Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ»

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
к магистерской диссертации

«Мультимодальный кроссплатформенный анализ данных для выявления индивидуальных и групповых атрибутов пользователей социальных сетей»

Автор: Самборский Иван Михайлович _____
Направление подготовки (специальность): 01.04.02 Прикладная математика и информатика
Квалификация: Магистр
Руководитель: Фильченков А. А., канд. физ.-мат. наук _____

К защите допустить
Зав. кафедрой Васильев В.Н., докт. техн. наук, проф. _____

«___» _____ 20___ г.

Санкт-Петербург, 2017 г.

Студент Самборский И. М. **Группа** М4238 **Кафедра** компьютерных технологий **Факультет** информационных технологий и программирования

Направленность (профиль), специализация Технологии проектирования и разработки программного обеспечения

Консультанты:

а) Фарсеев А.И., магистр естественных наук _____

Квалификационная работа выполнена с оценкой _____

Дата защиты «___» _____ 20___ г.

Секретарь ГЭК *Павлова О.Н.* Принято: «___» _____ 20___ г.

Листов хранения _____

Демонстрационных материалов/Чертежей хранения _____

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

За последние несколько лет произошел огромный скачок популярности социальных медиа. Это может быть связано с тем, что многие пользователи сейчас используют несколько социальных сетей одновременно. Так например, согласно статистике GlobalWebIndex [1], люди в среднем ежедневно пользуются приблизительно тремя социальными сетями. Эти социальные сети связаны посредством так называемой «прошитой функциональности» (cross-linking functionality) [2], которая предоставляет информацию об одних и тех же пользователях, но с разных точек зрения [3]. Например, используя Twitter[1], можно узнать множество различных демографических признаков (пол, возраст и так далее) [4], а по Instagram[2] — особенности характера человека [5]. В попытке анализировать подавляющее количество данных, были разработаны различные системы рекомендаций для извлечения релевантной информации о пользователях. Однако, большинство из них использует данные только одной модальности (текст, изображения) или только из одного источника, игнорируя при этом потенциал мультимодальной рекомендации из нескольких источников.

Исследования, проведенные ранее, установили, что эффективная рекомендация может быть достигнута на основе комбинации индивидуальных и групповых знаний [6]. Индивидуальные знания полезны для учета личного опыта [7], в то время как групповые улучшают разнообразие рекомендации [8].

В этом исследовании мы сосредоточились на проблеме **кроссплатформенной рекомендации**, основанной на данных из нескольких социальных сетей. В частности, эта работа посвящена новой теме — **рекомендации категории места (venue)** [9–11], где мы рекомендуем ранжированный список категорий мест из социальной сети Foursquare[3] для пользователей, использующих три социальные сети одновременно, а именно Twitter, Instagram и Foursquare [4]. Рекомендация выполняется на основе мест, размещенных вокруг текущего местоположения пользователя, что делает ее независимой от какой-либо конкретной геогра-

[1] https://www.twitter.com
[2] https://www.instagram.com
[3] https://www.foursquare.com

фической области. В свою очередь, это означает, что такая модель может использоваться во многих реальных сценариях, таких как планирование туристических маршрутов [12], планировка строительного объекта [13] или интерактивная мобильная помощь [10]. Например, в зависимости от предпочтений пользователя, мы можем рекомендовать ему определенное место (например, китайский ресторан, кинотеатр и так далее) рядом с его текущим местоположением, а не аналогичное место, далеко расположенное от него, даже если оно немного лучше соответствует его предпочтениям. Наконец, рекомендация категории, а не конкретного места, помогает преодолеть трудности с оценкой, которые часто возникают из-за разреженности наборов данных.

Несмотря на возможный потенциал, который мы получаем от использования мультимодальной кроссплатформенной рекомендации, она также привносит некоторые трудности. Одна из проблем — **интеграция данных**. Различные источники данных часто описывают пользователя с разных сторон. В то же время, социальные сети содержат различные типы данных, такие как текст, изображения или видео, которые также демонстрируют различные аспекты жизни пользователей. Такие мультимодальные данные из нескольких источников необходимо уметь интегрировать в одной модели, что является открытой проблемой для исследования. Другая проблема относится к **групповому представлению**, которое может быть выражено в виде сообществ пользователей. Однако обнаружение таких пользовательских сообществ из мультимодальных данных является непростой задачей из-за необходимости правильного моделирования взаимосвязей внутри и между источниками.

Основываясь на предыдущих исследованиях и вышеперечисленных проблемах, в данной работе мы стремимся дать ответы на следующие вопросы:

а) Возможно ли улучшить точность рекомендаций, объединив индивидуальные и групповые знания?
б) Позволяет ли информация о взаимоотношениях между источниками данных найти более релевантные сообщества пользователей?
в) Каков вклад каждого источника данных в рекомендательную систему?

Для ответа на указанные вопросы, в рамках данной работы, мы представляем новую рекомендательную систему $\mathbf{C^3R}$, которая совмещает индивидуальные и групповые знания для выполнения кроссплатформенной рекомендации категории места, основанной на пользовательских сообществах (**C**ross-Source User **C**ommunity-Based **C**ollaborative venue category **R**ecommendation). Индивидуальные знания представлены в виде распределения по категориям мест, полученного из профиля пользователя Foursquare. Для добавления групповых знаний в модель, были извлечены сообщества пользователей из латентного пространства, построенного на кроссплатформенных данных, где связь между пользователями моделируется как многослойный граф. Результаты экспериментов показывают высокую точность нашей модели в трех географических регионах по сравнению с современными подходами.

Основной вклад данной работы состоит в следующем:

а) мы представили **модель для рекомендации категории места**, которая использует как индивидуальные, так и групповые знания;

б) мы предлагаем **новый подход к поиску сообществ пользователей из нескольких источников данных**, который использует как взаимосвязь между источниками, так и способность источников дополнять друг друга посредством эффективной регуляризации;

в) мы предлагаем **новый подход для автоматического построения графа отношений между источниками данных**, что устраняет необходимость наличия экспертных знаний.

ГЛАВА 1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

В этой главе мы опишем текущее состояние предметной области, а также сделаем обзор последних исследований по следующим направлениям: рекомендация и обнаружение сообществ пользователей.

1.1. Социальные медиа

Социальные медиа — веб-ресурсы, созданные для общения пользователей в сети [14]. Ярчайшими представителями данного типа коммуникации являются социальные сети. Миллионы пользователей ежедневно публикуют свои сообщения в разных сетях, таких как Twitter, Instagram, Foursquare и так далее. Например, сегодня более 56% взрослых американцев старше 65 лет пользуются социальными сетями, что свидетельствует о росте более чем в три раза по сравнению с 2010 годом, когда было зарегистрировано только 11% [15]. Пользователи сетей пользуются широкой степенью свободы, в следствии чего, они могут свободно публиковать свое мнение и легко общаться со своими друзьям. В результате чего люди постоянно обмениваются сообщениями и обсуждают различные темы: от личных событий, таких как вечеринка по случаю дня рождения, до глобальных, таких как вспышка гриппа.

Все это способствует тому, что исследователи прибегают к использованию социальных сетей в качестве источников данных в достаточно большом числе работ. Так, данные, извлеченные из Twitter, используются для решения следующих задач:

а) определение демографических признаков пользователя [16, 17];
б) определение влияния пользователей социальных сетей на составление маркетинговые кампании [18];
в) определение поведения пользователя [19].

1.2. Задача рекомендации

За последние несколько лет произошел огромный рост информации в Интернете, из-за чего становится все сложнее извлекать значимые сведения. В свою очередь, рекомендательные системы помогают фильтровать нерелевантную информацию. Они встречаются повсеместно:

а) крупнейшие веб-сайты о кинематографе, такие как IMDb[1] и Кинопоиск[2], используют рекомендательные системы, чтобы советовать пользователям новые фильмы для просмотра;
б) социальные сети, такие как Facebook[3] и Twitter, применяют их для помощи пользователям с поиском людей, которых они могут знать, или просто близких к ним по интересам.

Рекомендательная система — модель, предсказывающая, что будет интересно пользователю (фильмы, музыка и так далее), используя определенную информацию о его профиле или данные о других пользователей. Такая информация может быть разделена на две группы: явную и неявную. К явным относятся такие данные, как оценка объекта пользователем; объекты, которые понравились пользователю. А к неявным: поведение пользователя онлайн; история посещений сторонних сервисов.

В данной работе рассматривается рекомендация категории места. Такая задача имеет отличительную особенность, в сравнении с обычной задачей рекомендации, так как в результирующем векторе не исключается присутствие уже посещенных пользователем мест. Для решения таких задач часто используют методы, использующие историю пользователя, в данном случае историю посещения пользователем различных мест. Такой подход обладает значительными преимуществами: простота реализации и скорость работы, но также существует и весомый недостаток, а именно: такой подход не способен учитывать изменения интересов пользователя, что может сильно сказаться на эффективности рекомендации.

1.3. Определение сообществ пользователей

Поиск сообществ пользователей в сети, известный как *обнаружение сообществ (community detection)* [20], является фундаментальной проблемой социального компьютинга, которая привлекла большое внимание в последние годы в связи с большим интересом к исследованиями *интеллектуального анализа данных (big data)*. Полученные сообщества используются в различных приложениях, например:

[1] https://www.imdb.com
[2] https://www.kinopoisk.ru
[3] https://www.facebook.com

- для проведения таргетированных рекламных компаний;
- как признаки для более детального профилирования пользователей;
- для визуализации больших данных.

В терминах машинного обучения, задача обнаружения сообществ пользователей в социальном графе может быть представлена как задача кластеризации. Однако, с развитием социальных сетей и Интернета становится все сложнее извлекать значимые признаки для эффективной кластеризации. Так как, например, современные алгоритмы обработки данных не способны справиться с высокой размерностью получаемого представления данных. Что, в свою очередь, приводит нас к необходимости создания новых подходов, способных обрабатывать такие данные.

1.4. Обзор сопутствующих работ

Вероятно, одно из первых исследований, направленных на повышение эффективности рекомендаций, основанных на данных с несколькими источниками, было проведено авторами [21]. Они обобщили профили пользователей из Flickr[4], Twitter и Delicious[5], чтобы продемонстрировать, что их стратегии кроссплатформенного моделирования пользователей оказывают большое влияние на эффективность рекомендации в случае «холодного старта». В то же время, [22] для рекомендации по интересам пользователя использовали признаки связанные с областью интересов. Позднее, [23] предложили двухэтапную рекомендательную систему для решения задачи рекомендации видео: во-первых, предпочтения пользователя были перенесены из вспомогательной сети, изучив корреляцию поведения между вспомогательной и целевой сетями; затем предпочтения были адаптированы к поведению пользователей в целевой сети. Одновременно с этим, [24] представили вероятностную структуру, которая решила проблему кроссплатформенной рекомендации, используя общие знания области и модальности для совместного обучения скрытого представления. Недавно [10] предложили подход для рекомендацию категории места из нескольких источников,

[4] https://www.flickr.com

[5] https://www.del.icio.us

где предпочтения пользователя были получены по средствам коллаборативной фильтрации (Collaborative Filtering) ближайших соседей. Указанные работы связаны с нашим исследованием относительно применяемых подходов к использованию информации из нескольких источников. Тем не менее, ни одна из представленных работ не использует одновременно групповые и индивидуальные знания для рекомендации, которые, в свою очередь, представлены в нашей работе и являются важным аспектом данного исследования.

В то же время, в нескольких исследованиях был показан потенциал работы с кроссплатформенными данными, демонстрируя способы поиска более значимых сообществ пользователей. Например, [25] продемонстрировали полезность обнаружения кроссплатформенных сообществ для различных приложений. [26] предложили подход для случая частично перекрывающихся графов социальных сетей. В тоже время, [27] идентифицировали мультимодальные перекрывающиеся сообщества пользователей в Foursquare. Наконец, [28] представили подход кроссплатформенной кластеризации, в котором расстояние между различными источниками данных измеряется с помощью многообразия Грассмана. Работы, упомянутые выше, предоставляют стратегии обнаружения кроссплатформенных сообществ. Однако данные модели имеют значительный недостаток, заключающийся в пренебрежении отношениями между источниками данных, что может привести к неоптимальным результатам в реальных условиях.

