алгоритма, работающего за полиномиаль-

ное время, до сих пор не было найдено. Таким образом, ее сложность остается неопределенной с середины 70-х годов. Недавняя работа предлагала квазиполиномиальный алгоритм, который проверяет подразделы графов на изоморфизм при помощи серии простых алгоритмов. Тем не менее, для структур с высокой степенью симметричности задача до сих пор остается очень сложной в плане вычисления.

На практике же численная оценка схожести графов дает намного больше информации о графах, нежели двоичный ответ на задачу об их изоморфизме. Оценка схожести имеет очень много применений из-за очень широкой области использования сетей в социальных науках, медицине, биологии, физике, экономике и так далее. Среди многих примеров, оценка может помочь различать неврологические расстройства путем количественной оценки топологического и функционального сходства, чтобы находить молекулы, имеющие похожие свойства, для разработки лекарств или же для определения численных изменений среди эволюционирующих с течением времени сетей.

Далее описаны метрики, используемые в данной работе. Стоит условиться, что все нижеописанные метрики сравнивают графы таким образом, что если графы идентичны, то возвращается ноль, а если же графы полностью противоположны, то возвращается единица. Иными словами, ниже приведенные методы возвращают нормализованное расстояние между двумя графами. Также далее будет использоваться константа $GSIZE$, которая является абстрактным обозначением максимального количества вершин в рассматриваемом графе.

Первой описываемой мерой является дистанция Хемминга, которая используется для сравнения двух бинарных объектов одинаковой длины на схожесть. Она определяется как отношение количества несовпадающих бит на одинаковых позициях в этих объектах к общему числу бит в этих объектах.

Более формально:

$$hamming(g_1, g_2) = \frac{\sum\limits_{i \neq j}^{GSIZE} g_{1_{i,j}} \neq g_{2_{i,j}}}{GSIZE(GSIZE-1)}.$$

Второй метрикой является мера Жаккара, схожая по своей сути с дистанцией Хемминга. Она является мерой сравнения схожести множеств и вы-

числяется как единица минус отношение размера пересечения этих множеств к размеру их объединения.

Более формально:

$$jaccard(g_1, g_2) = 1 - \frac{\sum\limits_{i \neq j}^{GSIZE} g_{1_{i,j}} \ \& \ g_{2_{i,j}}}{\sum\limits_{i \neq j}^{GSIZE} g_{1_{i,j}} \mid g_{2_{i,j}}},$$

где $\&$ — операция побитового И, а $\mid$ — операция побитового ИЛИ.

Далее определяется описанная в работе [6] Фуада Алескерова мера d, которая позволяет сравнить графы на схожесть, принимая во внимание только центральности вершин этих графов. Для подсчета центральностей вершин используется алгоритм PageRank, придуманный компанией Google в 1998 году и использующийся ей для оценки релевантности результатов поиска в поисковых системах. Он очень хорошо масштабируется на различные виды графов и проверен временем.

d-мера вычисляется при помощи вспомогательной матрицы r, определяемую следующим образом:

$$r_{i,j} = \begin{cases} 1, c_i - c_j > \varepsilon \\ 0, otherwise. \end{cases}$$

для всех $1 \leq i, j \leq GSIZE$, где c_i равно величине центральности i-й вершины графа (в данной работе для определения центральности используется вышеупомянутый алгоритм PageRank). Значение ε вычисляется при помощи одной из формул, описанных в работе, и равно некоторому проценту выборочной дисперсии центральности.

Далее значение d-меры между графами g_1 и g_2 определяется следующим образом:

$$d(g_1, g_2) = \frac{\sum\limits_{i \neq j}^{GSIZE} |r_{1_{i,j}} - r_{2_{i,j}}|}{GSIZE(GSIZE - 1)},$$

где r_1 — матрица r для графа g_1, а r_2 — матрица r для графа g_2.

Следующей мерой является δ-мера. Она оценивает схожесть двух графов

в плане их топологии. Для δ-меры вспомогательная матрица c для графа g определена следующим образом:

$$c_{i,j} = g_{i,j}/maxdeg(g)$$

для всех $1 \leq i, j \leq GSIZE$, где $maxdeg(g) = \max\limits_{i=1}^{GSIZE} \sum\limits_{j=1}^{GSIZE} g_{i,j}$, иными словами, максимальная степень вершины среди всех вершин графа. Данная формула адаптирована под конкретную задачу этой работы. В общем случае стоит заметить, что максимальная степень берется именно в случае невзвешенных графов, в случае взвешенных графов необходимо взять для нормализации вес максимального ребра в графе вместо степени.

Тогда δ-мера для графов g_1 и g_2 вычисляется следующим образом:

$$\delta(g_1, g_2) = \frac{\sum\limits_{i,j}^{GSIZE} |c_{1_{i,j}} - c_{2_{i,j}}|}{GSIZE^2 \cdot \gamma},$$

где $\gamma = \max\limits_{i,j}(c_{1_{i,j}}, c_{2_{i,j}})$, иными словами γ является коэффициентом нормирования.

