УДК 519.677 **DOI** 10.15622/sp.61.4

В.В. Печенкин, М.С. Королёв, Л.В. Димитров

ПРИКЛАДНЫЕ АСПЕКТЫ ИСПОЛЬЗОВАНИЯ АЛГОРИТМОВ РАНЖИРОВАНИЯ ДЛЯ ОРИЕНТИРОВАННЫХ ВЗВЕШЕННЫХ ГРАФОВ (НА ПРИМЕРЕ ГРАФОВ СОЦИАЛЬНЫХ СЕТЕЙ)

Печенкин В.В., Королёв М.С., Димитров Л.В. **Прикладные аспекты использования алгоритмов ранжирования для ориентированных взвешенных графов (на примере графов социальных сетей).**

Аннотация. Рассматриваются прикладные аспекты использования предварительного ранжирования вершин ориентированного взвешенного графа. Особое внимание уделяется широкому использованию такого приема в разработке эвристических алгоритмов дискретной оптимизации. Задача ранжирования имеет непосредственное отношение к проблеме определения центральности в социальных сетях, обработке больших массивов данных реального мира, но, как показано в статье, явно или косвенно используется при разработке алгоритмов решения прикладных задач в качестве начального этапа построения решения. Приводятся примеры использования предварительного ранжирования, в которых продемонстрировано повышение эффективности решения некоторых прикладных задач, имеющих широкое применение в математических методах оптимизации. Дано описание структуры первой фазы вычислительного эксперимента, которая связана с получением тестовых наборов данных. Полученные данные представлены взвешенными графами, которые соответствуют нескольким группам социальной сети ВКонтакте с числом вершин в диапазоне от 9000 до 24 тысяч участников. Показано, что структурные характеристики полученных графов по числу компонент связности существенно различаются. Продемонстрированы некоторые характеристики центральности (распределения степенных последовательностей), которые имеют экспоненциальный характер. Основное внимание уделяется анализу трех алгоритмов построения иерархии ранжирования вершин графов, предлагаются новые подходы к вычислению рангов вершин с использованием информации об активности пользователей в социальных сетях. Проводится сравнение распределений полученных совокупностей рангов. Вводится понятие сходимости алгоритмов ранжирования вершин графов, а также обсуждаются различия их использования при рассмотрении данных большой размерности и необходимости построения решения в случае учета только локальных изменений.

Ключевые слова: ранжирование, ориентированный граф, взвешенный граф, инкрементальный алгоритм, локальный алгоритм.

1. Введение. Рассматривается задача ранжирования вершин ориентированного взвешенного графа с точки зрения использования предварительного ранжирования при решении некоторых оптимизационных задач. Отметим, что в работе в качестве исходных эмпирических данных используются графы социальных сетей, что несколько сужает прикладное значение полученных результатов. Тем не менее ниже продемонстрировано, что алгоритм предварительного ранжирования является универсальным подходом во многих оптимизационных задачах, но в каждом конкретном случае требуется дополнительный анализ.

Постановка задачи имеет достаточно длинную историю и разнообразные подходы к ее решению. Алгоритмы ранжирования предложены в большом количестве, и имеется анализ их использования и реализации. В настоящее время в связи с тем, что появляются структурированные данные большой размерности — графы с десятками, сотнями тысяч вершин, эта задача вновь становится актуальной. Такого рода структурированные данные необходимо обрабатывать и анализировать их структуру. При этом важной задачей становится снижение временной сложности алгоритмов обработки за счет распараллеливания и обработки только локальной окрестности той части графа, в которой произошли изменения. То есть в этом случае решение задачи для всего графа можно получить из имеющегося глобального решения за счет его локального изменения на ограниченной окрестности вершин. Такая постановка задачи в современных исследованиях связана с понятием инкрементального алгоритма поиска решения, который для коррекции глобального решения использует обработку вершин, находящихся в локальной окрестности части графа, подвергшейся изменению [1, 2].

Далее будут приведены примеры использования предварительного ранжирования вершин графа, что приводит к повышению эффективности решения уже конкретной задачи. Другой рассматриваемой задачей статьи является анализ результатов конкретных алгоритмов ранжирования на предмет типа распределения полученной совокупности рангов.

2. Формальная постановка задачи. Мы рассматриваем задачу построения такого отображения множества вершин ориентированного взвешенного графа в множество вещественных чисел, которое отражает доминирование одной вершины по отношению к другой. Алгоритмы такого рода используются при ранжировании сайтов в сети Интернет и имеют приложения к разнообразным задачам. В качестве определения ориентированного взвешенного графа мы используем стандартное понятие графа $G=(V, E, w)$, где V — множество вершин графа, E — множество дуг (ориентированных ребер), $w: V \to R$ — функция веса дуг. Смысловым источником такого рода задач является классическая задача топологической сортировки, которая имеет формальную постановку, опирающуюся на следующие определения. В том случае, если функция веса действует тривиальным способом, задавая вес любой дуги равным 1, мы будем опускать ее при определении графа.

Определение 1. Дан граф $G=(V, E)$, говорим, что вершина v достижима из вершины u, если $v = u$, или $(u, v) \in E$, или существует путь $p = v_1, v_2, \ldots, v_n$ в графе G такой, что $v_1 = u$ и $v_n = v$. Факт достижимости в этом случае обозначаем как *u ~> v.* Для понятия «взвешенный граф», которое предполагает наличие функции веса для ориентированных дуг, используется такое же определение.

Близкой по своей природе к задаче ранжирования является задача топологической сортировки (ТС), которая может быть сформулирована следующим образом: топологической сортировкой графа является отображение $f : V \rightarrow R$ такое, что справедлива формула (1):

$$\forall v, u \in V(v \sim> u) \rightarrow (f(v) < f(u)). \tag{1}$$

Классическая постановка задачи о ТС накладывает серьезные ограничения на граф — он должен быть ациклическим. Заметим, что в общем случае такие графы являются достаточно редкими, и необходимы процедуры ранжирования, пригодные для произвольных графов. Задача ТС может быть решена в ходе обхода графа в ширину и имеет сложность $O(n+m)$, где n — количество вершин графа, m — количество дуг.

Под формальным определением задачи ранжирования вершин ориентированного взвешенного графа мы будем понимать задачу, описанную следующим образом. Под ранжированием понимается определение функции f, как это сделано при определении топологической сортировки, определенной выше, но имеющей следующие существенные отличия:

1. Допускается существование таких пар вершин графа $v, u \in V$, что $f(v) = f(u)$.
2. Отношение ранжирования строится на основе композиции отношений ранжирования для локального окружения каждой пары вершин.
3. Снимается ограничение об отсутствии циклов в графе.

Анализируемые далее алгоритмы могут быть разбиты в соответствии с нашими установками на два класса — локальные и глобальные. *Локальные* алгоритмы при вычислении ранга вершины учитывают структурные характеристики локальной окрестности вершины в графе. При вычислении ранга вершины в этом случае учитывается только направленность и вес дуг, инцидентных данной вершине. *Глобальные* алгоритмы получают решение в результате последовательности итераций, учитывающих изменения текущего решения для всех вершин графа. Использованные далее способы вычисления рангов были адаптированы для нашей задачи и нормализованы к диапазону сравнимых значений.