Были также приложены усилия по изучению вклада различных источников данных в других приложениях. Например, [4] предложили модель, призванную объединить мультимодальные данные из нескольких источников для определения демографических атрибутов пользователей, в то время как, [3] и [29] разработали структурированные многозадачные подходы для оценки интересов пользователей и классификации событий, связанных со здоровьем, соответственно. В этих трех исследованиях отношение между источниками было автоматически выведено из данных и использовалось для обучения модели, что также может быть применено и для решения нашей задачи.

Выводы по главе 1

Задача кроссплатформенной мультимодальной рекомендации малоисследованна и, вместе с тем, актуальна. Цель данной работы — предложить и программно реализовать рекомендательную модель, которая позволит с высокой точностью производить предсказание категории места пользователя, используя для тренировки данные разной модальности и из разных источников.

ГЛАВА 2. ЭКСПЕРИМЕНТАЛЬНАЯ СРЕДА

В данной главе представлено описание корпуса данных, который используется для оценки эффективности предложенной модели; метрика, с помощью которой эта оценка будет производиться; а также особенности программной реализации модели.

2.1. Набор данных пользователей

Для оценки полученной модели были выбраны мультимодальные наборы данных из нескольких источников: NUS-MSS [4] и список пользователей из TwiSty [30]. Они объединяют в себе данные из трех социальных сетей одновременно (Twitter, Foursquare и Instagram). NUS-MSS был собран для периода с 10 июля 2014 по 20 декабря 2014 для трех городов: Лондон, Нью-Йорк и Сингапур. А TwiSty — для периода с 22 декабря 2010 по 27 апреля 2017 без учета регионального признака.

Данные были разделены на две выборки: *обучающую* и *тестовую*. Дата, относительно которой было произведено разделение, выбиралась с учетом максимизации пересечения пользователей обучающей и тестовой выборок. В случае NUS-MSS такой датой стало 10 октября 2014, а для TwiSty — 3 мая 2013. Для оценки моделей используются только пользователи, которые имеют записи во всех трех социальных сетях, а также обучающей и тестирующей выборках. В итоге, было получено:

- 813 пользователей для Лондона (NUS-MSS);
- 1602 для Нью-Йорка (NUS-MSS);
- 1801 для Сингапура (NUS-MSS);
- 803 без учета регионального признака (TwiSty).

Число объектов рекомендации (категорий места) одинаковое для всех трех географических регионов и равно 764.

2.1.1. Признаки Twitter

Twitter представляет текстовый тип данных, для которого посчитаны следующие виды признаков:

а) *LDA признаки*. С помощью латентного размещения Дирихле (**L**atent **D**irichlet **A**llocation, LDA) [31] для каждого пользователя были составлены вектора распределения по латентным темам. Для этого была использована реализация LDA из программного паке-

та для анализа данных KNIME[1] со следующими параметрами: $\alpha = 0{,}5$, $\beta = 0{,}1$, $T = 50$ и $N_{iter} = 1000$;

б) *Лингвистические признаки*. Из сообщений пользователей также были извлечены лингвистические признаки, используя LIWC (**L**inguistic **I**nquiry and **W**ord **C**ount) [32];

в) *Эвристические признаки*. Данные вид представлен такими признаками, как число URL ссылок, хэштегов (hashtags), упоминаний пользователей (mentions), сленговых слов, эмоциональных слов, смайлов (emoticons), орфографических ошибок и так далее.

2.1.2. Признаки Instagram

Визуальные данные представлены социальной сетью Instagram, которые были обработаны с помощью глубокой сверточной нейронной сети ImageNet [33]. Данная нейронная сеть отражает степень принадлежности обрабатываемого изображения набору концептов (таких как цветок, стул и так далее). В итоге, для каждого пользователя был получен вектор распределения по 1000 концептам.

2.1.3. Признаки Foursquare

Данные из Foursquare описывают передвижения пользователя и представлены в виде нормированного распределения по 764 различным категориям мест. Например, пользователь отметил свое присутствие в трех местах: два раза в ресторане, один в аэропорту и ни разу в любом другом месте. В таком случае, вектор c, представляющий распределение данного пользователя, содержит значения $c_r = \frac{2}{3}, c_a = \frac{1}{3}$, где r, a — индексы, отвечающие за посещение ресторана и аэропорта соответственно, когда все остальные значения вектора равняются 0.

2.2. Дополнительно извлеченные признаки

В дополнение к признакам, предоставляемым наборами данных, которые были описаны в пункте 2.1, мы дополнительно извлекли признаки основанные на времени и мобильности пользователей. Известно, что онлайн-активность пользователей социальных медиа тесно связана с временными и мобильными аспектами [34], что делает разумным

[1] `https://www.knime.org`

включение таких признаков в процесс обнаружения сообществ. Пользователи с одинаковыми моделями мобильности (часто появляющиеся в похожих местах) и схожие по временной активности (часто выполняющие действия с одинаковыми временными интервалами) могут иметь схожие интересы и могут формировать сообщество на основе интересов. Ниже приведено краткое описание признаков основанных на времени и мобильности пользователей, которые мы использовали.

Признаки мобильности вычислялись на основе *областей интереса* (AoI, **A**rea **o**f **I**nterests) пользователя [35], которые представляют из себя географические области высокой плотности относительно передвижения пользователя. Они не зависят от семантического значения геопозиции, которое, в нашем случае, может быть сообщением в Twitter, постом в Instagram с привязкой по геолокации или отметкой в Foursquare. AoI были получены путем выполнения кластеризации на основе передвижений пользователя [36], используя выпуклую оболочку каждого кластера в качестве нового AoI. Для этого был использован алгоритм кластеразиации DBScan, где ε был рассчитан путем анализа среднего расстояния между соседями в графе [37], $MinPts = 3$ был выбран эмпирическим путем. AoI представляют шаблоны мобильности пользователей [34, 35], которые могут быть связаны с их образом жизни и интересами. Для полученных AoI мы получили следующие признаки:

- **Число AoI**. Этот признак отражает физическую активность пользователя. Например, пользователи с большим числом AoI могут иметь физически напряженный образ жизни. AoI также представляют частые области деятельности пользователей, например, дом, офис, университет, школа [35] и могут указывать на то, как далеко пользователи готовы путешествовать ежедневно;
- **Средний размер AoI пользователя** — отображает мобильность пользователя внутри AoI. Размер AoI определяется как среднее расстояние между центром масс и всеми точками внутри AoI;
- **Число отметок вне привычных AoI** — показывает, как часто пользователь посещает места вне своих AoI, как часто он отклоняются от своих обычных шаблонов мобильности;

- **Среднее расстояние между AoI**. Этот признак показывает, как часто и как далеко пользователи перемещаются между зонами своей деятельности и могут быть полезны для выведения сообществ пользователей, связанных с путешествиями.

В качестве временных признаков было выбрано **среднее число сообщений в определенный период времени**. Всего таких периодов 8, длительность каждого из которых 3 часа. Признаки были рассчитаны для каждого источника данных с учетом буднего и выходного дня. Всего было получилось $8 \cdot 3 \cdot 2 = 48$ признака. Такие признаки указывают на временную онлайн-активность пользователей [34].

2.3. Особенности программной реализации

Модель реализована с использованием языка программирования Java 8 и следующего списока библиотек:

- WEKA[2] — одна из наиболее популярных библиотек машинного обучения для Java. Также был использован ряд библиотек расширяющих стандартный функционал WEKA: XMeans[3], DBScan[4];
- jblas[5] — библиотека линейной алгебры, которая содержит огромное число различных операций над матрицами;
- Apache Mahout[6] — библиотека для создания распределенных алгоритмов машинного обучения.

Исходный код проекта выложен в открытый доступ на GitHub[7].

2.4. Метрика оценки работы алгоритмов

Для оценки эффективности модели, мы используем широко распространенную при работе с рекомендательными системами метрику **NDCG** (**N**ormalized **D**iscounted **C**umulative **G**ain). NDCG — метрика, показывающая корректность рекомендации. В отличие от многих других распространенных метрик ранжирования, она позволяет учитывать порядок объектов, что важно для оценки результатов рекомендации кате-

[2] http://www.cs.waikato.ac.nz/ml/weka/
[3] http://weka.sourceforge.net/packageMetaData/XMeans/index.html
[4] http://weka.sourceforge.net/packageMetaData/optics_dbScan/
[5] http://jblas.org/
[6] http://mahout.apache.org/
[7] https://github.com/Nilera/C3R

гории места. Она определяется как:

$$\mathrm{NDCG@}p = \frac{\mathrm{DCG@}p}{\mathrm{IDCG@}p}, \mathrm{DCG@}p = \sum_{i=1}^{p} \frac{2^{rel_i}}{\log_2(i+1)}, rel_i = \frac{Cat_i}{N_{Cat}},$$

где p — максимальное число объектов, которое можно рекомендовать, IDCG — максимально возможное (идеальное) значение DCG для конкретного запроса, rel_i — релевантность результата на позиции i, Cat_i — число раз, которое пользователь был в месте категории i, $N_{Cat} = 764$ — общее число категорий в Foursquare.

Выводы по главе 2

Описаны наборы данных, которые используются нами при исследовании задачи кроссплатформенной мультимодальной рекомендации категории места, особенности программной реализации, а также метрика для оценки эффективности полученной модели.

ГЛАВА 3. МУЛЬТИМОДАЛЬНАЯ, КРОССПЛАТФОРМЕННАЯ РЕКОМЕНДАЦИЯ

Известно, что на процесс принятия стратегических решений влияют различные факторы, такие как личный опыт и общественное мнение [38]. Общественное мнение может быть выражено через сообщества пользователей, которые могут формироваться на основе социальных отношений, а могут по средствам схожести пользователей между собой. Такое явление используется для повышения эффективности рекомендаций. Для того чтобы проверить данное предположение, наша рекомендательная система будет учитывать как персональные, так и групповые знания о пользователе, что в свою очередь, моделирует естественное влияние общества на поведение человека при выборе чего-то нового. Формально наш подход к рекомендации $\mathrm{C^3R}$ записывается следующим образом:

$$\mathrm{rec}(u) = sort\left(\gamma \cdot vec_u + \theta \frac{\sum_{v \in C_u} vec_v}{|C_u|}\right) \tag{1}$$

где vec_u — распределение объектов рекомендации (категорий мест) для пользователя u, $\frac{\sum_{v \in C_u} vec_v}{|C_u|}$ — нормализованное распределение всех представителей сообщества между объектами, γ — контролирует личный аспект рекомендации, а θ — групповой.

Помимо описанных выше преимуществ, использование сообществ пользователей значительно уменьшает пространство поиска во время процесса рекомендации и оставляет только лучших кандидатов для проведение дальнейшей коллаборативной фильтрации [39]. В то время как, личная информация, такая как предыдущая активность пользователей, часто доступна, в большинстве случаев сообщества пользователей не указаны явно, что порождает проблему обнаружения сообществ пользователей.

3.1. Обнаружение сообществ пользователей

Первым шагом в поиске репрезентативных групп пользователей является моделирование взаимоотношений пользователей в виде графа, в таком случае плотные подграфы являются сообществами пользователей. Такой граф может быть построен несколькими способами:

а) с помощью социальных связей между пользователями (подписки/-подписчики), которые часто скрыты за настройками конфиденциальности;

б) используя пользовательский контент, где сходство между пользователями оценивается как расстояние между представлением данных пользователей и каждым источником данных, смоделированным как слой многослойного графа.

Многослойный граф — граф , представляющий из себя множество графов с одним набором вершин, но разным набором ребер. Формально данное определение записывается следующим образом: $G = \{G_i\}, G_i = (V, E_i)$. Пример такого графа изображен на рисунке 1.