Оставшиеся две метрики используют d-меру и δ-меру для получения такой меры, которая учитывает значения центральностей вершин и топологию графов для более точного сравнения двух графов. Первая метрика была названа d_1-мерой, а вторая — d_2-мерой. d_1-мера вычисляется по следующей формуле:

$$d_1(g_1, g_2) = \sqrt{\frac{d(g_1, g_2)^2 + \delta(g_1, g_2)^2}{2}}.$$

Здесь же $d_1(g_1, g_2)$ равно 1, если графы абсолютно различны, и 0, если идентичны.

И, наконец, вторая метрика d_2 выглядит следующим образом:

$$d_2(g_1, g_2) = \alpha d(g_1, g_2) + (1 - \alpha)\delta(g_1, g_2),$$

где α — параметр, отвечающий за то, в каком отношении необходимо брать d-меру и δ-меру. Эта формула очень хорошо подходит для таких задач, где приоритет нужно отдать одной из характеристик схожести.

Таким образом, были рассмотрены многие подходы к сравнению графов на схожесть, была упомянута задача об изоморфизме, которая предоставляет намного меньше информации, нежели задача численной оценки схожести графов, были представлены примеры использования решения задачи сравнения графов в реальных науках и описаны некоторые метрики для сравнения и улучшения.

2 Получение данных и сравнение метрик

Этот раздел посвящен реализации описанных в теоретической части метрик, вспомогательных алгоритмов для их подсчета и программного кода, позволяющего привести в необходимый для работы вид реальные данные, на которых потом было проведено тестирование реализованных алгоритмов.

Изначально данные заданы в виде одного файла в формате `.csv`, в котором первая строка содержала номер строки (0), затем название второго поля для дат (Dates), отделенное слева и справа через точку с запятой, а затем названия всех 194 компаний, разделенных между собой через точку с запятой. В следующих строках (до конца файла) находились данные в формате: номер строки, торговый день в формате `dd.mm.yyyy`, а далее цена акций каждой из компаний в порядке описания их в первой строке входных данных, также разделенных через точку с запятой. Цены акций — действительные числа, записанные с разным количеством знаков после десятичной точки.

Для каждой пары (день, компания) были получены цены акций заданной компании в заданный день. Пусть это значение равно $P_{i,j}$, где i — номер компании, если пронумеровать их от 1 до 194 в порядке входных данных, а j — номер дня, если нумеровать их в порядке входных данных, начиная с 1 и заканчивая 1665 (всего в рассматриваемом промежутке времени с 10.01.2012 по 09.04.2019 находится 1665 торговых дней).

После обработки входных данных в предыдущем листинге был сделан переход к логарифмическим доходностям для каждой компании. Это необходимо для того, чтобы сгладить разницу между большими и маленькими компаниями, разница в доходах между которыми может быть колоссальной, но это не означает, что какую-то из этих компаний следует учитывать в большей мере, чем другую. Логарифмическая доходность компании i в день j равна $LP_{i,j} = \log(\frac{P_{i,j}}{P_{i,j-1}})$. Очевидно, что таких значений ровно 1664 и строятся они со дня 2 по день 1665.

Затем были построены неориентированные невзвешенные графы, описывающие корреляции между компаниями. Было взято плавающее окно в 250 торговых дней, для этого окна между каждой парой компаний была посчитана корреляция их логарифмических доходностей в рассматриваемый период времени по формуле Пирсона, которая для выборок длины n двух случайных величин $x = (x_1, x_2, \ldots, x_n)$ и $y = (y_1, y_2, \ldots, y_n)$ выглядит следующим

образом:

$$r_{x,y} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}},$$

где $\bar{x} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$, а $\bar{y} = \frac{1}{n}\sum\limits_{i=1}^{n} y_i$.

Затем по корреляции было определено расстояние между компаниями по формуле $d_{x,y} = \sqrt{1 - 2r_{x,y}}$, которое является по сути обратным к значению корреляции, и если оно не превосходило определенный порог $\theta = 1.2$, то это означало, что в графе корреляций между этой парой компаний должно быть неориентированное ребро (эти величины коррелируют между собой).

Это окно сдвигалось на 20 следующих дней до тех пор, пока в нем находилось 250 торговых дней (то есть после построения графа для первых 250 дней с 1 до 250 происходило построение с 21 до 270 и так далее). Таким образом, при построении получилось построить 71 граф, каждый из которых состоит из 194 вершин.

В следующей части были реализованы все описанные в теоретической части работы метрики. Для подсчета d-меры был реализован алгоритм PageRank и нахождение вспомогательной матрицы r, для δ-меры также была посчитана вспомогательная матрица c, затем при помощи вспомогательной программы был произведен подсчет схожестей по каждой из шести метрик для реальных экономических данных, полученных в вышеописанной части работы.