3. Алгоритмы ранжирования в решении прикладных задач. Многие приложения (анализ структуры указателей при организации блоков памяти операционной системы, инкрементальная компиляция) сводятся к реализации ТС ациклических ориентированных графов в условиях их динамического изменения [3]. Вычисление различных характеристик динамически изменяющихся графов представляет определенный интерес, и в этом плане наибольшее значение имеют так называемые инкрементальные алгоритмы, которые позволяют сконструировать глобальное решение для всего графа, учитывая только локальные изменения и имеющееся решение задачи до момента изменения. Так для решения задачи построения ТС вершин графа разработаны эффективные алгоритмы динамической сортировки, которые учитывают операции добавления/удаления ребер графа и позволяют корректировать имеющееся решение сразу после изменений [4, 5].

Опишем ряд результатов, которые показывают, что задача ранжирования вершин ориентированного графа имеет большое значение для повышения эффективности оптимизационных алгоритмов. Предварительное ранжирование и индексирование данных является традиционным способом повышения скорости выполнения запросов к реляционным базам данных. Аналогичная ситуация имеет место и при работе с NoSQL базами данных (Neo4j, DEX), которые содержат информацию о динамических структурах с большим количеством отношений различного типа [6, 7].

Повышение скорости работы алгоритмов на основе ранжирования вершин графа предлагается и для решения классических оптимизационных задач. В работе [8] описан муравьиный алгоритм решения обобщенной оптимизационной задачи (рассматривается задача коммивояжёра) на основе набора эвристик. На основе инкрементального определения концентрации искусственного феромона определяются вероятности выбора следующей вершины графа при выборе продолжения промежуточного решения. В процессе работы алгоритма все вершины графа ранжируются и вероятность выбора следующей вершины зависит от ее ранга.

Работа [9] содержит описание двух алгоритмов ранжирования вершин ориентированных и неориентированных графов, которые используют матрицу Лапласа. Особое внимание к этой проблеме связано с задачей обучения интеллектуальных систем принятия решений в условиях динамического изменения данных различной природы. В качестве примеров, используемых для демонстрации эффективности алгоритмов, рассматриваются два набора данных. Первый имеет отношение к компьютерной биологии и содержит иерархическую классификацию протеинов. Второй содержит разбиение текстовых новостных статей на группы по связанным с ними тегам. Результаты вычислительного эксперимента показывают уменьшение суммарной ошибки ранжирования при использовании динамической процедуры. К особенностям рассматриваемой задачи относится то, что на структуру графов накладываются серьезные ограничения — они должны быть *к*-дольными, а в качестве используемых массивов для вычислительного эксперимента выбраны данные именно такой природы.

В работе [10] рассматривается оптимизационная задача поиска кратчайшего пути на местности, описываемой географической картой. Особенностью предложенного в работе подхода является использование математического аппарата нечетких графов, у которых функция принадлежности ребра имеет вид:

$$\mu : V^2 \to [0,1].$$

Решаемая задача связана с нечеткостью определения на географической карте объектов (дорог и полигонов) и, соответственно, расстояний между ними. Предлагается переход от карты к нечеткому графу, учитывая параметры сетей дорог и полигонов, подлежащие оптимизации. Длина пути в графе вычисляется подобно обычному алгоритму для взвешенного графа (сумма весов дуг, входящих в путь, должна быть минимальной), но используются операции сложения и нахождения экстремумов для нечетких чисел. Ранжирование вершин используется при вычислении кратчайших путей в графах, соответствующих картам.

Похожий подход используется в работе [11], в которой задается граф с нечеткими весами дуг и трапециевидной функцией принадлежности для решения задачи повышения скорости передачи данных в оптических сетях. Кратчайшие пути в этом случае ранжируются предлагаемым авторами алгоритмом. Результаты описанного вычислительного эксперимента показывают повышение производительности маршрутизации, которое связано с уменьшением времени на поиск оптимального пути и увеличением пропускной способности сети от 10 до 20%.

Ранжирование как один из этапов оптимизационного алгоритма часто используется в математическом моделировании систем видеонаблюдения. В работе [12] решается задача повторной идентификации человека в видеопотоке, который получается с нескольких камер видеонаблюдения. Полученное множество оценок видеоизображений людей преобразуется в графовую модель, для которой решается задача ранжирования, упрощающая построение консенсусной оценки. В [13] на основе ранжирования решается задача поиска скрытых объектов при визуальном отслеживании некоторой сцены.

4. Исходные данные и сравнение результатов ранжирования. Для проведения вычислительного эксперимента используются наборы данных, которые были получены при анализе структурных характеристик виртуальных сообществ социальной сети ВКонтакте. Схема проведенного эксперимента представлена на рисунке 1.

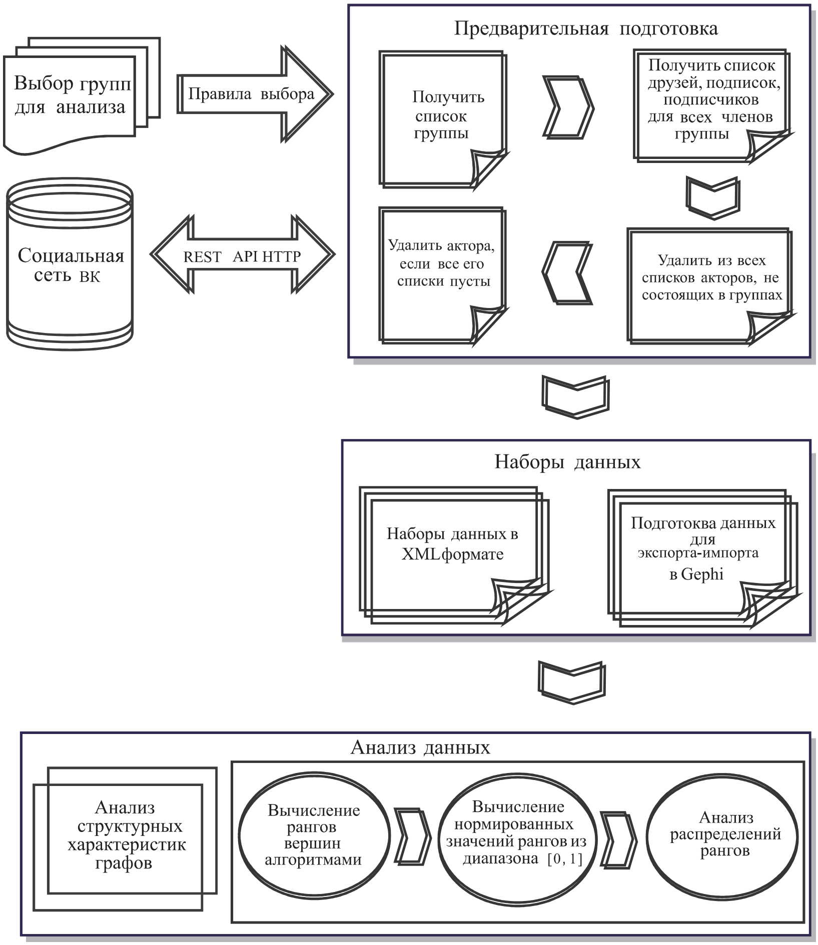


Рис. 1. Структура вычислительного эксперимента

Специально разработанное программное обеспечение было использовано для извлечения данных о пользователях групп, постах, комментариях и лайках к ним. Приложение вызывает методы HTTP интерфейса (REST HTTP API), предоставляемого социальной сетью. Разработанная программа позволяет получить данные обо всех интересующих нас интеракциях пользователей в рамках одной группы. Для преобразования данных из объектной модели в XML файлы и обратно используется свободно распространяемая библиотека JAXB. В качестве целевых выбирались группы численностью порядка 10-30 тысяч человек.