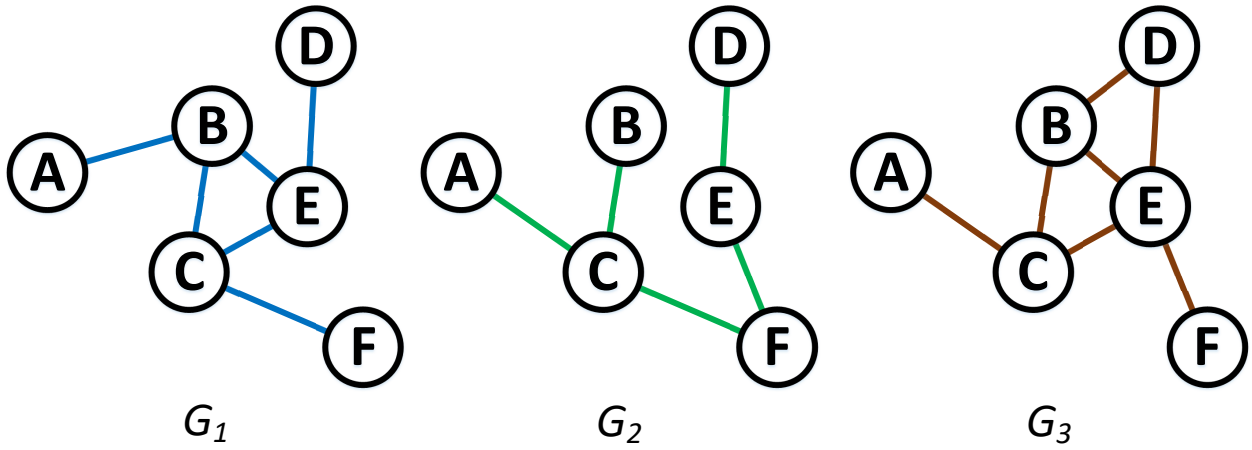

Рисунок 1 – Пример многослойного графа G, состоящего из трех слоев, где V — множество вершин, E_i — множество ребер i-ого слоя

В нашей работе мы используем построение графа на основе пользовательского контента, чтобы избежать проблем конфиденциальности и ограничений при явном использовании социальных связей. Так как часто пользователи скрывают часть своей информации, что усложняет построение социального графа. Мы взвешиваем рёбра между вершинами таким образом, что ребро получает наибольший вес, если соединяет наиболее похожие вершины (пользователей). Для достижения данного свойства веса присваиваются в соответствии с радиально-базисной функцией ядра (RBF kernel, **R**adial **B**asis **F**unction kernel) [40]:

$$K(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right),$$

где $||\cdot||$ — евклидова норма, а σ — положительный параметр, влияющий на окрестность точек, в которой ребра получают больший вес.

Такой подход к построение графа обладает некоторыми преимуществами [41]. Например, такой подход не страдает от недостатка информации о взаимоотношениях пользователей, поскольку отношения между пользователями просто моделируются на основе сходства пользовательского контента.

3.2. Формулировка проблемы

Одной из наиболее часто используемых формулировок проблемы обнаружения сообщества является представление в виде NCut, что позволяет минимизировать сумму весов ребер в каждом подграфе (сообществе) [42]. Другими словами, это означает, что все найденные сообщества формируются пользователями, которые наиболее похожи друг на друга, что хорошо применимо к нашей задаче. Определение NCut приведено ниже:

$$\mathrm{NCut}(C_1, ..., C_k) = \sum_{i=1}^{k} \frac{\mathrm{W}(C_i, \overline{C}_i)}{\mathrm{vol}(C_i)} = \sum_{i=1}^{k} \frac{\mathrm{cut}(C_i, \overline{C}_i)}{\mathrm{vol}(C_i)},$$

где $\mathrm{vol}(C_i)$ — сумма весов всех ребер в подграфе C_i.

Однако, доказано что проблема NCut является $\mathcal{NP}$-трудной [43]. К счастью, существует приближение, которое определяется как минимизация следа (также известная как *спектральная кластеризация*) [44]:

$$\min_{U \in R^{n \times k}} \mathrm{tr}(U^\intercal L_{sym} U), U^\intercal U = I, \tag{2}$$

где L_{sym} — матрица Кирхгофа, I — единичная матрица.

По теореме Рэлея-Ритца, решением уравнения 2 являются первые k собственных векторов матрицы Кирхгофа:

$$L_{sym} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}},$$

где W — матрица смежности, а D — матрица степеней вершин графа [45]. Далее используя, например, k-means кластеризацию в простран-

стве собственных векторов, мы получим итоговые кластеры (сообщества) [44].

3.3. Спектральная кластеризация многослойного графа

Известные недостатки раннего и позднего агрегирования данных [3, 29] побуждают нас выполнять кластеризацию всех источников данных (слоев графа) одновременно. Окончательное латентное представление данных должно быть согласовано со всеми слоями графа, чтобы кластеризация не была смещена в сторону отдельных источников. Один из способов сохранить латентное представление, совместимое со всеми слоями многослойного графа, — это применить регуляризацию, которая учитывала бы расстояние между слоями графа и конечным представлением. Для измерения расстояния между конечным латентным пространством U и латентным представлением каждого слоя U_i мы будем использовать многообразие Грассмана (рисунок 2) [46]:

$$\begin{aligned} d^2_{proj}(U_1, U_2) &= \sum_{i=1}^{k} \sin^2 \theta_i \\ &= k - \sum_{i=1}^{k} \cos^2 \theta_i \\ &= k - \operatorname{tr}(U_1 U_1^\intercal U_2 U_2^\intercal) \\ &= \frac{1}{2}(\operatorname{tr}(U_1 U_1^\intercal) + \operatorname{tr}(U_2 U_2^\intercal) - 2\operatorname{tr}(U_1 U_1^\intercal U_2 U_2^\intercal)) \\ &= \frac{1}{2}||U_1 U_1^\intercal - U_2 U_2^\intercal||^2_F, \end{aligned}$$

где $||A||_F$ — норма Фробениуса [47] матрицы A.

Таким образом, расстояние между конечным пространством U и всеми индивидуальными пространствами $\{U_i\}_{i=1}^M$ определяются как сумма квадратов расстояний между U и пространствами $\{U_i\}_{i=1}^M$ [28]:

$$\begin{aligned} d^2_{proj}(U, \{U_i\}_{i=1}^M) &= \sum_{i=1}^{M} d^2_{proj}(U,U_i) \\ &= kM - \sum_{i=1}^{M} \operatorname{tr}(UU^\intercal U_i U_i^\intercal). \end{aligned} \tag{3}$$

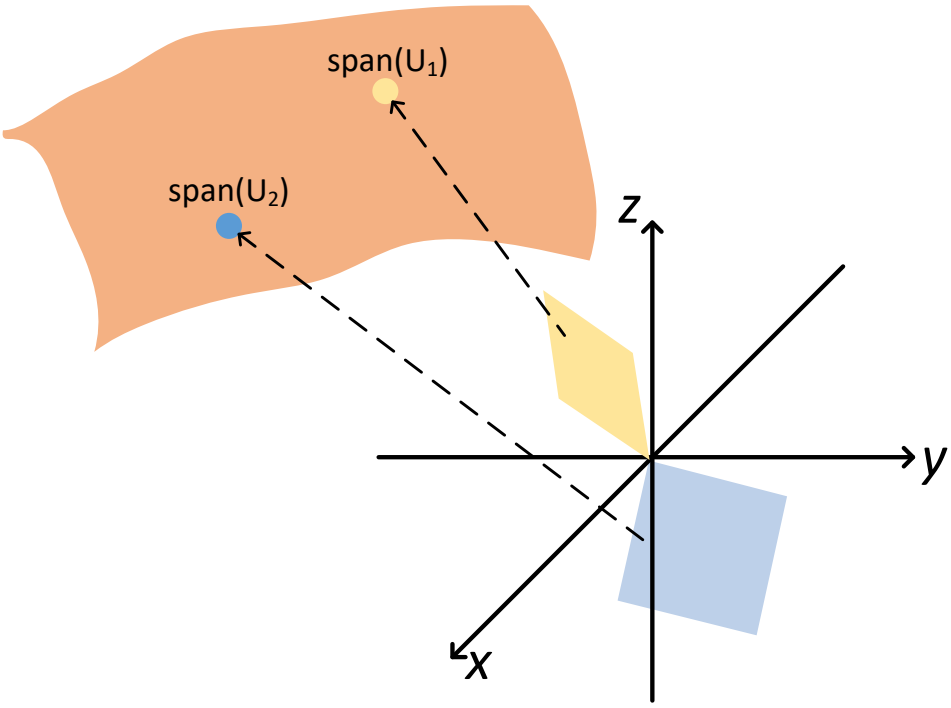

Рисунок 2 – Пример использования многообразия Грассмана для перевода из $\mathbb{R}^3$ в точки на $\mathcal{G}(3, 2)$

Используя уравнение 3, уравнение 2 может быть расширино, добавив регуляризацию:

$$\min_{U \in \mathbb{R}^{n \times k}} \sum_{i=1}^{M} \operatorname{tr}(U^\intercal L_i U) + \alpha \left(kM - \sum_{i=1}^{M} \operatorname{tr}(UU^\intercal U_i U_i^\intercal) \right), U^\intercal U = I, \qquad (4)$$

где α — коэффициент регуляризации.

С помощью несложных преобразований уравнение 4 приводится к виду проблемы минимизации следа:

$$\begin{aligned}
&\min_{U\in\mathbb{R}^{n\times k}} \sum_{i=1}^{M} \operatorname{tr}(U^\intercal L_i U) + \alpha \left(kM - \sum_{i=1}^{M} \operatorname{tr}(UU^\intercal U_i U_i^\intercal) \right) \\
&= \min_{U\in\mathbb{R}^{n\times k}} \sum_{i=1}^{M} \operatorname{tr}(U^\intercal L_i U) - \alpha \sum_{i=1}^{M} \operatorname{tr}(UU^\intercal t U_i U_i^\intercal) \\
&= \min_{U\in\mathbb{R}^{n\times k}} \sum_{i=1}^{M} \operatorname{tr}(U^\intercal L_i U) - \alpha \sum_{i=1}^{M} \operatorname{tr}(U^\intercal U_i U_i^\intercal U) \\
&= \min_{U\in\mathbb{R}^{n\times k}} \operatorname{tr} \sum_{i=1}^{M} (U^\intercal L_i U - \alpha U^\intercal U_i U_i^\intercal U) \\
&= \min_{U\in\mathbb{R}^{n\times k}} \operatorname{tr} \left(U^\intercal \sum_{i=1}^{M} (L_i U - \alpha U_i U_i^\intercal U) \right) \\
&= \min_{U\in\mathbb{R}^{n\times k}} \operatorname{tr} \left(U^\intercal \sum_{i=1}^{M} (L_i - \alpha U_i U_i^\intercal) U \right),
\end{aligned} \tag{5}$$

Уравнение 5, по теореме Рэлея-Ритца [45], может быть решено поиском первых k собственных векторов в измененной матрице Кирхгофа:

$$L_{mod} = \sum_{i=1}^{M} (L_i - \alpha U_i U_i^\intercal).$$

Собственное разложение матрицы L_{mod} может быть эффективно вычислено, используя уже известные алгоритмы [48].

3.4. Использование отношений между источниками данных

Несмотря на то, что представленный выше подход с использованием регуляризации хорошо работает для синтетических наборах данных [28], в реальном мире часто необходимо учитывать взаимосвязи между источниками [3]. Такое требование вызвано различием между источниками данных и тем как они описывают пользователей [10]. В частности, в некоторых ситуациях одни источники данных могут быть более информативными, чем другие. Например, данные из Foursquare могут быть более полезны для рекомендации категории места, чем текстовые сооб-

щения, в то время как текстовые данные имеют решающее значение для демографического профилирования [4].

Основываясь на нашем наблюдении и предыдущих работах [3, 10, 49], мы делаем еще один шаг в кроссплатформенной кластеризации, введя новую модель регуляризации, которая будет учитывать отношение между источниками данных для лучшей рекомендации.

Учитывая многослойный граф пользователей G, построенный, как описано в пункте 3.1, предположим, что существует полный неориентированный граф подобия R с матрицей смежности W_R, где каждое значение матрицы $w_{i,j}$ представляет подобие между i-ым и j-ым слоем графа G. Построение графа R рассмотрено в пункте 3.5.

