информационные системы

Компьютерные инструменты в образовании, 2014
№ 6: 28–38
УДК: 004.942
http://ipo.spb.ru/journal

РАЗРАБОТКА КРОССПЛАТФОРМЕННОЙ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ НА ОСНОВЕ ИЗВЛЕЧЕНИЯ ДАННЫХ ИЗ СОЦИАЛЬНЫХ СЕТЕЙ

Фарсеев Александр Игоревич,
Жуков Николай Николаевич,
Государев Илья Борисович,
Заричняк Юрий Петрович

Аннотация

Статья содержит:
– структуру и описание реализации кроссплатформенного сбора данных;
–обзор и описание данных, использованных для обнаружения сообществ пользователей;
–описание рекомендательной системы для пользователей различных социальных сетей на основе обнаружения их сообществ.

Ключевые слова: *геопозиционирование, социальная сеть, извлечение данных, анализ сообществ пользователей, кластеризация, кроссплатформенный сбор данных, геоинформационная социальная сеть, геосоциальная сеть.*

В настоящее время происходит рост популярности геосоциальных сетей (location based social networks). В таких сетях одни пользователи разделяют с другими различную информацию: свое местоположение, мнение об объектах (*точках*), опыт посещения каких-либо точек, а также изображения и видеоролики, используя для этого современные возможности мобильных устройств (например смартфонов и планшетов). Пользователь, посещая определенную точку, фиксирует свое местоположение в ней (эта операция обозначается термином *check-in*; далее для простоты будет использоваться термин *чекин*). Пользователям, выполнившим эту операцию в одной и той же точке, доступен просмотр информации друг о друге. Под *точкой* здесь и далее в тексте понимается местоположение и описание объекта (например кафе, площадь в городе, определенное здание). Такую точку можно описать с помощью географических координат и набора релевантной метаинформации. При этом указанная информация и метаинформация становятся доступны одновременно в нескольких социальных сетях за счёт использования ими интерфейса программирования приложений (application programming interface — далее API) и объединяющего идентификатора пользователя (например адрес электронной почты).

Цель данной статьи: описать масштабируемую кроссплатформенную систему рекомендации точек для посещения, работающую в режиме онлайн.

Большинство рекомендательных сервисов анализируют информацию о пользователях на основе *одного* источника [5, 8, 10], однако широко известно, что многие пользователи используют несколько социальных сервисов в их повседневной жизни[1], а значит, актуальна задача создания рекомендательного сервиса, позволяющего анализировать пользовательские данные из *различных* источников, в том числе из социальных сетей, использующих геопозиционирование. Решение такой задачи в комбинации с использованием мобильных приложений позволит создать рекомендательный сервис для пользователей различных социальных сетей, работающий в режиме онлайн. Иными словами, такой сервис позволит производить рекомендацию точек для пользователей с заранее неизвестными геопредпочтениями, что, в свою очередь, позволит решить проблему, известную как *проблема холодного старта* (Cold start problem[2]). Например, для пользователя социальной сети, для которого информация о его геопредпочтениях не известна, рекомендация точек будет производиться на основе его «схожести» с другими пользователями (из данной социальной сети), для которых такая информация может быть получена на основе их активности в геосоциальных сетях.

Однако задача рекомендации точек для посещения не тривиальна. В статье [5] Leung выделил две основные проблемы создания механизма рекомендации точек:

1. Большая вычислительная сложность из-за большой размерности данных.

2. Низкая релевантность профилирования пользователя из-за использования необработанных данных GPS/ГЛОНАСС (долгота, широта).

Относительно первой проблемы установлено, что «наиболее интересные геоинформационные социальные сервисы реализуют механизм рекомендации точек, используя алгоритмы кластеризации данных о физическом расположении пользователей, представленных в матричном виде» [5].

Проблема низкой релевантности профилирования пользователя может быть преодолена за счёт использования геоинформационных социальных сетей, в которых фиксируются данные о каждом действии, совершенном пользователем, в то время как проблема большой вычислительной сложности требует решения для организации рекомендательного сервиса в режиме онлайн. Релевантность профилирования может быть увеличена благодаря использованию дополнительной информации, такой как взаимосвязь между пользователями и сообщениями, к которым они принадлежат, а также дополнительной информации о посещенных пользователями точках (например описание точки, тип точки); пользователи геоинформационных социальных сетей косвенным образом взаимодействуют друг с другом, посещая различные точки и публикуя комментарии и рекомендации о них. Эти разнородные взаимодействия передают скрытую информацию для идентификации значимых и поддающихся интерпретации социальных сообществ, которые проявляют уникальные геоориентированные характеристики. Данные, описывающие каждое такое сообщество, могут быть ранжированы, например, по параметру «наиболее посещаемые точки» и использованы для рекомендации пользователям.