Для каждой из заданных шести метрик была построена матрица s размера 71×71, где $s_{i,j}$ обозначало схожесть графов i и j по этой метрике. Очевидно, что элементы на диагонали матрицы равны нулям.

Полученные результаты можно применить для нахождения важных экономических событий (например, сильных потрясений на рынке). Если построить графики по последовательным значениям $s_{i,i+1}$, то сильные всплески на этих графиках будут означать сильные различия между двумя соседними во времени графами. Это и будет значить то, что в этот период произошли какие-то сильные изменения на рынке.

Стоит отдельно заметить, что дистанция Хемминга очень плохо работает на разреженных сетях из-за того, что считается отношение какого-то небольшого количества ребер к максимально возможному количеству ребер в графе,

которое может достигать практически квадрата от количества вершин. Мера Жаккара в этом плане работает лучше, но стоит понимать, что она никак не учитывает никак центральности вершин, а также при небольшом количестве ребер в сети тоже может вести себя непредсказуемо из-за того, что каждое добавленное или удаленное ребро может очень сильно изменить отношение.

Сравнительный анализ по времени работы:

Дистанция Хемминга вычисляется за время $O(GSIZE^2)$, так как вычисления содержат два вложенных цикла, которые выполнят не более $GSIZE$ действий каждый. Мера Жаккара также вычисляется за время $O(GSIZE^2)$ по той же самой причине. d-мера вычисляется за время работы $O(GSIZE^2)$ (два вложенных цикла, каждый делает не более $GSIZE$ дейтвий) при вычисленной матрице центральностей, которая, в случае с алгоритмом PageRank, не имеет четкого асимптотического времени работы, но эмпирически было показано, что на достаточно больших графах она работает не дольше, чем $O(GSIZE^2 \log GSIZE)$ (каждая итерация работает за $O(GSIZE^2)$, потому что имеет в себе два вложенных цикла, каждый из которых делает не более $GSIZE$ действий, а всего итераций для достаточной сходимости алгоритма требуется не более, чем $O(\log GSIZE)$). δ-мера вместе с вспомогательной матрицей вычисляется за время $O(GSIZE^2)$. d_1-мера и d_2-мера вычисляются за $O(1)$ (подсчет простой формулы) при посчитанных d-мере и δ-мере.

Таким образом, был проведен анализ шести реализованных метрик сравнения графов, основанный на результатах, полученных при анализе реальных экономических данных, а также было проведено сравнение реализованных метрик между собой по различным критериям.

ЗАКЛЮЧЕНИЕ

Были проанализированы шесть различных метрик оценки численного сходства для пары графов, реализован код, позволяющий вычислять их, и проведен анализ времени работы каждого из этих методов. Также реализованный код был применен на реальных данных, взятых из экономической сферы, и результаты также были проанализированы и сопоставлены с реальной информацией за прошедший период времени.

В ходе написания работы были решены следующие задачи:

- Были найдены реальные данные, на которых можно провести анализ реализованных программ;
- эти данные были структурированы и приведены в вид, готовый к применению реализованных алгоритмов;
- было реализовано шесть метрик сравнения графов на схожесть;
- были реализованы вспомогательные алгоритмы для применения вышеперечисленных метрик;
- реализованные алгоритмы были применены к полученным данным, результаты были проанализированы;
- было проведено сравнение реализованных алгоритмов в плане сложности реализации и времени работы.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 *Sidorov, S.* Measuring long-range correlations in news flow intensity time series / S. Sidorov, A. Faizliev, V. Balash // *International Journal of Modern Physics C*. — 2017. — Vol. 28, no. 08. — P. 1750103. https://doi.org/10.1142/S0129183117501030.

2 *Bunke, H.* A Graphtheoretic Approach to Enterprise Network Dynamics / H. Bunke, P. Dickinson, M. Kraetzl, W. Wallis. — Birkhauser, Boston, 2007.

3 *Kruskal, J. B.* Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis / J. B. Kruskal // *Psychometrika*. — Mar 1964. — Vol. 29, no. 1. — Pp. 1–27. https://doi.org/10.1007/BF02289565.

4 *Papadimitriou, P.* Web graph similarity for anomaly detection / P. Papadimitriou, A. Dasdan, H. Garcia-Molina // *Journal of Internet Services and Applications*. — May 2010. — Vol. 1, no. 1. — Pp. 19–30. https://doi.org/10.1007/s13174-010-0003-x.

5 Graph structure in the web / A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wienerc // *Computer Networks*. — 2000. — Vol. 33. — Pp. 309 – 320. http://snap.stanford.edu/class/cs224w-readings/broder00bowtie.pdf.

6 *Aleskerov, F.* Stability and similarity in networks based on topology and nodes importance // Complex Networks and Their Applications VII / Ed. by L. M. Aiello, C. Cherifi, H. Cherifi, R. Lambiotte, P. Lió, L. M. Rocha. — Cham: Springer International Publishing, 2019. — Pp. 94–103.