Наша цель состоит в том, чтобы регуляризовать спектральную кластеризацию таким образом, что представление каждого слоя $\hat{U}_i$ вычислялось бы с учетом подобия между источниками данных $w_{i,j}$, которое взято из матрицы смежности графа подобия R. Заметим также, что подход в уравнении 5 опирается на предварительно вычисленные матрицы Кирхгофа $\{L_i\}_{i=1}^M$ и соответствующие им спектральные пространства $\{U_i\}_{i=1}^M$. В нашей работе мы применяем регуляризацию между источниками данных в процессе вычисления спектральных пространств для каждого из слоев, так что конечное многослойное спектральное пространство вычисляется так же, как и в уравнении 5. Используя ранее определенное расстояние на многообразии Грассмана, мы определяем новую целевую функцию для i-ого слоя следующим образом:

$$\min_{\hat{U}_i \in \mathbb{R}^{n \times k}} \operatorname{tr}(\hat{U}_i^\intercal L_i \hat{U}_i) + \beta_i \left(kM - \sum_{j=1, j \neq i}^{M} w_{i,j} \operatorname{tr}(\hat{U}_i \hat{U}_i^\intercal U_j U_j^\intercal) \right), \quad \hat{U}_i^\intercal \hat{U}_i = I, \tag{6}$$

где $\hat{U}_i^\intercal$ — новое спектральное представление i-ого слоя, β_i — параметр, который контролирует регуляризацию между источниками данных для i-ого слоя, $w_{i,j}$ — коэффициент подобия между i-ого и j-ого слоя. После преобразований, схожих с преобразованиями для уравнения 5, уравне-

ние 6 может быть сведено к проблеме минимизации следа:

$$\min_{\hat{U}_i \in \mathbb{R}^{n\times k}} \operatorname{tr}(\hat{U}_i^\intercal L_i \hat{U}_i) + \beta_i \left(kM - \sum_{j=1, j\neq i}^{M} w_{i,j} \operatorname{tr}(\hat{U}_i \hat{U}_i^\intercal U_j U_j^\intercal) \right)$$
$$= \min_{\hat{U}_i \in \mathbb{R}^{n\times k}} \operatorname{tr} \left(\hat{U}_i^\intercal \left(L_i - \beta_i \sum_{j=1, j\neq i}^{M} w_{i,j} U_j U_j^\intercal \right) \hat{U}_i \right),$$

что по теореме Рэлея-Ритца, может быть решено с помощью первых k собственных векторов матрицы Кирхгофа $\hat{L}$:

$$\hat{L}_i := L_i - \beta_i \sum_{j=1, j\neq i}^{M} w_{i,j} U_j U_j^\intercal.$$

Мы представили все необходимые компоненты, чтобы определить наш подход к кроссплатформенному обнаружению сообществ пользователей со следующей целевой функцией:

$$\min_{U \in \mathbb{R}^{n\times k}} \sum_{i=1}^{M} \operatorname{tr}(U^\intercal \hat{L}_i U) + \alpha \left(kM - \sum_{i=1}^{M} \operatorname{tr}(U U^\intercal \hat{U}_i \hat{U}_i^\intercal) \right)$$
$$= \min_{U \in \mathbb{R}^{n\times k}} \operatorname{tr} \left(U^\intercal \sum_{i=1}^{M} (\hat{L}_i - \alpha \hat{U}_i \hat{U}_i^\intercal) U \right), \tag{7}$$
$$U^\intercal U = I.$$

Чтобы сделать процедуру кластеризации ясной, мы привели псевдокод на листинге 1. Из псевдокода видно, что оптимизация уравнения 7 и дальнейшая кластеризация состоит из пяти основных этапов:

а) выполнение спектральной кластеризации на каждом слое для получения L_i и U_i;
б) выполнение спектральной кластеризации с учетом регуляризации на каждом слое, используя граф R, для получения $\hat{L}_i$ и $\hat{U}_i$;
в) выполнение очередной спектральной кластеризации, используя $\hat{L}_i$ и $\hat{U}_i$, для получения $\hat{L}_{mod}$ и U;
г) нормализация U для получения U_{norm};

д) используя U_{norm}, выполнить кластеризацию, например x-means [50].

Листинг 1 – $\mathrm{C^3R}$ кластеризация

function cluster($\{W_i\}_{i=1}^M$, W_R, k, α, $\{\beta_i\}_{i=1}^M$)
▷ $\{W_i\}_{i=1}^M$ — матрицы смежности слоев многослойного графа,
W_R — матрица смежности графа подобия,
k — итоговое число кластеров,
α и $\{\beta_i\}_{i=1}^M$ — параметры регуляризации

---

    for $i \leftarrow [0; M-1]$ **do**
        Вычислить L_i и U_i для G_i [44]
        ▷ L_i — матрица Кирхгофа слоя i,
        U_i — спектрально разложение L_i,
        G_i — i-ый слой графа
        Вычислить $\hat{L}_i \leftarrow L_i - \beta_i \sum_{j=1, j\neq i}^{M} w_{i,j} U_j U_j^\intercal$
        ▷ $\hat{L}_i$ — регуляризованная матрица Кирхгофа для i-ого слоя
        Вычислить $\hat{U}_i \in \mathbb{R}^{N\times k}$
        ▷ $\hat{U}_i$ — спектральное разложение матрицы $\hat{L}_i$,
        представляющее из себя первые k собственных векторов $\hat{L}_i$ [48]
    end for
    Вычислить $\hat{L}_{mod} \leftarrow \sum_{i=1}^{M} (\hat{L}_i - \alpha \hat{U}_i \hat{U}_i^\intercal)$
    ▷ $\hat{L}_{mod}$ — измененная матрица Кирхгофа [28]
    Вычислить $U \in \mathbb{R}^{N\times k}$
    ▷ U — матрица, состоящая из первых k собственных векторов $\hat{L}_{mod}$
    Нормализовать U чтобы получить U_{norm}
    $\{C\}_{i=1}^k \leftarrow$ simpleCluster(U_{norm})
    ▷ simpleCluster() — любой алгоритм кластеризации,
    например, k-means или x-means

---

    return $\{C\}_{i=1}^k$
    ▷ $C_1, ..., C_k$ — итоговые кластеры
end function

Значение параметров α и β_i находится с помощью метода спуска (hill climbing) [51].

3.5. Граф подобия

Граф подобия R должен представлять сходство между слоями. Мы определяем подобие $\mathrm{sim}(w, q)$ между слоями q и w как нормализованное

различие между матрицами $M_{q,k}, M_{w,k} \in N \times N$, где

$$m_{i,j} = \begin{cases} 1, & \text{если пользователи } i \text{ и } j \text{ в одном кластере,} \\ 0, & \text{иначе.} \end{cases}$$

Такие матрицы получаются путем выполнения однослойной спектральной кластеризации на каждом слое в отдельности при различных значениях k ($k = 2..K$, где $K = \sqrt{N}$ [52]). Затем мы берем среднее значение по всем k:

$$\mathrm{sim}(w, q) = \frac{\left(\sum_{k=2}^{K} 1 - \frac{||M_{w,k} - M_{q,k}||}{\sqrt{N(N-1)}} \right)}{K - 1}.$$

Приведенная выше формулировка является модифицированной и нормализованной версией измерения разности разделов (Partition Difference measurement) [53]. Будучи усредненной по различным значениям k, она может служить надежным индикатором сходства между различными социальными сетями с точки зрения результатов кластеризации. Мы явно хотели бы отметить, что предложенный нами подход построения графов подобия является автоматизированным и не требует каких-либо экспертных знаний, что предполагает его дальнейшее использование в других подходах с обучением без учителя.

3.6. Оценка времени вычислений

Для оценки времени кластеризации нам необходимо оценить сложность каждого шага алгоритма 1. Пусть N — число пользователей, M — число источников данных (слоев графа), а k — конечное число кластеров. Тогда сложность времени кластеризации может быть оценена как $\mathcal{O}(N^2(M^2k + MN + k^2))$. Ниже приведено вычисление получившейся оценки более подробно.

Во-первых, вычислительная сложность каждой матрицы Кирхгофа L_i и матрицы собственных векторов U_i — $\mathcal{O}(N^3)$, что в сумме по всем слоям графа — $\mathcal{O}(MN^3)$. Вычисление каждой матрицы $\hat{L}_i$ стоит $\mathcal{O}(MN^2k)$, что в итоге дает $\mathcal{O}(M^2N^2k)$. Итоговая стои-

мость вычисления $\hat{U}_i$ — $\mathcal{O}(MN^3)$. $\hat{L}_{mod}$ вычисляется за $\mathcal{O}(MN^2k)$ времени. Вычисление матрицы U занимает $\mathcal{O}(N^3)$ времени, в то время как x-means кластеризация в пространстве U_{norm} — $\mathcal{O}(N^2k^2)$ [50]. $\mathcal{O}(M^2N^2k) + \mathcal{O}(MN^3) + \mathcal{O}(N^2k^2) = \mathcal{O}(N^2(M^2k + MN + k^2))$ — конечное время, которое необходимо затратить на кластеризацию.

Выводы по главе 3

Предложена модель $\mathrm{C}^3\mathrm{R}$, которая использует как индивидуальные, так и групповые знания для решения задачи рекомендации категории места. Индивидуальные знания моделируются как распределение среди категорий места, которое посетил пользователь в прошлом, в то время как для моделирования групповых был разработан подход мультимодальной кластеризации из нескольких источников для эффективного поиска сообществ пользователей.

ГЛАВА 4. ЭКСПЕРИМЕНТЫ

Существует два основных подхода для оценки алгоритмов кластеризации: явный и неявный [27]. Явная оценка проводится по средствам вычисления какой-либо метрики, например, модульность (modularity). Однако, нет какой-либо общепринятой метрики для оценки качества кроссплатформенной кластеризации. Более того, многие такие метрики слабо связаны с реальными кластерами [20]. Неявная оценка, в свою очередь, сравнивает результаты, достигнутые подходами из других областей, например, рекомендация, классификация и так далее. Такие подходы должны быть созданы на основе ранее полученных кластеров. Случай неявной оценки хорошо согласуется с нашим исследованием и позволяет оценить как наш предложенный подход кроссплатформенной рекомендаций, так подход для поиска кроссплатформенных сообществ пользователей. Таким образом, в данной работе мы сравниваем подходы, используя неявную оценку.

4.1. Взаимосвязь между источниками данных

Как упоминалось ранее, различные социальные сети описывают пользователя с разных точек зрения. Однако для поиска более релевантных сообществ необходимо чтобы объединяемые источники данных полностью не противоречили друг другу [8]. Поэтому важно понимать, что общего у различных источников данных.

Чтобы проанализировать взаимосвязь между источниками, обнаруженными из каждого источника данных независимо, мы используем коэффициент корреляции r-Пирсона, который поможет нам показать взаимосвязь между сообществами пользователей. Для обнаружения таких сообществ, мы используем обычную спектральную кластеризацию [45]. Следует отметить, что в этом эксперименте мы устанавливаем число кластеров $k = 4$, поскольку это позволяет нам достичь наилучшей эффективности рекомендаций в случае, когда спектральная кластеризация выполняется для каждой социальной сети в отдельности.

Корреляционный анализ выполняется на основе данных из набора NUS-MSS [4], а в частности, на базе пользователей из Сингапура, которые имеют учетные записи во все трех социальных сетях: Twitter, Instagram

Таблица 1 – Пример векторов распределения сообществ

Пользователь	Сообщества Foursquare				Сообщества Instagram			
	4sq0	4sq1	4sq2	4sq3	inst0	inst1	inst2	inst3
user 1	0	0,27	0	0	0,23	0	0	0
...	...	...	...	...	...	...	...	...
user k	0,14	0	0	0	0	0	0,33	0

и Foursquare. Коэффициент r подсчитан на основе векторов распределения сообществ, который вычисляется следующим образом: для каждого полученного сообщества значение в позиции i равно косинусному расстоянию между центроидом сообщества и i-ым пользователем, если i-ый пользователь принадлежит этому сообществу, иначе значение равно 0. Чтобы прояснить эту идею, мы приводим пример в таблице 1. *user 1* принадлежит сообществу *4sq1* в Foursquare и *inst0* в Instagram, а *user k* — *4sq0* и *inst2*, соответственно.

Полученные статически значимые результаты[1] представлены в виде корреляционной матрицы на рисунке В.1. Для того чтобы проще интерпретировать полученные результаты мы произвели профилирование некоторых сообществ пользователей, которые представлены в таблице 2.

Из корреляционной матрицы видно, что сообщество *inst1* из Instagram (связанное с досугом) положительно коррелирует с сообществом *4sq1* из Foursquare («Покупки/Еда»), но отрицательно коррелирует с временным сообществом *temp2*, пользователи которого активны в выходные дни с 12 по 15 часов, и сообществом Foursquare *4sq3* («Бизнес/Путешествия»). Первое отношение объясняется сходством между образами жизни членов сообщества, для которых досуг часто сопровождается покупками или посещениями ресторанов. На основе отрицательной корреляции, в свою очередь, может быть выдвинута гипотеза, что у пользователей из сообщества «Досуг» отсутствуют активности связанные с работой.