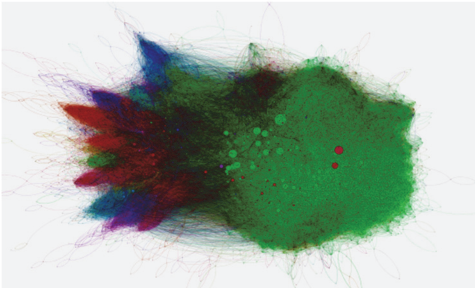
Далее приводятся результаты анализа графов, которые получены для конкретных групп с условными названиями «Волонтеры», «Шесть рукопожатий», «Письма добра», имеющие от 10 до 24 тысяч участников. Описание процедуры получения данных и некоторые структурные характеристики графов для соответствующих групп приведены в работе [14].

Последним шагом этапа предварительной подготовки данных является удаление всех изолированных вершин из полученных графов. Этот шаг позволяет «очистить» данные от вершин, которым могут быть присвоены минимальные ранги, так как они не связаны ни с одной вершиной графа. Такая процедура выглядит естественной и обоснованной.

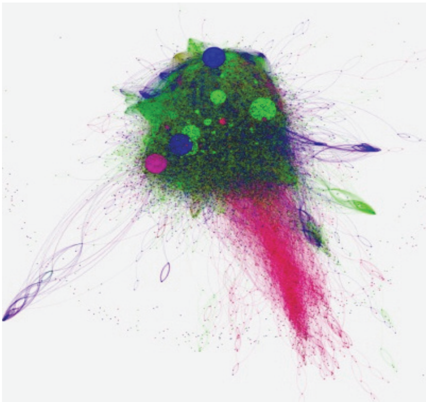
Второе замечание касается представленных далее визуальных образов графов. В случае, когда размеры графов исчисляются десятками тысяч вершин, их визуальное представление позволяет увидеть структуру графа и характерные для них структурные различия. Здесь нужно отметить, что различные алгоритмы укладки графов могут сформировать принципиально различные визуализации, которые подчеркивают разные особенности их структурных характеристик. Для получения визуализации был использован алгоритм ForceAtlas 2, реализованный в открытой платформе для визуализации графов Gephi (https://gephi.org/), который объединяет несколько теоретических подходов к укладке. Этот алгоритм основан на минимизации энергии (вершины притягиваются или отталкиваются друг от друга в зависимости от их взаимного расположения и наличия связей), которая приписывается графу в целом, его вершинам и ребрам. Алгоритм хорошо подходит для построения изображений, подчеркивающих структуру графа, а также для визуализации подмножеств вершин с высокой степенью взаимодействия. Визуальное представление графов приведено на рисунке 2. Для каждого графа также указывается число вершин и компонент связности. Эти величины лишний раз подчеркивают различия в структурных характеристиках графов. Графы на рисунке 2 визуализируют отношение «подписка» (односторонняя дуга) и «дружба» (двусторонняя дуга), то есть между двумя вершинами может быть либо однонаправленная стрелка, либо двунаправленная.

Графики на рисунке 3 дают представление о виде распределения полустепеней исхода и захода для вершин соответствующего графа, представляющего первый набор данных «Волонтеры». На рисунке 3(а) представлено распределение полустепеней захода, на рисунке 3(b) — распределение полустепеней исхода. Для других наборов данных графики выглядят подобным образом.

Волонтеры
N = 23 651
Компонент
связности: 77



Шесть
рукопожатий
N = 23 979
Компонент
связности 399



Письма добра
N =10464
Компонент
связности: 121

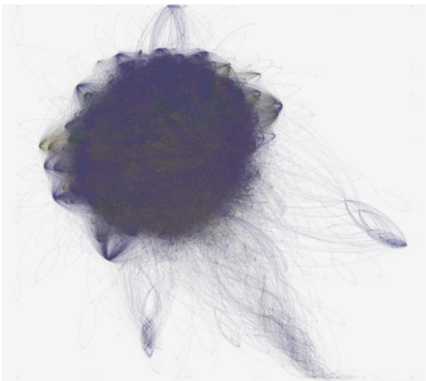
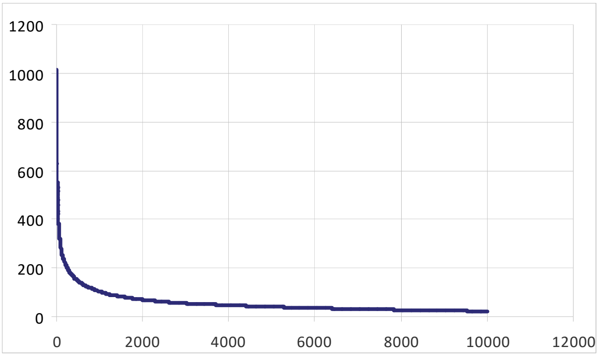
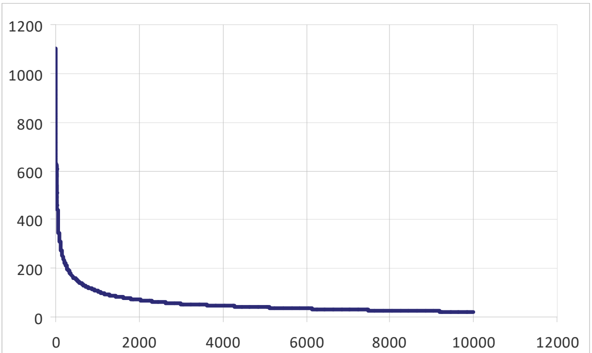


Рис. 2. Визуальное представление структуры трех групп с использованием алгоритма «ForceAtlas 2»



a)



b)

Рис. 3. Распределение полустепеней захода (a) и исхода (b) вершин графа «Волонтеры»

4.1. Анализируемые алгоритмы ранжирования. В качестве алгоритмов ранжирования рассматриваются три алгоритма, которые были использованы для построения иерархий в различных ситуациях при анализе структуры ссылочного пространства глобальной сети или структуры социальных сетей. В работе анализируется популярный и подробно описанный в литературе алгоритм ранжирования PageRank, локальный алгоритм Twitter, и глобальный алгоритм Freeman, осуществляющий рекуррентный пересчет рангов, но дающий похожие результаты. Краткое описание этих алгоритмов дано в работе [15].

В общем виде задача ранжирования, решаемая в рассматриваемых далее подходах, применяется к моделям, описывающим взаимодействие объектов глобальной сети, которые представляют акторов социальной сети либо страницы сети Интернет.

Имеющиеся данные представляют акторов социальной сети, для которых будут предложены формулы, справедливые для определения рангов вершин, исходя из весов дуг и модифицированные формулы, которые учитывают такие стандартные действия участников групп, как «Понравилось» («Лайки»).

