МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

МЕТОДЫ РАНЖИРОВАНИЯ В СЛОЖНЫХ СЕТЯХ И ИХ ПРИЛОЖЕНИЯ В РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМАХ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студентки 2 курса 248 группы
направления 09.04.03 — Прикладная информатика

механико-математического факультета
Шевченко Алины Игоревны

Научный руководитель
доцент, к. ф.-м. н., доцент _____ В. Р. Шебалдин

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2026

ВВЕДЕНИЕ

Актуальность темы. В последнее время в различных областях науки и практики наблюдается быстрый рост объёмов данных, требующих анализа для принятия управленческих, исследовательских и технических решений. Накопление больших объёмов слабоструктурированных и связанных данных приводит к необходимости поиска новых методов их обработки, поскольку классические статистические и аналитические подходы не всегда применимы для структур, характеризующихся высокой размерностью и сложными связями. Естественной моделью представления взаимосвязанных данных выступают графовые структуры, в которых вершины соответствуют объектам, а рёбра задают отношения или связи между ними.

К числу фундаментальных задач в данной области относится проблема ранжирования вершин графа, направленная на определение их относительной значимости в соответствии с заданными критериями. Алгоритмы ранжирования нашли применение в самых разных сферах: в поисковых системах — для оценки важности веб-страниц, в социальных сетях — для выявления влиятельных пользователей, в биоинформатике — для идентификации ключевых белков, в наукометрии — для оценки вклада авторов и научных статей. Особую актуальность методы ранжирования приобрели в контексте рекомендательных систем: современные сервисы — музыкальные стриминговые платформы, социальные сети, онлайн-магазины — используют графовые модели для предсказания предпочтений пользователей. Это подчёркивает актуальность и перспективность развития методов графового ранжирования.

Целью магистерской работы является исследование методов ранжирования вершин графа, демонстрация их практической эффективности на задаче определения наиболее значимых объектов в музыкальных сетевых данных и анализ их приложения в рекомендательных системах.

Объектом исследования является задача ранжирования вершин графовых моделей высокой размерности. Исследование фокусируется на разработке и применении алгоритмов, способных анализировать структуру связей между объектами и определять их относительную значимость.

Предметом исследования является процесс применения алгорит-

мов графового ранжирования (PageRank, HITS, Katz centrality, Personalized PageRank, TrustRank и др.) к задачам построения рекомендательных систем на больших наборах пользовательских данных.

Для достижения цели были поставлены следующие задачи:

— изучить основные понятия и определения теории графов, рассмотреть математическую постановку задачи ранжирования вершин;
— исследовать существующие алгоритмы ранжирования вершин графа (PageRank, HITS, Katz centrality, TrustRank, Personalized PageRank), а также классические меры центральности;
— провести сравнительный анализ методов с точки зрения вычислительной сложности, устойчивости и применимости к графам большой размерности;
— проанализировать принципы работы рекомендательных систем и способы интеграции в них графовых методов;
— реализовать рассмотренные алгоритмы и применить их к данным открытого музыкального датасета Yandex Yambda;
— разработать гибридную рекомендательную систему на основе графовых методов ранжирования и провести экспериментальную оценку её качества.

Практическая значимость работы заключается в применимости полученных результатов к построению современных рекомендательных систем. Разработанные программные решения и вычисленные весовые коэффициенты могут быть использованы как самостоятельная рекомендательная система или как источник дополнительных признаков для обучаемых ML-моделей ранжирования в составе промышленных рекомендательных пайплайнов. Графовые методы ранжирования позволяют решать проблему разреженности данных, характерную для матричных методов, и обеспечивать качественные рекомендации даже для редких или новых объектов, если они связаны с популярными сущностями через граф знаний.

Структура и содержание работы. Работа состоит из введения, 5 разделов, заключения и списка использованных источников, содержащего 20 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении раскрывается актуальность темы работы, формулируется цель работы и задачи, которые необходимо решить, отмечается практическая значимость результатов.

В первом разделе приводятся теоретические основы анализа графов. Граф $G(V, E)$ определяется как совокупность непустого множества вершин V и множества рёбер $E \subseteq V \times V$. Рассматриваются основные способы задания графа: аналитический, геометрический, матрица смежности $A = (a_{ij})$ и матрица инцидентности $B = (b_{ik})$. Выделяются ориентированные и неориентированные графы; для каждой вершины вводится характеристика степени, а для ориентированных графов — входящей и исходящей степени.