[1] Уровень α значимости теста равен $0{,}05$

Таблица 2 – Профили сообществ пользователей

№	Имя	Профиль
Twitter		
1	Работа	#misspellings, #words/tweet, work
Foursquare		
1	Покупки/Еда	Mall, Chinese Restaurant, Fast Food, Japanese Restaurant, Food Court
3	Бизнес/Путешествия	Airport, Café, Hotel, Bakery, Bus Station
Instagram		
1	Досуг	Lotion, Sunscreen, Sea slug, Swan
3	Дом/Семья	Gong, Backpack, Dog, Pot, Library (Book)
Temporal		
2	-	Weekends 12 – 15
3	-	Weekdays 16 – 22
Mobility		
3	Исследование	#AOI outliers

В то же время сообщество из Twitter *tw1* («Работа») положительно коррелирует со временными сообществами *temp2* и *temp3*, что объясняется графиком работы ориентированных на работу пользователей.

Наконец, сообщество Foursquare *4sq3* («Бизнес/Путешествия») положительно коррелирует с сообществом Instagram *inst3* («Дом/Семья»), но отрицательно коррелирует с сообществом *mob3* («Исследование»). Оба примера объясняется образом жизни пользователей сообщества «Бизнес/Путешествия». Например, это может быть вызвано тенденцией ориентированных на бизнес пользователей проводить больше времени с семьей, а также привычками посещать знакомые места.

Стоит отметить, что большинство сообществ, извлеченных из одних и тех же слоев, отрицательно коррелируют друг с другом, что показывает хорошую эффективность спектральной кластеризации. Большое число значительных корреляций между сообществами показывает, что источники данных разных модальностей могут дополнять друг друга, что поддерживает идею совместного обнаружения сообщества из кроссплатформенных мультимодельных данных.

4.2. Базовые алгоритмы

В данной секции мы опишем различные подходы к рекомендации и различные модификации $\mathrm{C^3R}$.

4.2.1. Базовые алгоритмы рекомендации

В качестве базовых алгоритмов рекомендации были выбраны следующие методы:

- **Popular (POP)** — выполняет рекомендацию, основываясь только на прошлом опыте пользователя, в нашем случае это распределение по категориям мест, в которых отмечался пользователь в прошлом. Стоит отметить, что это является частным случаем $\mathrm{C^3R}$ рекомендации (уравнение 1), где коэффициент $\theta = 0$;
- **Popular**$_{All}$ **(POP**$_{All}$**)** — рекомендация основывается на агрегировании опыта всех пользователей, что приводит к рекомендации не зависящий от пользователя;
- **Multi-Source Re-Ranking (MSRR)** [10] — линейно объединяет результаты рекомендаций из всех источников данных, используя веса определенные с помощью метода спуска: $w_{tw} = 0{,}42$, $w_{fsq} = 0{,}95$, $w_{isnt} = 0{,}36$, $w_{temp} = 0{,}65$, $w_{mob} = 0{,}02$;
- **Коллаборативная фильтрация ближайшего соседа (CF)** [54] — производит рекомендацию, основываясь на $k = 20$ наиболее близких пользователей из источника рекомендации, в нашем случае Foursquare, используя меру подобия, в нашем случае косинусное расстояние;
- **Раннее объединение (EF, Early Fusion)** [55] — объединяет данные с несколькими источниками в единый вектор признаков и выполняет рекомендацию, используя CF;
- **Сингулярное разложение, усиленное неявной обратной связью (SVD++, Implicit Feedback-Enhanced Singular Value Decomposition)** [56] — рекомендация с использованием неявной обратной связью в факторизационной модели, где $\lambda = 0{,}67$, $\gamma = 0{,}06$, $k = 277$, $iter = 55$ подобраны с помощью метода спуска;
- **Факторизационные машины (FM, Factorization Machines)** [57] — объединяет преимущества различных моделей основанных на

факторизации посредством эффективной регуляризации. В нашем исследовании мы обучали FM на основе 764 различных категорий места, использовали технику оптимизации *MCMC* [57] и регрессионную функцию потерь (regression loss), где $k = 30, iter = 140$ подобраны с помощью метода спуска.

4.2.2. C³R и его модификации

В четырех модификациях $\mathrm{C^3R}$ ниже рекомендация выполняется в соответствии с уравнением 1:

- $\mathbf{C^3R}$ — предложенный нами подход (уравнение 1), где сообщества пользователей определяются минимизацией уравнения 7 с помощью алгоритма 1. Используя следующие параметры: $k = 10$, $\alpha = 0{,}922$, $\beta_i = 1\ \forall i \in [1, M]$, $\gamma = 1$, $\theta = 0{,}248$, подобранные методом спуска. Матрица смежности графа подобия W_R построена автоматически, как описано в пункте 3.5:

$$W_R = \left(\begin{array}{c|ccccc} & \textbf{txt} & \textbf{loc} & \textbf{vis} & \textbf{tmp} & \textbf{mob} \\ \hline \textbf{txt} & 1 & 0{,}632 & 0{,}621 & 0{,}643 & 0{,}561 \\ \textbf{loc} & 0{,}632 & 1 & 0{,}614 & 0{,}631 & 0{,}570 \\ \textbf{vis} & 0{,}621 & 0{,}614 & 1 & 0{,}621 & 0{,}551 \\ \textbf{tmp} & 0{,}643 & 0{,}631 & 0{,}621 & 1 & 0{,}560 \\ \textbf{mob} & 0{,}561 & 0{,}570 & 0{,}551 & 0{,}560 & 1 \end{array}\right),$$

  где **txt** — текстовые данные (Twitter), **loc** — данные о геолокации (Foursquare), **vis** — визуальные (Instagram), **tmp** — временные, **mob** — данные о мобильности пользователей;
- $\mathbf{C^3R_{-\hat{L}}}$ — C^3R рекомендация без регуляризации, которая учитываем отношение между источниками данных ($\beta_i = 1\ \forall i \in [1, M]$), где $k = 10$, $\alpha = 0{,}521$, $\gamma = 1$, $\theta = 0{,}1$ подобраны методом спуска;
- $\mathbf{C^3R_{-\hat{L}-L_{mod}}}$ — C^3R без регуляризаций ($\beta_i = 1\ \forall i \in [1, M]$, $\alpha = 0$), где $k = 22$, $\gamma = 1$, $\theta = 0{,}147$ подобраны методом спуска;
- $\mathbf{C^3R_{-omm}}$ — C^3R рекомендация без использования сообществ пользователей, то есть все пользователи участвуют во время рекомендации.

Также, для проверки качества кластеризации с помощью предложенного нами алгоритма, были проверено несколько базовых кластеризаций:

- **C^3R (DBScan)** — C^3R рекомендация, где поиск сообществ осуществляется с помощью DBScan кластеризации [36] со следующими параметрами: $\varepsilon = 0{,}9$ и $MinPts = 6$ — подобранными методом спуска;
- **C^3R (x-means)** — поиск сообществ осуществляется с помощью x-means кластеризации [50];
- **C^3R (Hierarchical)** — поиск сообществ осуществляется с помощью иерархической кластеризации. Параметры $k = 10$ и «односвязное» расстояние между кластерами получено методом спуска.

4.3. Сравнение с базовыми алгоритмами

Чтобы ответить на один из вопросов исследования: возможно ли улучшить точность рекомендаций, объединив индивидуальные и групповые знания, мы оцениваем модель C^3R относительно других рекомендательных систем. Результаты оценки для набора данных NUS-MSS представлены на рисунке 3, для TwiSty — на рисунке А.1.

Во-первых, интересным наблюдением является плохая эффективность подхода EF, что в очередной раз [10] указывает на необходимость использования надлежащих подходов при работе с данными из нескольких источников. С другой стороны, POP показывает достаточно хорошую эффективность рекомендации, особенно для небольших значений p, что и предполагалось в пункте 1.2. Это означает, что пользователи Foursquare, как правило, повторно посещают те места, которые они уже посещали в прошлом, что подчеркивает важность индивидуальных знаний для рекомендации категории места и позволяет POP достичь высокой точности. В то же время POP_{All} превосходит подходы FM и EF, что показывает полезность групповых знаний для рекомендации. Наконец, видно, что комбинация индивидуальных и групповых знаний в C^3R значительно превосходит другие базовые подходы. Что доказывает тот факт, что **комбинация индивидуальных и групповых знаний значительно улучшает точность рекомендации**.

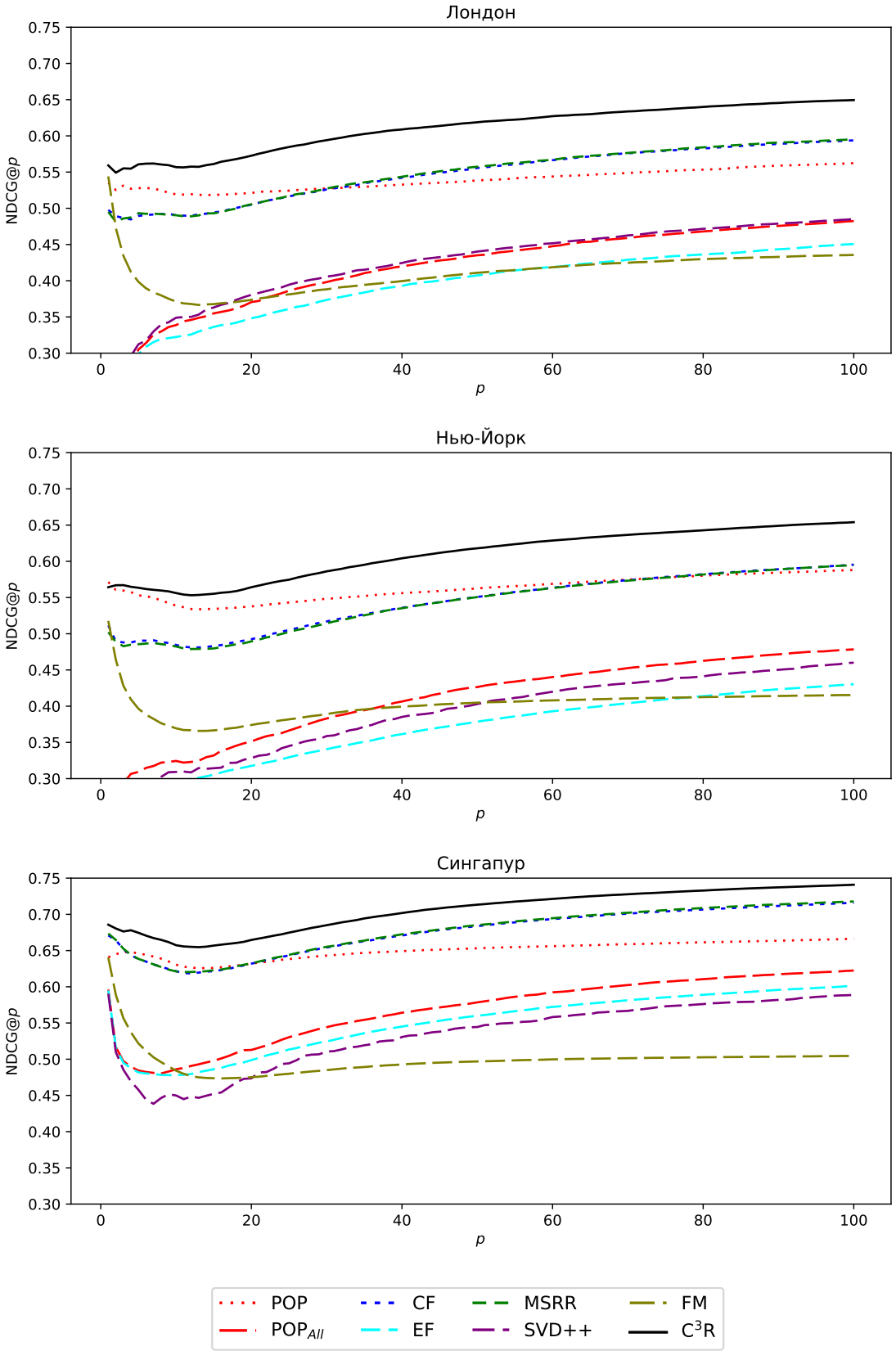

Рисунок 3 – Оценка NDCG для $\mathrm{C}^3\mathrm{R}$ относительно других базовых алгоритмов

4.4. Сравнение модификаций C^{3}R

Для того чтобы ответить на следующий вопрос: позволяет ли информация о взаимоотношениях между источниками данных найти более релевантные сообщества пользователей, мы сравниваем модификации $\mathrm{C}^3\mathrm{R}$, которые включают в себя различные варианты предлагаемого подхода, а также другие подходы к кластеризации. Результаты оценки представлены на рисунке 4 для набора данных NUS-MSS, для TwiSty — на рисунке Б.1.