Первый алгоритм построения иерархии ранжирования вершин ориентированного взвешенного графа использует только характеристики их локального окружения и с точки зрения нашей терминологии является локальным [16]. Оригинальный подход решает задачу вычисления формально определенной меры доверия между пользователями микроблога Twitter. На основании этой меры моделируется и предсказывается последовательность переходов при использовании сервиса. В этом случае степень доверия пропорциональна количеству взаимодействий между двумя акторами и определяется только параметрами связывающих их дуг, то есть является локальной характеристикой в принятых нами обозначениях. Расчет меры доверия актора *j* актору *i* осуществляется по формуле (2):

$$r_{ij} = \frac{wS_{ij} + (1-w)F_{ij}}{N_i}, \tag{2}$$

где $S_{ij}=1$, если актор *j* подписан на актора *i*, и $S_{ij}=0$ в противном случае; $F_{ij}=1$, если актор *i* подписан на актора *j*, и $F_{ij}=0$ в противном случае; N_i — общее количество подписок и подписчиков актора *i*; *w* — коэффициент степени влияния S_{ij} и F_{ij}.

При проведении вычислений в работе рекомендуется принять значение *w* в диапазоне от 0,5 до 1. При использовании значения $w = 0,5$ приравнивается степень влияния наличия собственных подписок и подписчиков на итоговый результат.

Для учета влияния «лайков» постов и комментариев необходимо дополнительно изменить формулу (2) расчета ранга. Для этого добавляются новые слагаемые в числитель формулы (2):

$$r_{ij} = \frac{wS_{ij} + (1-w)F_{ij} + cL_{ij} + (1-c)M_{ij}}{N_i}, \tag{3}$$

где *c* — коэффициент, показывающий степень влияния одного «лайка» на уровень доверия между двумя вершинами; L_{ij} — количество «лайков», которые актор *i* поставил под постами и комментариями *j*; M_{ij} — количество «лайков», которые актор *j* поставил под постами и комментариями *i*.

В рамках проведенного вычислительного эксперимента коэффициент c принят равным 0,25.

Для вычисления ранга актора i суммируются степени доверия ему всех акторов, которые с ним взаимодействуют. Таким образом, ранг пользователя будет рассчитываться по формуле:

$$r_i = \sum_{\forall j \in N_i} r_{i,j}. \tag{4}$$

Далее обозначаем алгоритм ранжирования вершин, основанный на формуле (4), как ***TW***.

В работе [17] предлагается способ итеративного построения иерархии ранжирования, основанный на вероятностных характеристиках взаимодействия в социальных сетях. В этом случае учитываются не только взаимодействие между актором сети и его локальным окружением, но и текущий усредненный уровень рангов вершин из этого окружения. То есть на значение ранга на каждой итерации алгоритма влияют и внешние по отношению к его окружению интеракции. Следующая формула (5) используется для подсчета ранга вершины i:

$$r_i = k\frac{F_i - S_i}{N_i} + Q_i, \tag{5}$$

где k — постоянный коэффициент, значение которого в цитируемой статье принято равным 2; F_i — количество подписчиков актора i; S_i — количество подписок актора i; N_i — общее количество подписок и подписчиков актора i (рассчитывается как количество уникальных акторов, которые являются подписчиками и тех, на кого подписан актор); Q_i — усредненный ранг всех акторов, с которыми взаимодействует актор i.

Итерации алгоритма продолжаются до тех пор, пока не будет достигнута нужная точность (максимальная разница между значениями ранга вершины на двух последовательных итерациях не будет превышать наперед заданного значения ε). Если алгоритм не сходится, и разница между значениями рангов в последовательных итерациях не стремится к 0, то необходимо ограничить количество итераций. Максимальное количество итераций в нашем алгоритме было установлено в значение 500, $\varepsilon = 0{,}1$. Формальное определение сходимости будет дано далее.

Для учета «лайков» комментариев и постов акторов добавляется дополнительное слагаемое в числителе формулы (5):

$$r_i = \frac{k(F_i - S_i) + \frac{k}{2}(M_i - L_i)}{N_i} + Q_i, \tag{6}$$

где L_i — количество лайков, которые поставил актор; M_i — количество лайков, которые были поставлены на записи и комментарии текущего актора; S_i , F_i — множество всех вершин, с которыми взаимосвязана вершина I.

Далее алгоритм ранжирования, основанный на формуле (6), обозначается как ***FR***.

Третий использованный в работе алгоритм ранжирования PageRank описан в работе [18]. Итеративный алгоритм подсчитывает ранг каждой вершины по следующей формуле:

$$r_i = c\sum \frac{r_j}{N_i} + cE_i, \tag{7}$$

где r_i — ранг i-ой вершины; c — коэффициент нормализации; F_i — множество вершин, ссылающихся на вершину i; S_i — множество вершин, на которые ссылается вершина i; $N_i = S_i \cup F_i$; E_i — некоторое первоначальное значение ранга вершины. При описании алгоритма утверждается, что оптимальными значением этого параметра является 0,15.

Далее алгоритм ранжирования, основанный на формуле (7), обозначается как ***PR***.

Алгоритм ***PR*** и его модификации подробно рассматриваются в литературе, проанализированы преимущества и недостатки его использования. Но особенное внимание в последнее время уделяется практике использования модификаций ***PageRank*** для моделей данных, имеющих большую размерность. Для этого случая разработаны специальные алгоритмы, использующие только локальные фрагменты всей модели и распределенную среду вычисления [19, 20]. В работе [21] рассматривается вычисление иерархии ранжирования в случае, когда граф может содержать «определенные» и «неопределенные» дуги. Авторами применяется традиционная для таких задач терминология, имеющая отношение к центральности. «Определенные» дуги являются традиционными для ориентированного графа с одной начальной и одной конечной вершинами. «Неопределенные» дуги имеют единственную вершину в качестве начальной, но множественные варианты конечной вершины, называемые «потенциальными» конечными. В качестве допустимого варианта потенциальной конечной рассматривается и вершина, которая отсутствует в заданном графе, но имеет специальное обозначение и добавлена в формальную математическую модель графа. Предложенный алгоритм ориентирован на динамический инкрементальный пересчет рангов при изменении графа в противовес статическому, когда вся совокупность рангов вычисляется заново при любом изменении.

Подсчет рангов вершин графов с использованием алгоритма ***PR*** производился с помощью программы Gephi 0.82 beta. Два других индекса (***TW*** и ***FR***) рассчитывались с использованием программного обеспечения собственной разработки.

4.2. Сходимость алгоритма ранжирования для заданного графа. Дадим определение понятия сходимости алгоритма ранжирования для заданного набора данных. Такое определение необходимо для корректного обсуждения временной сложности обсуждаемых алгоритмов, которая зависит не только от размерности исходных графов, но и от количества итераций, выполняемых алгоритмами. Последнее определяется либо наперёд заданным числом, либо выполнением некоторого условия, позволяющего прекратить вычисления.