Помимо описанных выше проблем, существует проблема выдачи рекомендаций неизвестным пользователям. Как уже было упомянуто, большинство алгоритмов основываются на предположении, что какая-то часть информации о георасположении пользователя заранее известна. Организация выдачи рекомендаций, основанная на использовании совместной фильтрации в геоинформационных сетях, предполагает анализ данных о пользователях и точках, посещенных им ранее. Однако в большинстве ситуаций такая информация не может быть получена в силу ограничений, заложенных разработчиками таких сервисов или в силу того, что получение такой информации нарушает конфиденциальность пользователя.

[1] http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/.

[2] https://en.wikipedia.org/wiki/Cold_start.

Таким образом, масштабируемый механизм, одновременно использующий алгоритмы кластеризации и информационного поиска, предложенный в статье, опирается на данные двух крупных социальных сетей Foursquare/Swarm[3] (крупнейшая геосоциальная сеть) и Twitter[4] (микроблог с наибольшим числом англоязычных пользователей).

Данные для анализа были предоставлены лабораторией медиа-поиска, входящей в состав школы вычислений при Национальном университете Сингапура (NUS School of Computing Lab of Media Search (LMS))[5], и включают в себя информацию геоинформационной социальной сети Foursquare/Swarm о точках и посещении их пользователями [2]. Структура данных представлена в табл. 1–4.

Всего набор данных содержит более 240 тысяч пользовательских чекинов, выполненных на более чем 6.5 тысячах точек, представляющих более 400 категорий в городе Сингапур. Эта база данных была собрана с использованием API социальной сети Twitter: в течение четырех месяцев отслеживались публичные данные пользователей о посещении различных точек.

Подсказками («tips», см. табл. 4) называют записи, относящиеся к конкретной точке и опубликованные пользователями геоинформационной социальной сети Foursquare/Swarm. Они могут содержать текст и изображения.

Социальная сеть Foursquare/Swarm с помощью использования специального API позволяет получать удобный доступ к данным с возможностью сохранения и последующей их обработки. Для того чтобы обеспечить конфиденциальность пользователей, Foursquare/Swarm ограничивает доступ к данным о посещении пользователями точек так, что доступными для обработки становятся данные только тех пользователей, которые в текущий момент активны.

Табл. 1. Структура описания чекина

Идентификатор чекина	Идентификатор пользователя	Идентификатор точки	Данные
5087d167e4b0ede6cc95b746	5993539	4b0bd124f964a520e03323e3	1351078248000

Табл. 2. Структура описания точки

Идентификатор точки	Название точки	Расположение точки	Категория точки
4b0bd124f964a520e03323e3	Changi International Airport (SIN)	"latitude" : 1.356, "longitude" : 103.988	4bf58dd8d48988d1ed 931735

Табл. 3. Структура описания категорий точек

Идентификатор категории	Название категории
4bf58dd8d48988d1ed931735	Airport

Табл. 4. Структура описание подсказок для точек

Идентификатор подсказки	Идентификатор пользователя	Идентификатор точки	Текст
4961bc4770c603bbff0c8eb4	5993539	4b0bd124f964a520e03323e3	Check out Tribeca…

[3] https://ru.foursquare.com.
[4] https://twitter.com.
[5] http://lms.comp.nus.edu.sg.

Однако, благодаря тому, что Foursquare/Swarm позволяет связывать учетные записи в данной сети с учетными записями в других социальных сетях (в частности, с системой микроблогов Twitter), существуют альтернативные пути получения данных о пользователях и их чекинах.

При выполнении чекина в определенной точке первым действием (1) является отправка широты и долготы точки на сервер Foursquare/Swarm. Дальнейшие действия таковы: (2) сравнение данных, полученных от пользователя и находящихся в базе данных Fousquare/Swarm, (3) выдача списка точек, находящихся поблизости и упорядоченных по возрастанию расстояния до точки, (4) выбор точки, в которой пользователь находится, и (5) отправка информации обратно на сервер Foursquare/Swarm.