Во-первых, отметим, что нерегуляризованный вариант подхода ($\mathrm{C}^3\mathrm{R}_{-\hat{L}-L_{mod}}$) имеет худшую эффективность по отношению к другим алгоритмам кластеризации. Что доказывает тот факт, что такой нерегуляризуемый подход к кроссплатформенной кластеризации неспособен создать латентное пространство, которое учитывало бы межсетевые отношения, что приводит к плохой эффективности рекомендации. В то же время подход $\mathrm{C}^3\mathrm{R}$ выиграл все базовые подходы во всех трех городах, чего не удалось достичь с помощью нерегуляризованных версий ($\mathrm{C}^3\mathrm{R}_{-\hat{L}-L_{Mod}}$, $\mathrm{C}^3\mathrm{R}_{-\hat{L}}$, $\mathrm{C}^3\mathrm{R}_{-Comm}$). Вышеупомянутое наблюдение свидетельствует о важности как регуляризации промежуточных пространств с помощью многообразия Грассмана, так и регуляризации отношений между источниками данных для эффективной кластеризации. Это также позволяет нам утверждать, что **информация о взаимоотношениях между источниками данных найти более релевантные сообщества пользователей**.

Также стоит отметить хорошую эффективность других подходов кластеризации. И то что рекомендация, основанная на кластеризации x-means ($\mathrm{C}^3\mathrm{R}$ (x-means)), работает несколько лучше, чем другие базовые подходы к кластеризации. Вышесказанное согласуется с предыдущим исследованием [10] и может быть объяснено возможностью алгоритма автоматически выводить число кластеров, что позволяет находить более характерные сообщества пользователей и, в конечном итоге, предоставлять более эффективную рекомендацию.

4.5. Сравнение различных комбинаций источников

Чтобы получить представление о роли различных источников данных для рекомендации категории места и ответа на вопрос: каков вклад

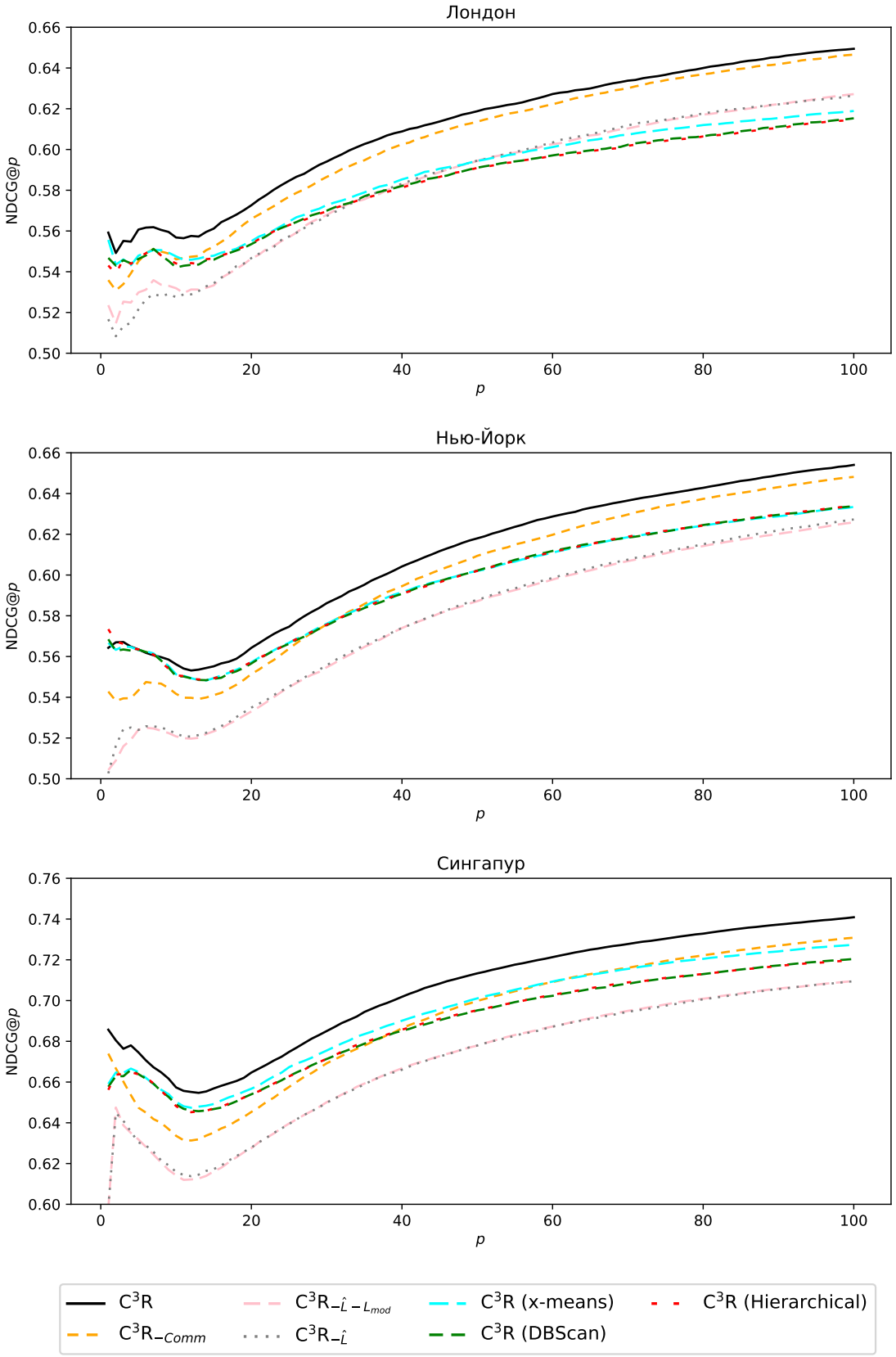

Рисунок 4 – Оценка NDCG для различных модификаций $\mathrm{C^3R}$ и алгоритмов кластеризации

каждого источника данных в рекомендательную систему? Мы оцениваем наш подход на всех комбинациях источников данных. Полученные результаты представлены на рисунке 5. Стоит отметить, что временной слой и слой, связанный с передвижениями пользователей, включены в каждую из комбинаций, например, в случае комбинации «Текст» используются и текстовые данные, и временные данные, и данные о мобильности пользователей.

Неудивительно, что рекомендация, основанная исключительно на распределении категорий мест, наилучшим образом соответствует всем исходным точкам с одним источником во всех трех городах [10]. Причина в том, что данные из целевого источника (Foursquare) содержат явные знания о распределении элементов рекомендации в обучающем наборе и, таким образом, могут точно прогнозировать распределение элементов в тестовой выборке. Результаты комбинации двух источников также дают интересные наблюдения. Например, комбинация «Текст + Геолокация» обеспечивает лучшую рекомендацию в регионах Лондона и Нью-Йорка, но не в Сингапуре, где «Текст» уступает комбинации «Изображения». Возможная причина такого поведения — различия в использовании социальных сетей в разных географических регионах: пользователи из Лондона и Нью-Йорка в основном используют Twitter, а пользователи Сингапура, где Twitter широко не используется, любят делится моментами своей жизни, загружая фотографии в Instagram. Вышесказанное также подтверждается предыдущим исследованием [4], где данные изображения играют решающую роль в задаче кроссплатформенного демографического профилирования в Сингапуре. Наконец, мы также можем отметить, что сочетание всех источников данных обеспечивает наилучшие результаты для всех трех географических регионов, где максимальная эффективность рекомендаций достигается в Сингапуре. Это может быть связано с тем, что в Сингапуре имеется больше открытых данных в используемых социальных сетях по сравнению с Нью-Йорком и Лондоном [4]. Суммируя вышесказанное, мы можем ответить на поставленный ранее вопрос следующим образом: **вклад каждого источника данных сильно зависит от географического региона, в то вре-**

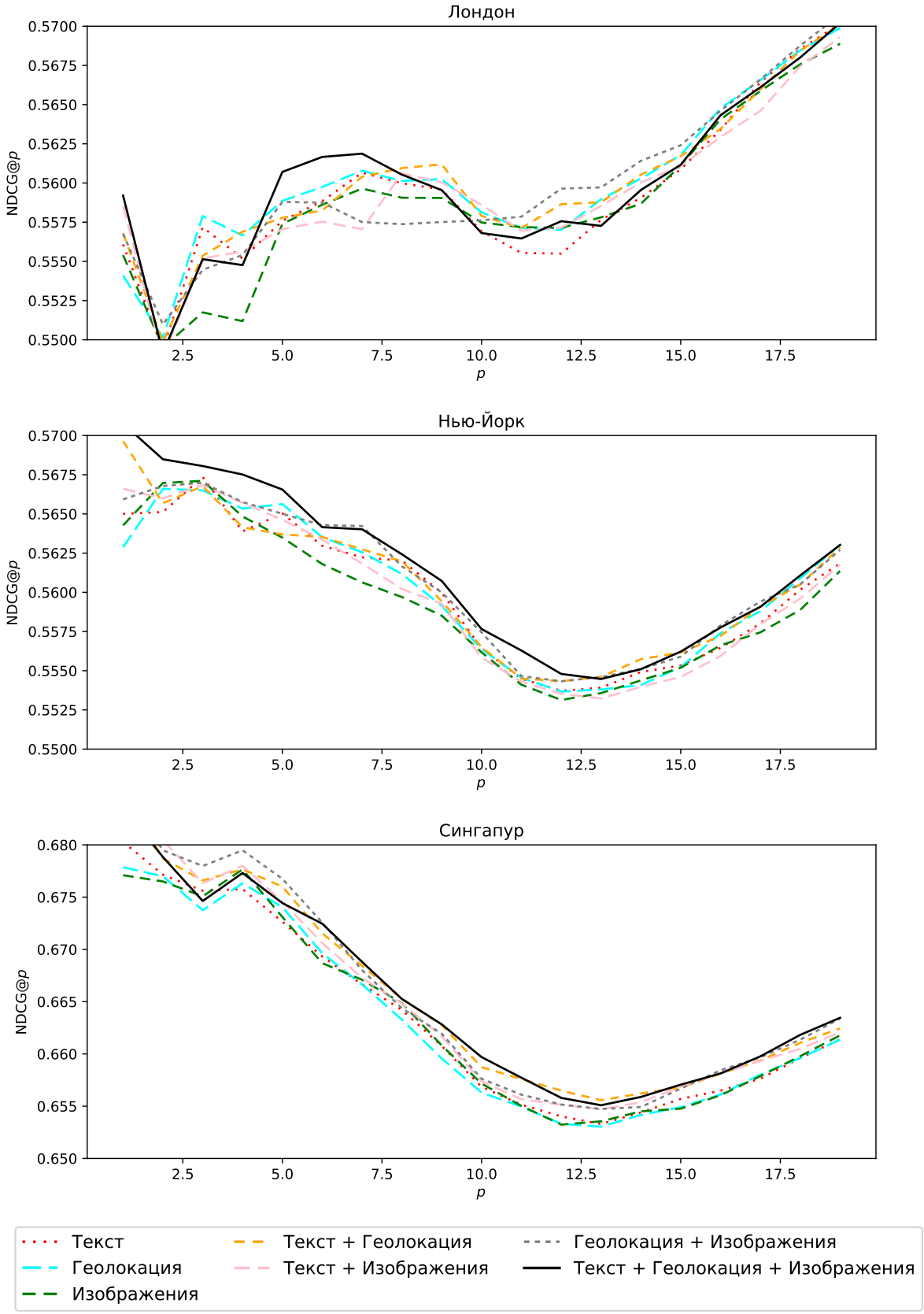

Рисунок 5 – Оценка NDCG для $\mathrm{C}^3\mathrm{R}$ на различных комбинациях источников

мя как, в большинстве случаев комбинация всех источников данных позволяет достичь наилучшей точности рекомендации.