Рассмотрены важнейшие частные виды графов: тривиальные, пустые, полные (K_n с числом рёбер $n(n-1)/2$), регулярные степени k, двудольные и полные двудольные ($K_{m,n}$), циклы, деревья, подграфы, планарные графы.

Особое внимание уделено математической постановке задачи ранжирования вершин. Формально цель ранжирования формулируется как поиск функции $f : V \to \mathbb{R}$, ассоциирующей каждой вершине $v \in V$ число $f(v)$, отражающее её значимость. Эквивалентно вводится вектор $x \in \mathbb{R}^n$, где $x_i = f(v_i)$. В литературе встречаются несколько формальных вариантов постановки задачи ранжирования: собственная задача $Ax = \lambda x$, поиск стационарного распределения Маркова $\pi P = \pi$, итерационная формула $x^{(t+1)} = F(A, x^{(t)})$, оптимизационная постановка вида $\max_{\|x\|=1} x^\top Ax$. Функция $f(v)$ должна обладать инвариантностью к изоморфизму, устойчивостью, масштабируемостью и масштабной инвариантностью.

В разделе подробно описаны классические меры центральности: степенная центральность $C_D(v) = d(v)/(n-1)$, центральность по близости $C_C(v) = 1/\sum_u \mathrm{dist}(v, u)$, центральность по посредничеству $C_B(v) = \sum_{s \neq v \neq t} \sigma_{st}(v)/\sigma_{st}$, центральность по собственному вектору, определяемая как главный собственный вектор матрицы смежности. Для каждой меры приведены математическая формулировка, обсуждение свойств, преимуществ и ограничений, а также вычислительная сложность. Показано, что выбор меры зависит от характера приложения и структуры графа.

Во втором разделе рассматриваются конкретные алгоритмы ранжи-

рования вершин графа. Описывается эволюция методов ранжирования — от локальных мер на основе свойств вершин до спектральных методов, мер с учётом многошаговой структуры графа (Katz centrality), стохастических методов на основе моделей случайного блуждания (PageRank), двупараметрических спектральных подходов (HITS), а также современных методов на основе обучения представлений и графовых нейронных сетей.

Алгоритм PageRank представляет собой математическую модель ранжирования вершин ориентированного графа, основанную на идее случайного блуждания. Вектор рангов $r \in \mathbb{R}^n$ определяется как решение уравнения фиксированной точки

$$r = (1-d)\frac{1}{n}\mathbf{1} + dPr,$$

где $P = D_{\text{out}}^{-1}A$ — матрица переходов, $d \in (0,1)$ — дампинг-фактор. Рассматриваются практические уточнения: обработка висячих вершин (dangling nodes), персонализированная версия с произвольным вектором телепортации v вместо равномерного распределения. При $0 < d < 1$ и корректно скорректированной матрице переходов гарантируется существование единственного неотрицательного стационарного вектора и сходимость итераций к нему.

Алгоритм HITS (Hyperlink-Induced Topic Search), предложенный Дж. Клейнбергом, рассматривает каждую вершину в двух ролях: как авторитет (authority) и как хаб (hub). Авторитет — вершина, на которую указывают многие хабы; хаб — вершина, ссылающаяся на многие авторитеты. Векторы авторитетов a и хабов h удовлетворяют системе $a = A^\top h$, $h = Aa$. Подстановкой получаем $a = A^\top Aa$, $h = AA^\top h$, откуда следует, что a — собственный вектор матрицы $A^\top A$, а h — собственный вектор матрицы $AA^\top$. По теореме Фробениуса–Перрона итерационный процесс сходится к главным собственным векторам.

Другие методы ранжирования. Центральность Катца

$$C_K(i) = \sum_{k=1}^{\infty} \alpha^k (A^k \mathbf{1})_i = \left((I - \alpha A)^{-1} - I\right)\mathbf{1}$$

учитывает все возможные пути с экспоненциальным затуханием по длине. Алгоритм TrustRank — модификация PageRank, в которой телепортация производится не равномерно, а на множество заранее отобранных «доверенных»

вершин; используется для борьбы со спамом и манипуляциями. Персонализированный PageRank (PPR) обобщает классический PageRank: вектор телепортации задаётся индивидуально для каждого пользователя, что позволяет получать персонализированные результаты ранжирования.