Предполагаем, что алгоритм выполняет последовательное вычисление всей совокупности рангов вершин. Пусть $r^i=\left(r_1^i,r_2^i,...,r_n^i\right)$ — набор рангов для вершин заданного графа G, полученный на i-ой итерации алгоритма. В качестве меры расстояния D^i между двумя последовательными наборами r^i и r^{i+1}, рассматриваемыми как векторы, удобно использовать метрику Манхэттена, определенную формулой (8):

$$D^i=d(r^i,r^{i+1})=\sum_{j=1}^{n}\left|r_j^i-r_j^{i+1}\right|. \tag{8}$$

Более «мягким» вариантом метрики, которая предъявляет менее жесткие условия к процедуре вычисления рангов, является метрика Чебышева, определяемая формулой:

$$D^i=d(r^i,r^{i+1})=\max_{j=1,..,n}\left|r_j^i-r_j^{i+1}\right|. \tag{9}$$

Определение 2. Говорим, что алгоритм ранжирования сходится для заданного графа G, если:

$$\lim_{i \to \infty} D^i = 0. \tag{10}$$

Очевидно, что проблема сходимости (поточечной или равномерной) алгоритма ранжирования для заданного графа актуальна только для алгоритмов ***FR*** и ***PR***, так как они используют рекуррентную процедуру вычисления рангов. Алгоритм ***TW*** при вычислении рангов вершин использует только характеристики их локального окружения и не требует пересчета всей совокупности рангов при локальном изменении в графе.

4.3. Результаты вычислительного эксперимента. Диапазоны полученных рангов для трех алгоритмов, которые вычисляются по представленным выше формулам, для одних и тех же данных различаются. Например, при использовании алгоритма ***FR***, ранги могут становиться отрицательными из-за структуры формул (5), (6). По этой причине значения всех рангов были приведены к диапазону [0,1] преобразованием (11):

$$r_i' = \frac{r_i - \min\limits_{i=1,N}(r_i)}{\max\limits_{i=1,N}(r_i) - \min\limits_{i=1,N}(r_i)}, \tag{11}$$

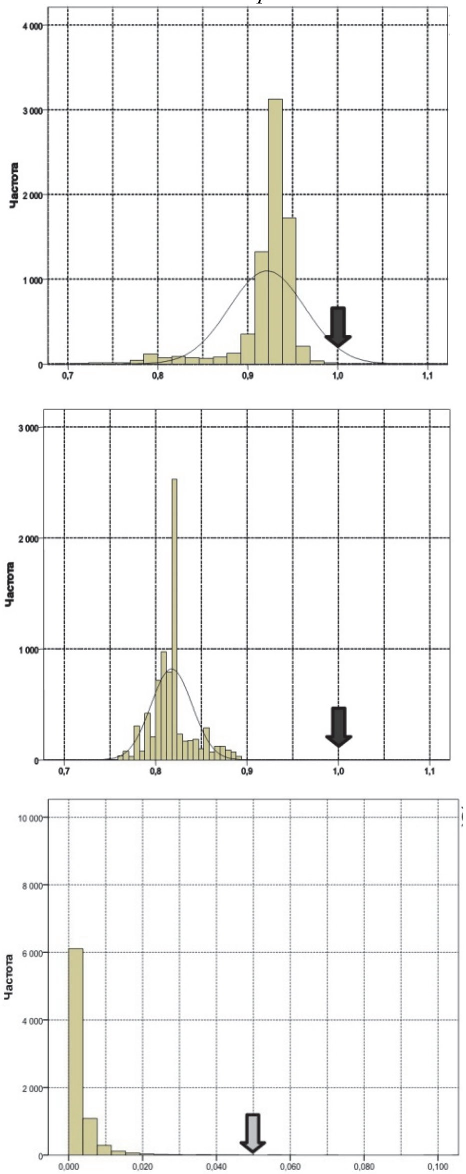
где максимум и минимум берутся по всему множеству рангов.

На рисунке 4 (a, b, c) показаны гистограммы распределений рангов вершин для графа, представляющего группу «Волонтеры».

Гистограммы сформированы для совокупности рангов, полученных алгоритмами ***FR***, ***TW*** и ***PR***. На графиках по оси абсцисс отложено нормализованное значение ранга вершин, на оси ординат — частота появления рангового диапазона. Для наборов данных «Шесть рукопожатий» и «Письма добра» получены похожие результаты, демонстрирующие такие же закономерности в распределении полученных рангов. Гистограммы распределений для этих графов показаны на рисунках 5 и 6 в уменьшенном масштабе, но позволяют подтвердить выявленные и описанные далее закономерности. Отдельно подчеркнем существенные различия в структурных характеристиках полученных графов (количество компонент связности, статистики степенных последовательностей и некоторые другие), описанных более подробно в работе [14].

Подобная задача решается в работе [22], в которой рассматриваются структурные характеристики виртуальных сообществ, представленных графами.

Волонтеры



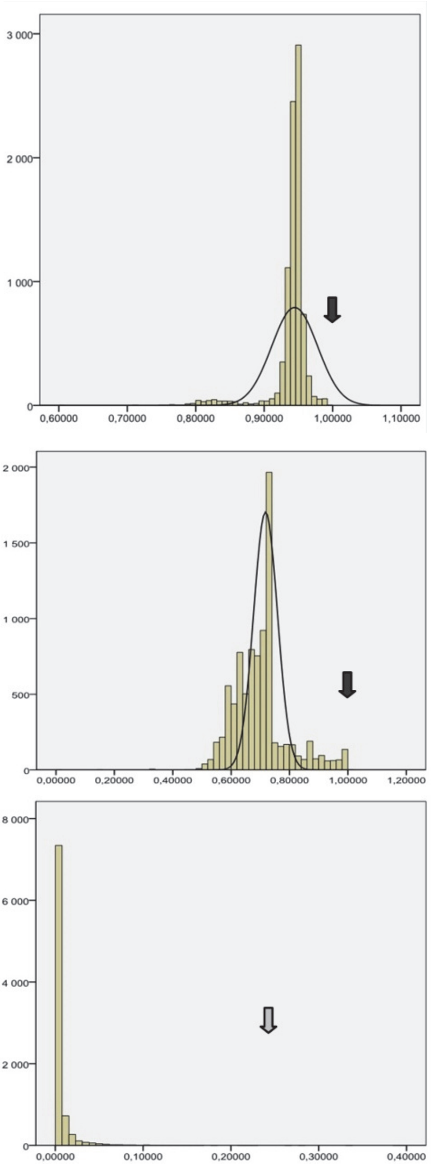
***FR** (a)*

***TW** (b)*

***PR** (c)*

Рис. 4. Гистограммы распределений рангов вершин графа «Волонтеры» для ***FR***(a), ***TW***(b), ***PR***(c). На a, b стрелка показывает положение 1, c — позиция значения 0,05