4.6. Обзор получившихся сообществ

Ключевой идеей нашего подхода к рекомендациям является включение групповых знаний в рекомендации путем выявления соответствующих сообществ пользователей из нескольких источников мультимедийных данных. Результаты экспериментов косвенно показывают способность модели $\mathrm{C^3R}$ обнаруживать важные пользовательские сообщества. Чтобы явно продемонстрировать эффективность обнаружения сообществ, в таблице 3 мы перечислили профили нескольких сообществ, обнаруженных в Сингапуре. Пользовательские профили строятся как «мешок слов» (bag-of-words) над текстовыми, визуальными и данными о геолокации. Из таблицы видно, что большинство «слов» во всех модальностях данных согласуются друг с другом и представляют различные пользовательские сообщества. Например, *Сообщество 1* представлено словами: *device*, *launcher*, *android*; визуальными концептами: *mouse*, *digital clock*, *hard disk*; и категориями места: *electroincs store*, *tech startup*, *technology building*. Таким образом, мы можем сказать, что пользователи в данном сообществе интересуются технологиями.

Выводы по главе 4

Качество обнаруженных сообществ позволяет $\mathrm{C^3R}$ осуществлять коллаборативную фильтрацию на основе тех пользователей социальных сетей, которые семантически близки к другим пользователям системы, что помогает повысить эффективность рекомендации. Согласованность профилей сообществ показывает, что $\mathrm{C^3R}$ может выполнять сбалансированную кроссплатформенную кластеризацию. Общие результаты свидетельствуют о применимости предложенных методов для задачи обнаружения кроссплатформенных сообществ и стимулируют дальнейшие исследования в этом направлении.

Таблица 3 – Профили сообществ пользователей

Сообщество	**Мешок слов для разных модальностей**		
	Текст	**Изображения**	**Геолокация**
Сообщество 1 Технологии	device, launcher, android	mouse, digital clock, hard disc	electronics store, tech startup, technology building
Сообщество 2 Искусство	painting, landscape, reflection	obelisk, paintbrush, pencil box	arts & crafts store, arts & entertainment, museum
Сообщество 3 Еда	dining, coffee, cooking	pineapple, microwave, frying pan	italian restaurant, pizzeria, macanese restaurant

ЗАКЛЮЧЕНИЕ

В данной работе было проведено исследование по мультимодальному, кроссплатформенному анализу данных для выявления индивидуальных и групповых атрибутов пользователей социальных сетей. Была предложена модель $\mathrm{C}^3\mathrm{R}$, которая использует как индивидуальные, так и групповые знания для решения задачи рекомендации. Для моделирования групповых знаний данной модели был разработан подход мультимодальной кластеризации из нескольких источников для эффективного поиска сообществ пользователей. Проведя комплексную оценку метода, сравнивая его с передовыми подходами, мы продемонстрировали, что наша модель может обеспечить достаточно эффективную рекомендацию.

Хотя $\mathrm{C}^3\mathrm{R}$ превосходит базовые подходы, есть несколько ограничений, которые мы хотели бы выделить. Во-первых, предлагаемый нами подход к кроссплатформенной кластеризации требует одновременной доступности данных из всех трех социальных сетей. Это не позволяет использовать его для решения проблемы «холодного старта». Но ее возможно решить, заполнив пропущенные данные до процесса кластеризации [3]. Другим ограничением является кубическая сложность подхода $\mathcal{O}(M^2N^3)$. Эта проблема может быть решена с помощью алгоритмов уменьшения размерности пространства [58], на котором будет в дальнейшем выполняться спектральная кластеризация.

За исключением устранения вышеуказанных ограничений, дальнейшая работа над задачей может включать в себя разработку подхода с частичным обучением, который сможет контролировать процесс обнаружения сообществ, используя знания об предметной области. Кроме этого, возможно дальнейшее увеличение числа источников данных и добавлением новых модальностей.

Результаты данной работы были представлены на конференциях ACM Multimedia 2016 [59], ISMW FRUCT 2016 [60] и SIGIR 2017 [61], а также опубликованы в журнале ACM SIGWEB Newsletter [41].

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 *Mander J.* Internet users have average of 5.54 social media accounts. — URL: http://www.globalwebindex.net/blog/internet-users-have-average-of-5-social-media-accounts ; (Дата обращения: 29.04.2017).

2 Understanding Cross-site Linking in Online Social Networks / Y. Chen [et al.] // Proceedings of the 8th Workshop on Soc. Net. Mining and Analysis. — 2014. — P. 6.

3 Interest inference via structure-constrained multi-source multi-task learning / X. Song [et al.] // Proceedings of the International Joint Conference on Artificial Intelligence. — 2015. — P. 2371–2377.

4 Harvesting multiple sources for user profile learning: a big data study / A. Farseev [et al.] // Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. — 2015. — P. 235–242.

5 Towards User Personality Profiling from Multiple Social Networks / K. Buraya [et al.] // Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. — 2017.

6 *Burke R.* Integrating knowledge-based and collaborative-filtering recommender systems // Proceedings of the Workshop on AI and Electronic Commerce. — 1999. — P. 69–72.

7 *Trewin S.* Knowledge-based recommender systems // Encyclopedia of library and information science. — 2000. — Vol. 69, Supplement 32. — P. 180.

8 *Hu R., Pu P.* Helping Users Perceive Recommendation Diversity. // DiveRS@ RecSys. — 2011. — P. 43–50.

9 Diversifying contextual suggestions from location-based social networks / M.-D. Albakour [et al.] // Proceedings of the 5th Information Interaction in Context Symposium. — ACM. 2014. — P. 125–134.

10 Cross-social network collaborative recommendation / A. Farseev [et al.] // Proceedings of the ACM Web Science Conference. — ACM. 2015. — P. 38.

11 *Veiga M. H.*, *Eickhoff C.* Privacy Leakage through Innocent Content Sharing in Online Social Networks // arXiv preprint arXiv:1607.02714. — 2016.

12 *Sadeghi-Niaraki A.*, *Kim K.* Corrigendum “Ontology based personalized route planning system using a multi-criteria decision making approach”[Experts Systems with Applications 36 (2P1)(2009)(1695–1705)] // Expert Systems With Applications. — 2009. — Vol. 36, no. 5. — P. 9604.

13 *Teixeira J. C.*, *Antunes A. P.* A hierarchical location model for public facility planning // European Journal of Operational Research. — 2008. — Vol. 185, no. 1. — P. 92–104.

14 *Балуев Д. Г.* Политическая роль социальных медиа как поле научного исследования // Образовательные технологии и общество. — 2013. — Т. 16, № 2.

15 Demographics of Key Social Networking Platforms / M. Duggan [и др.]. — URL: http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2; (Дата обращения: 30.04.2017).

16 Определение демографических атрибутов пользователей микроблогов / А. Коршунов [и др.] // Труды Института системного программирования РАН. — 2013. — Т. 25.

17 *Miller Z.*, *Dickinson B.*, *Hu W.* Gender prediction on twitter using stream algorithms with n-gram character features. — 2012.

18 Measuring user influence in twitter: The million follower fallacy. / M. Cha [et al.] // Icwsm. — 2010. — Vol. 10, no. 10–17. — P. 30.

19 Harbinger: An Analyzing and Predicting System for Online Social Network Users’ Behavior / R. Guo [et al.] // arXiv preprint arXiv:1312.2096. — 2013.

20 *Fortunato S.* Community detection in graphs // Physics reports. — 2010. — Vol. 486, no. 3. — P. 75–174.

21 Analyzing cross-system user modeling on the social web / F. Abel [et al.] // International Conference on Web Engineering. — Springer. 2011. — P. 28–43.

22 Cross social networks interests predictions based ongraph features / A. Tiroshi [et al.] // Proceedings of the 7th ACM conference on Recommender systems. — ACM. 2013. — P. 319–322.

23 *Yan M.*, *Sang J.*, *Xu C.* Unified youtube video recommendation via cross-network collaboration // Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. — ACM. 2015. — P. 19–26.

24 Cross-domain collaborative learning in social multimedia / S. Qian [et al.] // Proceedings of the 23rd ACM international conference on Multimedia. — ACM. 2015. — P. 99–108.

25 *Su W.-C.* Integrating and mining virtual communities across multiple online social networks: concepts, approaches and challenges // Digital Information and Communication Technology and it's Applications (DICTAP), 2014 Fourth International Conference on. — IEEE. 2014. — P. 199–204.

26 *Rhouma D.*, *Romdhane L. B.* An efficient algorithm for community mining with overlap in social networks // Expert Systems with Applications. — 2014. — Vol. 41, no. 9. — P. 4309–4321.

27 Detecting profilable and overlapping communities with user-generated multimedia contents in LBSNs / Y.-L. Zhao [et al.] // ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). — 2013. — Vol. 10, no. 1. — P. 3.

28 Clustering on multi-layer graphs via subspace analysis on Grassmann manifolds / X. Dong [et al.] // IEEE Transactions on signal processing. — 2014. — Vol. 62, no. 4. — P. 905–918.

29 From Tweets to Wellness: Wellness Event Detection from Twitter Streams. / M. Akbari [et al.] // AAAI. — 2016. — P. 87–93.

30 *Verhoeven B.*, *Daelemans W.*, *Plank B.* TwiSty: a multilingual twitter stylometry corpus for gender and personality profiling // 10th International Conference on Language Resources and Evaluation (LREC 2016). — 2016.

31 *Blei D. M.*, *Ng A. Y.*, *Jordan M. I.* Latent dirichlet allocation // Journal of machine Learning research. — 2003. — Vol. 3, Jan. — P. 993–1022.

32 *Pennebaker J. W.*, *Francis M. E.*, *Booth R. J.* Linguistic inquiry and word count: LIWC 2001 // Mahway: Lawrence Erlbaum Associates. — 2001. — Vol. 71, no. 2001. — P. 2001.

33 *Krizhevsky A.*, *Sutskever I.*, *Hinton G. E.* Imagenet classification with deep convolutional neural networks // Advances in neural information processing systems. — 2012. — P. 1097–1105.

34 Mining user mobility features for next place prediction in location-based services / A. Noulas [et al.] // Data mining (ICDM), 2012 IEEE 12th international conference on. — IEEE. 2012. — P. 1038–1043.

35 *Qu Y.*, *Zhang J.* Trade area analysis using user generated mobile location data // Proceedings of the 22nd international conference on World Wide Web. — ACM. 2013. — P. 1053–1064.

36 Density-based clustering in spatial databases: The algorithm gdbscan and its applications / J. Sander [et al.] // Data mining and knowledge discovery. — 1998. — Vol. 2, no. 2. — P. 169–194.

37 *Eppstein D.*, *Paterson M. S.*, *Yao F. F.* On nearest-neighbor graphs // Discrete & Computational Geometry. — 1997. — Vol. 17, no. 3. — P. 263–282.

38 Group versus individual decision-making: Is there a shift / A. Ambrus, B. Greiner, P. Pathak, [et al.] // Institute for Advanced Study, School of Social Science Economics Working Paper. — 2009. — Vol. 91.

39 Item-based collaborative filtering recommendation algorithms / B. Sarwar [et al.] // Proceedings of the 10th international conference on World Wide Web. — ACM. 2001. — P. 285–295.

40 *Vert J.-P.*, *Tsuda K.*, *Schölkopf B.* A primer on kernel methods // Kernel Methods in Computational Biology. — 2004. — P. 35–70.

41 360° user profiling: past, future, and applications by Aleksandr Farseev, Mohammad Akbari, Ivan Samborskii and Tat-Seng Chua with Martin Vesely as coordinator / A. Farseev [et al.] // ACM SIGWEB Newsletter. — 2016. — Summer. — P. 4.

42 *Shi J.*, *Malik J.* Normalized cuts and image segmentation // IEEE Transactions on pattern analysis and machine intelligence. — 2000. — Vol. 22, no. 8. — P. 888–905.

43 *Wagner D.*, *Wagner F.* Between min cut and graph bisection // Mathematical Foundations of Computer Science 1993. — 1993. — P. 744–750.

44 *Von Luxburg U.* A tutorial on spectral clustering // Statistics and computing. — 2007. — Vol. 17, no. 4. — P. 395–416.

45 *Lutkepohl H.* Handbook of matrices. // Computational Statistics and Data Analysis. — 1997. — Vol. 2, no. 25. — P. 243.