Проведён сравнительный анализ алгоритмов по критериям вычислительной сложности, масштабируемости и применимости. Установлено, что PageRank, TrustRank и Personalized PageRank — лучший выбор для больших разреженных графов благодаря сложности $O(k \cdot |E|)$ и быстрой сходимости. Алгоритмы HITS и Katz centrality хорошо подходят для анализа небольших или тематических подграфов, а меры на основе кратчайших путей (closeness, betweenness) требуют квадратичной сложности и применимы лишь к подграфам.

В третьем разделе обсуждается применение методов ранжирования в рекомендательных системах. Традиционные подходы делятся на три категории: коллаборативная фильтрация (анализ истории взаимодействий множества пользователей), контентный анализ (использование характеристик объектов) и гибридные подходы. Графовые модели позволяют объединить эти подходы в единую структуру, где пользователи и объекты представлены вершинами, а взаимодействия — рёбрами.

Рассмотрены варианты графовых рекомендательных систем: двудольные графы «пользователь–объект», случайные блуждания и Personalized PageRank, графы знаний с семантической информацией о жанрах, авторах, продюсерах. Приведены примеры использования графовых подходов ведущими стриминговыми платформами: Spotify (анализ совместного появления треков в плейлистах для «Discover Weekly»), Яндекс Музыка (коллаборативная фильтрация на графах последовательностей прослушиваний), ВКонтакте (алгоритмы на основе случайных блужданий с учётом социальных связей).

В четвёртом разделе приведена практическая часть работы. В качестве экспериментальной базы выбран открытый музыкальный датасет Yandex Yambda (YAndex Music Billion-interactions DAtaset), содержащий действия 1 миллиона пользователей на 9,39 миллиона треков за 10 месяцев, в общей сложности около 4,79 миллиарда событий. Все данные анонимизированы. Использовалась сокращённая версия Yambda-50M, содержащая 46 467 212 собы-

тий 10 000 пользователей.

Проведён разведочный анализ данных (EDA), включающий первичную оценку структуры, анализ полноты, статистический анализ атрибутов и исследование распределений. Выявлено доминирование событий типа `listen` (97,23%), бимодальный характер процента прослушивания (медиана 100%) и экстремально высокий эксцесс распределения длительности треков ($\approx 181{,}65$), что указывает на «тяжёлые хвосты» в характеристиках контента.

Построена графовая модель музыкальных взаимодействий в виде двудольного графа «пользователь–трек» с 9 238 пользователями и 877 168 треками (плотность связей 0,16%) и однородного графа переходов между треками с 877 168 вершинами и 11 232 567 рёбрами. Для эффективной работы с большими объёмами данных использовались категориальные типы данных Pandas и разреженные матричные представления.

Реализованы и экспериментально исследованы семь методов ранжирования:

1. *Алгоритм HITS* применён к бинарной матрице смежности размером 9238×877168. Сходимость достигнута за 19 итераций за время $\approx 5{,}68$ сек. Выявлены экспертные пользователи-хабы (топ-1: U398400) и треки-авторитеты (топ-1: T6901374). Проведены: эксперимент по масштабируемости (линейный рост времени работы от объёма графа), сравнение с базовым ранжированием по in-degree (совпадение 14 из 15 в топ-15 при содержательно различающемся порядке), бутстрэп-анализ устойчивости (среднее пересечение топ-10 с эталоном — 9,6 из 10), оценка эффекта холодного старта (hub-score новых пользователей в 671 раз ниже, чем у экспертов).
2. *Алгоритм PageRank* применён к графу переходов между треками с дампинг-фактором $\alpha = 0{,}85$ и порогом сходимости $\varepsilon = 10^{-6}$. Сходимость достигнута за 40 итераций за время $\approx 20{,}46$ сек. Лидер ранжирования — трек T3542184 со значением PageRank $5{,}09 \cdot 10^{-4}$. Показано, что PageRank оценивает не только количество входящих переходов, но и их качество: трек с меньшим числом переходов может иметь более высокий PageRank, если переходы поступают со значимых источников.
3. Дополнительно реализованы *Katz centrality* (13 итераций, 11,52 сек,

$\alpha = 2{,}38 \cdot 10^{-5}$), *Personalized PageRank* с телепортацией на историю пользователя U398400 (33 итерации, 10,59 сек), *TrustRank* с телепортацией на топ-50 доверенных треков (33 итерации, 9,31 сек). На сэмпле топ-2000 треков рассчитаны меры *closeness* (62,62 сек) и *betweenness* (70,30 сек, $k = 200$ источников).