Письма Добра



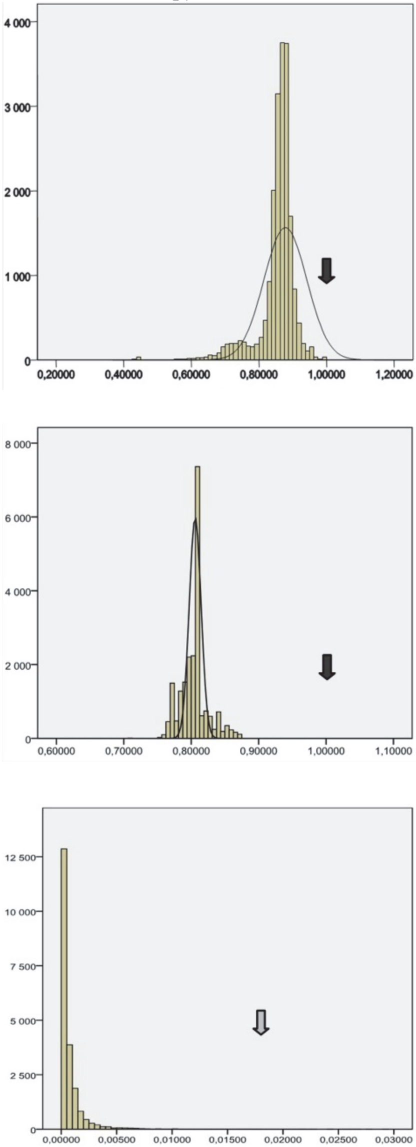
FR (a)

TW *(b)*

PR *(c)*

Рис. 5. Гистограммы распределений для наборов данных «Письма добра» при использовании алгоритмов ***FR***(a), ***TW***(b), ***PR***(c). На a, b стрелка показывает положение 1, c — позиция значения 0,25

Шесть рукопожатий



FR (a)

TW (b)

PR (c)

Рис. 6. Гистограммы распределений для наборов данных «Шесть рукопожатий» при использовании алгоритмов ***FR***(a), ***TW***(b), ***PR***(c). На a, b стрелка показывает положение 1, c — позиция значения 0,0175

Распределение полученных рангов подчиняются сходным закономерностям, которые могут быть описаны следующими пунктами.

1. В описанных в разделе 3 публикациях предложены решения оптимизационных задач, которые используют предварительное решение задачи ранжирования вершин. Этот подход позволяет повысить эффективность алгоритмов, улучшить их временные оценки.

2. Имеется существенное отличие распределений рангов, генерируемых различными алгоритмами ранжирования. Если алгоритмы ***FR*** и ***TW*** в качестве результата строят ранги, имеющие распределение близкое к нормальному, то ранги, полученные алгоритмом ***PR***, имеют распределение близкое к распределению, которое может быть описано как экспоненциальное (см. рисунки 4-6) или степенное с определенным набором взаимосвязанных параметров. Информация о типе распределения позволяет корректно выбирать параметрические тесты для анализа совокупности рангов.

3. Алгоритм ***FR*** дает совокупность рангов, распределение которых после нормирования смещено в сторону максимального значения 1, алгоритм же ***TW*** даёт меньшие значения распределения с меньшим разбросом значений. Этот вывод подтверждается данными таблица 1, которая позволяет сравнить средние значения, стандартное отклонение и скошенность распределений полученных рангов. Это различие можно объяснить тем, что алгоритм ***FR*** увеличивает ранги за счет дополнительного слагаемого Q_i в формулах 5, 6, что приводит к сдвигу центра распределения в сторону максимального значения 1.

Таблица 1. Статистические параметры распределения рангов

		FR	***TW***	***PR***
Волонтеры	Среднее	0,921	0,818	0,004
	Медиана	0,930	0,818	0,002
	Ст. отклонение	0,039	0,022	0,014
	Скошенность	-3,783	0,905	-
Шесть рукопожатий	Среднее	0,854	0,804	0,001
	Медиана	0,866	0,808	0,000
	Ст. отклонение	0,057	0,021	0,002
	Скошенность	-2,697	0,553	-
Письма добра	Среднее	0,938	0,703	0,006
	Медиана	0,946	0,702	0,007
	Ст. отклонение	0,050	0,089	0,015
	Скошенность	-13,552	1,051	-

4. Алгоритмы ***FR*** и ***PR*** используют рекуррентную схему вычисления следующего приближения рангов, что актуализирует проблему сходимости алгоритма для заданного набора данных. Фактически, в вычислительном эксперименте условием остановки процедуры вычисления рангов стало достижение наперед заданного максимального числа итераций. Алгоритм ***TW*** при вычислении ранга вершины использует только параметры ее локального окружения и при динамическом изменении графа (добавление/удаление дуги, изменение ее атрибутов) потребует только пересчета рангов двух вершин, которые соединены этой дугой. Все остальные ранги остаются неизменными. Это обстоятельство позволяет использовать эффективные инкрементальные/декрементальные процедуры вычисления рангов, сложность которых зависит только от размера локальных окружений вершин и сложности процедуры коррекции глобальной совокупности рангов.

5. Заключение. В работе на примере графов большой размерности проанализированы результаты применения алгоритмов ранжирования вершин. Показано, что процедура ранжирования имеет большое прикладное значение для графов большой размерности, представляющих отношения в социальных сетях. Для анализа применимости результатов в решении оптимизационных задач самой разнообразной природы необходимы дополнительные исследования, учитывающие особенности предметной области, используемых дискретных моделей.

Проанализированные алгоритмы ранжирования имеют различные вычислительную сложность и распределения полученных рангов. Если при решении прикладной задачи к распределению рангов предъявляются какие-либо требования (например, нормальность распределения является обязательным при использовании некоторых параметрических тестов), необходимо предварительно провести анализ результатов работы алгоритма, определить статистические характеристики полученного распределения рангов.

При использовании предварительного ранжирования вершин графа в случае большой размерности модели исходных данных важным моментом является возможность использования инкрементальных/декрементальных алгоритмов, которые минимизируют сложность обработки данных, как правило, учитывают при построении глобального решения только локальные изменения, не требуют пересчета всех рангов. В этом смысле алгоритм ***TW***, вычисляющий ранги только по данным локальной окрестности вершин, выгодно отличается от алгоритмов ***FR*** и ***PR***, которые требуют пересчета всей совокупности рангов даже при локальном изменении графа. Это обстоятельство не позволяет использовать последние два алгоритма в случае, когда размерность модели является большой, необходимо быстро получить решение, не пересчитывая всю совокупность рангов. Распределенная модель вычислений смягчает это требование, но не устраняет его принципиально.

В качестве дискуссионного момента можно выделить вопрос о свойствах графа, которые гарантируют сходимость алгоритма ранжирования. Алгоритмы ***FR*** и ***PR*** в проведенном вычислительном эксперименте качестве решения выдавали ранги после предельного, наперед заданного числа итераций. В статье предложено определение сходимости алгоритма ранжирования, оставляя за скобками интересный теоретический вопрос о критериях сходимости, определении нетривиальных структурных характеристик графов, наличие которых гарантирует достижение заданной точности вычисления рангов при выбранной мере расстояния, существование предела рекуррентной последовательности.

Этот же вопрос возникает и по поводу структуры формул, используемых для последовательного вычисления рангов. Ясно, что если второе слагаемое в Q_i в формулах (5) и (6) для алгоритма ***FR***, будет монотонно убывающей последовательностью в процессе итераций, то соответствующий алгоритм будет сходиться. Регулируя скорость уменьшения значений Q_i можно получить управляемую скорость сходимости всего процесса для графов любой структуры.