46 *Helgason S.* The Radon transform on Euclidean spaces, compact two-point homogeneous spaces and Grassmann manifolds // Acta Mathematica. — 1965. — Vol. 113, no. 1. — P. 153–180.

47 *Golub G. H.*, *Van Loan C. F.* Matrix computations. Vol. 3. — JHU Press, 2012.

48 *Lehoucq R. B.*, *Sorensen D. C.* Deflation techniques for an implicitly restarted Arnoldi iteration // SIAM Journal on Matrix Analysis and Applications. — 1996. — Vol. 17, no. 4. — P. 789–821.

49 Semantic-based location recommendation with multimodal venue semantics / X. Wang [et al.] // IEEE Transactions on Multimedia. — 2015. — Vol. 17, no. 3. — P. 409–419.

50 X-means: Extending K-means with Efficient Estimation of the Number of Clusters. / D. Pelleg, A. W. Moore, [et al.] // ICML. Vol. 1. — 2000. — P. 727–734.

51 *Renders J.-M.*, *Bersini H.* Hybridizing genetic algorithms with hill-climbing methods for global optimization: two possible ways // Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on. — IEEE. 1994. — P. 312–317.

52 *Hartigan J. A.*, *Wong M. A.* Algorithm AS 136: A k-means clustering algorithm // Journal of the Royal Statistical Society. Series C (Applied Statistics). — 1979. — Vol. 28, no. 1. — P. 100–108.

53 *Li T.*, *Ogihara M.*, *Ma S.* On combining multiple clusterings // Proceedings of the thirteenth ACM international conference on Information and knowledge management. — ACM. 2004. — P. 294–303.

54 *Adomavicius G.*, *Tuzhilin A.* Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions // IEEE transactions on knowledge and data engineering. — 2005. — Vol. 17, no. 6. — P. 734–749.

55 *Winoto P.*, *Tang T.* If you like the devil wears prada the book, will you also enjoy the devil wears prada the movie? a study of cross-domain recommendations // New Generation Computing. — 2008. — Vol. 26, no. 3. — P. 209–225.

56 *Koren Y.* Factorization meets the neighborhood: a multifaceted collaborative filtering model // Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. — ACM. 2008. — P. 426–434.

57 *Rendle S.* Factorization machines with libFM // ACM Transactions on Intelligent Systems and Technology (TIST). — 2012. — Vol. 3, no. 3. — P. 57.

58 *Sajda P.*, *Du S.*, *Parra L. C.* Recovery of constituent spectra using non-negative matrix factorization // Optical Science and Technology, SPIE's 48th Annual Meeting. — International Society for Optics, Photonics. 2003. — P. 321–331.

59 *Farseev A.*, *Samborskii I.*, *Chua T.-S.* bBridge: A Big Data Platform for Social Multimedia Analytics // Proceedings of the 2016 ACM on Multimedia Conference. — ACM. 2016. — P. 759–761.

60 *Самборский И.*, *Механиков Д.*, *Фарсеев А.* Рекомендация на основе мультимодальных данных из нескольких источников // Proceedings of the 2016 ISMW FRUCT Conference. — IEEE. 2016. — С. 94–96.

61 Cross-Network Recommendation Via Clustering On Multi-Layer Graphs / A. Farseev [et al.] // Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval. — ACM. 2017.

ПРИЛОЖЕНИЕ А. СРАВНЕНИЕ С БАЗОВЫМИ АЛГОРИТМАМИ НА НАБОРЕ ДАННЫХ TWISTY

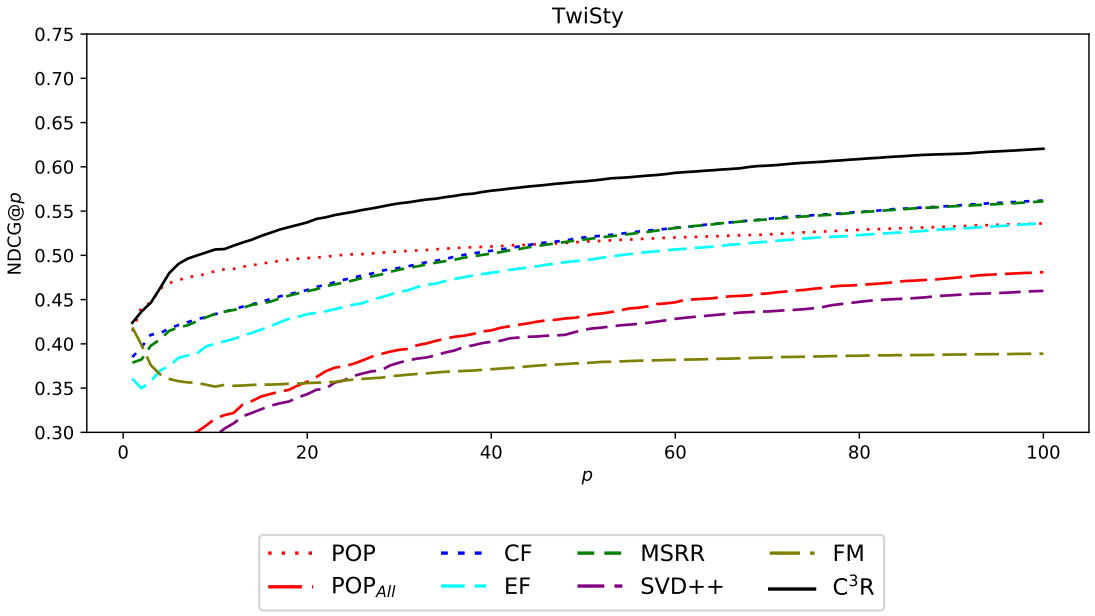

Рисунок А.1 – Оценка NDCG для $\mathrm{C}^3\mathrm{R}$ относительно других базовых алгоритмов на наборе данных TwiSty

ПРИЛОЖЕНИЕ Б. СРАВНЕНИЕ МОДИФИКАЦИЙ C³R НА НАБОРЕ ДАННЫХ TWISTY

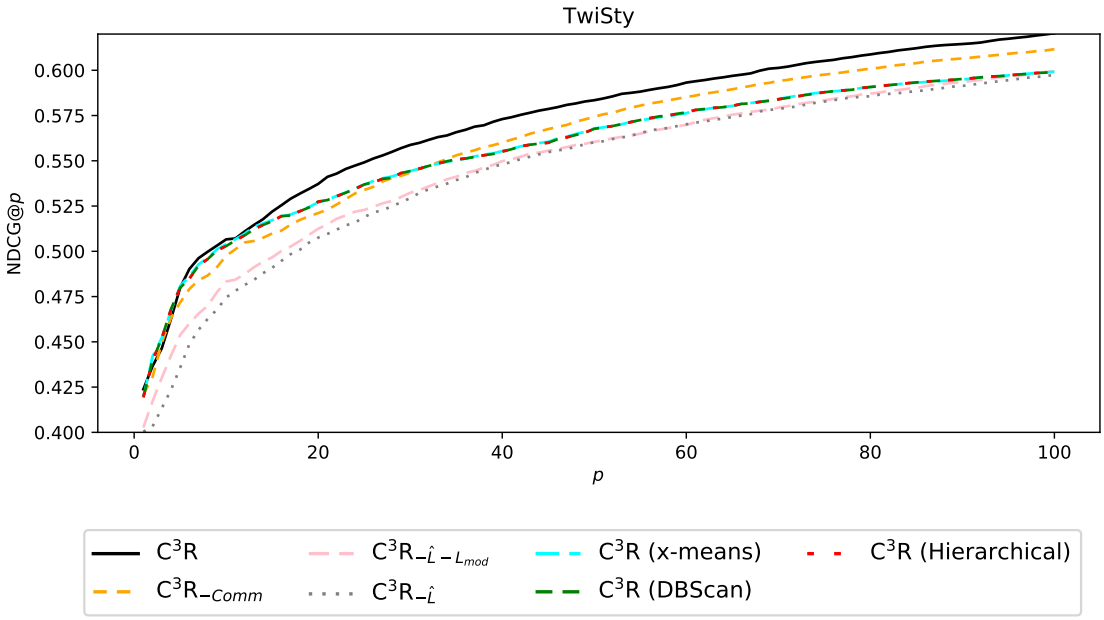

Рисунок Б.1 – Оценка NDCG для различных модификаций $\mathrm{C}^3\mathrm{R}$ и алгоритмов кластеризации на наборе данных TwiSty

ПРИЛОЖЕНИЕ В. КОРРЕЛЯЦИОННАЯ МАТРИЦА

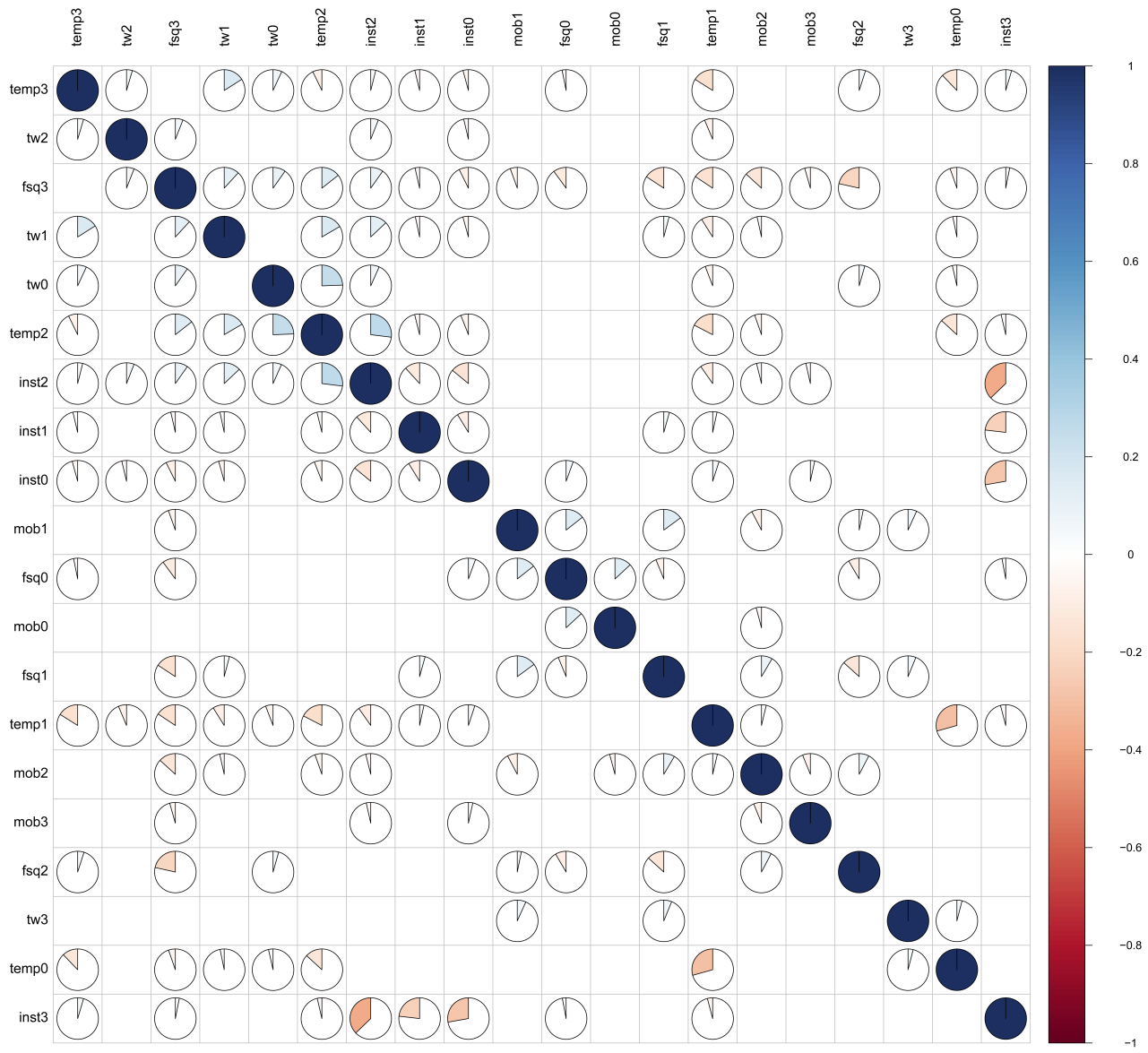

Рисунок В.1 – Матрица корреляций между сообществами