Построена матрица ранговых корреляций Спирмена между всеми реализованными методами на сэмпле топ-2000 треков. Все попарные корреляции положительны и находятся в диапазоне 0,74–1,00, что свидетельствует о согласованности выявляемой структуры графа разными методами.

В пятом разделе описана разработанная музыкальная рекомендательная система. Реализован класс `MusicRecommender` на языке Python, объединяющий пять стратегий формирования рекомендаций: случайные рекомендации (нижняя граница качества), по популярности (in-degree), по PageRank, Personalized PageRank, гибридная стратегия. Пользователи разделены на четыре сегмента по объёму истории: новые (1–9 треков), малоактивные (10–50), активные (51–200), экспертные хабы (более 200). Гибридная стратегия переключается между глобальным PageRank (при истории менее 10 треков) и персонализированным PageRank (при большей истории), что решает проблему холодного старта.

Для оценки качества рекомендаций использовано временное разделение истории каждого пользователя в пропорции 80/20 и стандартные метрики информационного поиска: точность Precision@K, полнота Recall@K, достижение HitRate@K и нормированная дисконтированная польза NDCG@K. Для активных пользователей гибридная стратегия обеспечивает Precision@10 $\approx 40\%$ и NDCG@10 $\approx 45\%$, что соответствует верхней границе результатов, приведённых в литературе по графовым рекомендательным системам для музыкальных каталогов.

Разработан веб-интерфейс на основе фреймворка Streamlit, обеспечивающий интерактивную работу с системой. Интерфейс содержит две функциональные вкладки: «Подобрать музыку» (персональные рекомендации с автоматическим выбором стратегии и отображением метрик качества) и «Найти похожие треки» (поиск треков, близких по слушательскому поведению, на основе того же алгоритма Personalized PageRank, но с вектором телепортации,

сосредоточенным на одном целевом треке). Использование одного алгоритма для разных задач демонстрирует универсальность графового подхода.

В заключении приведены результаты магистерской работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1. Изучены основные понятия и определения теории графов; рассмотрены различные способы задания графа, классические меры центральности и математическая постановка задачи ранжирования вершин в нескольких эквивалентных формах (собственная задача, стационарное распределение Маркова, оптимизационная постановка).
2. Исследованы современные алгоритмы ранжирования вершин графа: PageRank, HITS, Katz centrality, TrustRank, Personalized PageRank, а также меры центральности closeness и betweenness. Подробно описаны их математические модели, особенности применения, преимущества и ограничения.
3. Проведён сравнительный анализ алгоритмов по критериям вычислительной сложности, устойчивости и масштабируемости. Установлено, что PageRank-подобные методы имеют сложность $O(k \cdot |E|)$ и применимы к графам промышленного масштаба, тогда как меры на основе кратчайших путей имеют квадратичную сложность и требуют распределённых вычислений.
4. Реализованы все рассмотренные алгоритмы на языке Python с использованием разреженных матричных представлений; обеспечена эффективная обработка графа из 877 168 вершин и более 11 миллионов рёбер на одной рабочей станции.
5. Экспериментально подтверждены теоретические оценки вычислительной сложности и устойчивости алгоритмов: линейный рост времени работы от размера графа, устойчивость лидирующих позиций топ-10 к случайным возмущениям (среднее пересечение 9,6 из 10), согласованность результатов разных методов (попарные корреляции 0,74–1,00).
6. На основе реализованных алгоритмов разработана гибридная рекомендательная система, обеспечивающая Precision@10 $\approx$ 40% и NDCG@10 $\approx$ 45% для активных пользователей; для системы разработан веб-интерфейс на Streamlit с двумя режимами работы (персональные рекомендации и поиск похожих треков).
7. Показано, что комбинирование различных алгоритмов ранжирования (HITS для выявления экспертных пользователей, PageRank для гло-

бального ранжирования, Personalized PageRank для персонализации, TrustRank для фильтрации спама) является наиболее перспективным подходом к построению современных рекомендательных систем и превосходит по качеству любое из решений в отдельности.