Литература

1. *Кочетов Ю.А., Хмелев А.В.* Гибридный алгоритм локального поиска для задачи маршрутизации разнородного ограниченного автопарка // Дискретный анализ и исследование операций. 2015. Т. 22. № 5. С. 5–29.
2. *Кривошеин Д.Ю., Марченко А.М.* Алгоритмы пересчёта кратчайших путей в графе при изменении весов ребер // Проблемы разработки перспективных микро– и наноэлектронных систем. 2012. № 1. С. 263–266.
3. *Demetrescu C., Finocchi I., Italiano G.F.* Dynamic graphs. URL: www.diku.dk/PATH05/CRC-book1.pdf (дата обращения: 01.02.17).
4. *Pearce D.J., Kelly P.H.J.* A dynamic topological sort algorithm for directed acyclic graphs // Journal of Experimental Algorithmics (JEA). 2007. vol. 11. pp. 1–7.
5. *Deepak A., Tobias F.* Average-Case Analysis of Incremental Topological Ordering // Discrete Applied Mathematics. 2010. vol. 158. no. 4. pp. 240–250.
6. *Nicoara D., Kamali S., Daudjee K., Chen L.* Hermes: Dynamic Partitioning for Distributed Social Network Graph Databases // EDBT. 2015. pp. 25–36.
7. *Ammar A.B.* Query optimization techniques in graph Databases // International Journal of Database Management Systems (IJDMS). 2016. vol. 8. no. 4. 14 p.
8. *Курейчик В.В., Жиленков М.А.* Муравьиный алгоритм для решения оптимизационных задач с явно выраженной целевой функцией // Информатика, вычислительная техника и инженерное образование. 2015. № 2. С. 1–12.

9. *Agarw S.* Ranking on Graph Data // Proceedings of the 23rd International Conference on Machine Learning. 2006. pp. 25–32.
10. *Розенберг И.Н.* Использование нечетких представлений данных при определении медиан графа // Известия Южного федерального университета. Технические науки. 2001. № 4. С. 64–72.
11. *Adaikalam A., Manikandan S., Rajamani V.* Fuzzy graph based shortest path ranking method for optical network // Optical and Quantum Electronics. 2017. vol. 49. no. 9. 296 p.
12. *Barman A., Shah S.K.* SHaPE: A Novel Graph Theoretic Algorithm for Making Consensus-Based Decisions in Person Re-identification Systems // IEEE International Conference on Computer Vision (ICCV). 2017. pp. 1124–1133.
13. *Roffo G., Melzi S., Castellani U., Vinciarelli A.* Infinite Latent Feature Selection: A Probabilistic Latent Graph-Based Ranking Approach // IEEE International Conference on Computer Vision (ICCV). 2017. pp. 1407–1415.
14. *Печенкин В.В., Решетников Д.С., Ярская-Смирнова В.Н.* Визуализация сетевой структуры групповых отношений // 4М. Методология, методы, математическое моделирование. 2014. № 39. С. 40–61.
15. *Королёв М.С., Решетников Д.С.* Подходы к задаче ранжирования вершин в теории графов // Проблемы управления в социально-экономических и технических системах. 2017. С. 138–141.
16. *Lumbreras A., Gavaldà R.* Applying trust metrics based on user interactions to recommendation in social networks // IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2012. pp. 1159–1164.
17. *Jameson K.A., Appleby M.C., Freeman L.C.* Finding an appropriate order for a hierarchy based on probabilistic dominance // Animal Behaviour. 1999. vol. 57. no. 5. pp. 991–998.
18. Easy visualizations of PageRank and Page Groups with Gephi. URL: https://searchengineland.com/easy-visualizations-pagerank-page-groups-gephi-265716 (дата обращения: 07.02.2017).
19. *Sarma A.D., Molla A.R., Pandurangan G., Upfal E.* Fast distributed pagerank computation // Theoretical Computer Science. 2015. vol. 561. pp. 113–121.
20. *Dai L., Freris N.M.* Fully distributed PageRank computation with exponential convergence // 2017. arXiv preprint arXiv:1705.09927. 5 p.
21. *Nathan E., Fairbanks J., Bader D.* Ranking in Dynamic Graphs using Exponential Centrality // International Workshop on Complex Networks and their Applications. 2017. pp. 378–389.
22. *Рыков Ю.Г., Кольцова О.Ю., Мейлахс П.А.* Структура и функции онлайн-сообществ: сетевая картография ВИЧ-релевантных групп в социальной сети «ВКонтакте» // Социологические исследования. 2016. № 8. С. 30–42.

Печенкин Виталий Владимирович — д-р социол. наук, профессор, профессор кафедры прикладных информационных технологий института прикладных информационных технологий и коммуникаций, Саратовский государственный технический университет имени Гагарина Ю.А.. Область научных интересов: оптимизационные задачи на графах и сетях, визуализация графов, методы статистической обработки многомерных данных, многомерное шкалирование, анализ социальных сетей, использование математических и компьютерных подходов в социологических исследованиях. Число научных публикаций — 110. pechenkinvv@mail.ru; Политехническая, 77, Саратов, 41005; р.т.: +7(8452)99-87-15.

Королёв Михаил Сергеевич — аспирант кафедры прикладных информационных технологий института прикладных информационных технологий и коммуникаций, Саратовский государственный технический университет имени Гагарина Ю.А., ассистент кафедры кафедры прикладных информационных технологий института прикладных информационных технологий и коммуникаций, Саратовский государственный технический университет имени Гагарина Ю.А.. Область научных интересов: алгоритмы и методы оптимизации, виртуальная реальность, дополненная реальность, 3D-проектирование, 3D-моделирвоание, архитектурная визуализация, компьютерное зрение, компьютерный дизайн.. Число научных публикаций — 8. koroliow.mikhail@yandex.ru; Политехническая, 77, Саратов, 41005; р.т.: +7(8452)99-87-15.

Димитров Любомир Ванков — д-р физ.-мат. наук, профессор, проректор по учебной деятельности и аккредитации, Технический университет - София, профессор машиностроительного факультета, Технический университет - София. Область научных интересов: мехатроника, автоматика, микроэлектронные модули и системы и их применение (MEMS). Число научных публикаций — 230. lubomir_dimitrov@tu-sofia.bg; бул. Св. Климент Охридски, 8, София, 1756, Болгария; р.т.: +359 2 965-25 60.

DOI 10.15622/SP.61.4

V.V. PECHENKIN, M.S. KOROLEV, L.D. DIMITROV

APPLIED ASPECTS OF RANKING ALGORITHMS FOR ORIENTED WEIGHTED GRAPHS (ON THE EXAMPLE OF SOCIAL NETWORK GRAPHS)

Pechenkin V.V., Korolev M.S., Domotrov L.D. **Applied Aspects of Ranking Algorithms for Oriented Weighted Graphs (on the Example of Social Network Graphs).**

Abstract. The article deals with the applied aspects of the preliminary vertices ranking for oriented weighted graph. In this paper, the authors observed the widespread use of this technique in developing heuristic discrete optimization algorithms. The ranking problem is directly related to the problem of social networks centrality and large real world data sets, but as shown in the article ranking is explicitly or implicitly used in the development of algorithms as the initial stage of obtaining a solution for solving applied problems. Examples of such ranking application are given. The examples demonstrate the increase of efficiency in solving some optimization applied problems, which are widely used in mathematical methods of optimization, decision-making not only from the theoretical development point of view but also their applications. The article describes the structure of the first phase of the computational experiment, which is associated with the procedure of obtaining test data sets. The obtained data are presented by weighted graphs that correspond to several groups of the social network Vkontakte with the number of participants in the range from 9000 to 24 thousand. It is shown that the structural characteristics of the obtained graphs differ significantly in the number of connectivity components. Characteristics of centrality (degree's sequences), as shown, have exponential distribution. The main attention is given to the analysis of three approaches to graph vertices ranking. We propose analysis and comparison of the obtained set of ranks by the nature of their distribution. The definition of convergence for graph vertex ranking algorithms is introduced and the differences of their use in considering the data of large dimension and the need to build a solution in the presence of local changes are discussed.

Keywords: ranking, oriented graph, weighted graph, incremental algorithm, local algorithm.

Pechenkin Vitaly Vladimirovich — Ph.D., Dr. Sci., professor, professor of applied information technologies department of school of applied information technology and communication, Yuri Gagarin State Technical University of Saratov (SSTU). Research interests: optimization problems on graphs and networks, graph visualization, statistical processing of multidimensional data, multidimensional scaling, social networks analysis, application of mathematics and computer science in sociological studies. The number of publications — 110. pechenkinvv@mail.ru; 77, Politechnicheskaya, Saratov, 410054, Russia; office phone: +7(8452)99-87-15.

Korolev Mikhail Sergeevich — Ph.D. student of applied information technologies department of school of applied information technology and communication, Yuri Gagarin State Technical University of Saratov (SSTU), assistant of applied information technologies department of school of applied information technology and communication, Yuri Gagarin State Technical University of Saratov (SSTU). Research interests: optimization algorithms and methods, virtual reality, augmented reality, 3d-technology, computer vision, computer design.. The number of publications — 8. koroliow.mikhail@yandex.ru; 77, Politechnicheskaya, Saratov, 410054, Russia; office phone: +7(8452)99-87-15.

Dimitrov Lyubomir Vankov — Ph.D., Dr. Sci., vice-rector of learning activity and accreditation, Technical University of Sofia, professor of engineering faculty, Technical

University of Sofia. Research interests: mechatronics, automation, microelectronic modules and systems and their application(MEMS). The number of publications — 230. lubomir_dimitrov@tu-sofia.bg; 8, Kliment Orhidski Boulevard, Sofia, 1756, Bulgaria; office phone: +359 2 965-25 60.

References

1. Kochetov Y.A., Khmelev V.A. [A Hybrid algorithm of local search for routing problem with heterogeneous fleet limited]. *Diskretnyj analiz i issledovanie operatsij – Journal of Applied and Industrial Mathematics.* 2015. Issue 22. vol. 5. pp. 5–29. (In Russ.).
2. Krivoshein D.Yu., Marchenko A.m. [Algorithms for recalculating shortest paths in a graph when changing edge weights]. *Problemy razrabotki perspektivnykh mikro- i nanoehlektronnykh sistem* [Problems of perspective micro- and nanoelectronic systems development]. 2012. vol. 1. pp. 263–266. (In Russ.).
3. Demetrescu C., Finocchi I., Italiano G.F. Dynamic graphs. Available at: www.diku.dk/PATH05/CRC-book1.pdf (accessed: 01.02.17).
4. Pearce D.J., Kelly P.H.J. A Dynamic Topological Sort Algorithm for Directed Acyclic Graphs. *Journal of Experimental Algorithmics (JEA)*. 2007. vol. 11. pp. 1–7.
5. Deepak A., Tobias F. Average-Case Analysis of Incremental Topological Ordering. *Discrete Applied Mathematics.* 2010. vol. 158. no. 4. pp. 240–250.
6. Nicoara D., Kamali S., Daudjee K., Chen L. Hermes: Dynamic Partitioning for Distributed Social Network Graph Databases. EDBT. 2015. pp. 25–36.
7. Ammar A.B. Query optimization techniques in graph Databases. *International Journal of Database Management Systems (IJDMS).* 2016. vol. 8. no. 4. 14 p.
8. Kureychik V.V., Zhilenkov M.A. [Ant algorithm for solving optimization problems with explicit objective function information, computing and engineering education]. *Informatika, vychislitel'naya tekhnika i inzhenernoe obrazovanie – Computer science, computer engineering and engineering education*. 2015. vol. 2. pp. 1–12. (In Russ.).
9. Agarw S. Ranking on Graph Data. Proceedings of the 23rd International Conference on Machine Learning. 2006. pp. 25–32.
10. Rosenberg I.N. [The use of fuzzy representations of data when determining the medians of the graph]. *Izvestiya YUFU. Tekhnicheskie nauki – Izvestiya SFedU. Engineering sciences.* 2001. vol. 4. pp. 64–72. (In Russ.).
11. Adaikalam A., Manikandan S., Rajamani V. Fuzzy graph based shortest path ranking method for optical network. *Optical and Quantum Electronics*. 2017. vol. 49. no. 9. 296 p.
12. Barman A., Shah S.K. SHaPE: A Novel Graph Theoretic Algorithm for Making Consensus-Based Decisions in Person Re-identification Systems. IEEE International Conference on Computer Vision (ICCV). 2017. pp. 1124–1133.
13. Roffo G., Melzi S., Castellani U., Vinciarelli A. Infinite Latent Feature Selection: A Probabilistic Latent Graph-Based Ranking Approach. IEEE International Conference on Computer Vision (ICCV). 2017. pp. 1407–1415.
14. Pechenkin V.V., Reshetnikov D.S., Yarskaya-Smirnova V.N. [Visualization of the network structure of group relations] *4M. Metodologiya, metody, matematicheskoe modelirovanie – 4M. Methodology, methods, mathematical modeling.* 2014. vol. 39. pp. 40–61. (In Russ.).
15. Korolev M.S., Reshetnikov D.S. [Approaches to the problem of ranking vertices in graph theory]. *Problemy upravleniya v sotsial'no-ehkonomicheskikh i tekhnicheskikh sistemakh* [Problems of control in socio-economic and technical systems]. 2017. pp. 138–141. (In Russ.).
16. Lumbreras A., Gavaldà R. Applying trust metrics based on user interactions to recommendation in social networks. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2012. pp. 1159–1164.

17. Jameson K.A., Appleby M.C., Freeman L.C. Finding an appropriate order for a hierarchy based on probabilistic dominance. *Animal Behaviour.* 1999. vol. 57. no. 5. pp. 991–998.
18. Easy visualizations of PageRank and Page Groups with Gephi. Available at: https://searchengineland.com/easy-visualizations-pagerank-page-groups-gephi-265716 (accessed: 07.02.2017).
19. Sarma A.D., Molla A.R., Pandurangan G., Upfal E. Fast distributed pagerank computation. *Theoretical Computer Science.* 2015. vol. 561. pp. 113–121.
20. Dai L., Freris N.M. Fully distributed PageRank computation with exponential convergence. 2017. arXiv preprint arXiv:1705.09927. 5 p.
21. Nathan E., Fairbanks J., Bader D. Ranking in Dynamic Graphs using Exponential Centrality. International Workshop on Complex Networks and their Applications. 2017. pp. 378–389.
22. Rykov YU.G., Kol'tsova O.YU., Mejlakhs P.A. [Structure and Functions of Online Communities Network Mapping of HIV-relevant groups in VK]. *Sociological Studies – Sotsiologicheskie issledovaniya*. 2016. vol. 8. pp. 30–42.