После выполнения чекина пользователю доступна возможность опубликовать информацию о его местоположении в социальных сетях Twitter и Facebook[6] [4].

Заметим также, что сервис Twitter обеспечивает возможность доступа к пользователям и их сообщениям («твитам») с помощью специального потокового программного интерфейса, позволяющего, например, получить все твиты, совпадающие с ключевым словом или словами, заданными в поисковом запросе.

Признаком твитов, имеющих отношение к чекинам, является короткая ссылка на оригинальную веб-страницу сервиса Foursquare/Swarm с данными о пользователе, совершившем чекин (идентификатор, время и название точки).

Для того чтобы реализовать кроссплатформенную рекомендательную систему, помимо основных структур данных, указанных выше, необходимо собрать дополнительные данные. Одним из наиболее важных процессов на этом этапе является отображение (сопоставление) идентификаторов пользователей социальной сети Foursquare/Swarm и Twitter, в результате которого существенно увеличивается объем информации, доступный для использования при рекомендации. Данные для этого этапа могут быть получены через интерфейс (API) социальной сети Foursquare/Swarm, которая предоставляет информацию об учетных записях пользователей в других системах.

Процесс отображения (сопоставления) пользователей на основе имеющихся данных из социальных сетей Twitter и Foursquare/Swarm включил анализ более 1.5 миллионов соответствующих твитов. Для реализации этой процедуры использовался Twitter REST API[7].

Структура данных, полученных в результате анализа, представлена следующими полями:

– идентификатор твита;

– имя пользователя, опубликовавшего твит;

– ключевые слова (ключевое слово часто называют «хэштег» из-за символа «#»);

– текст твита.

Для вычисления схожести между пользователями необходимо организовать попарный анализ их профилей, однако данная операция с учетом большого количества пользователей имеет квадратичную вычислительную сложность ($O(n^2)$). Эта особенность усложняет вычисления для большого количества пользователей в режиме онлайн. Устранить этот недостаток возможно, организовав отбор «ключевых пользователей» для каждого сообщества. Термин *сообщество* в данной статье обозначает группу пользователей конкретной социальной сети, которых можно объединить по какому-либо критерию или набору критериев. Термином *ключевой* мы обозначаем тех пользователей, которые имеют высокую значимость в данном сообществе и активность которых схожа с активностью других пользователей из данного сообщества, то есть может служить разумной аппроксимацией показателя активности для всего сообщества. Наиболее активные пользователи в сообществе имеют высо-

[6] https://www.facebook.com.

[7] https://dev.twitter.com/rest/public.

кий показатель коэффициента «значимости». Формула и описание вычисления показателя «значимости» дана ниже.

Как было сказано ранее, для осуществления качественного сравнения интересов пользователей между собой необходимо иметь достаточный объем данных, сгенерированных пользователями в результате их деятельности в социальных сетях. В то же время, для обеспечения более полного описания точек был организован поиск и сбор данных с веб-сайтов этих точек (при наличии ссылки на них в сети Foursquare/Swarm), с веб-сайта свободной энциклопедии «Википедия», с веб-сайта сервиса Foursquare/Swarm; также использовался API поисковой системы Google. Решение задачи сбора дополнительных данных детальнее рассмотрено в статье Zhu [12].

При создании рекомендательной системы для точек, представляющих заведения, использовался подход, основанный на анализе сходства текста и предложенный Spaeth в работе [9]. В основу проектирования системы был положен принцип *масштабируемости*, то есть сохранения работоспособности при росте рабочей нагрузки. При одновременной работе с увеличением количества пользовательских сообществ или объема данных о пользователях сложность операций поиска и анализа похожих пользователей линейно возрастает, что делает этот процесс в больших базах данных крайне медленным. Для преодоления этой проблемы необходимо использовать алгоритм кластеризации пользователей социальной сети Foursquare/Swarm, позволяющий снизить размерность данных. Получившиеся кластеры профилируются при помощи наиболее посещаемых типов точек, отобранных среди всех зафиксированных чекинов. В результате для конкретного пользователя формируется рекомендация категорий и их точек.

С учетом вышеизложенного рекомендательная система проектировалась, исходя из разделения на онлайн и офлайн компоненты. В режиме офлайн система должна выделять пользовательские сообщества и ключевых пользователей в них, а в режиме онлайн осуществлять рекомендацию. Далее в тексте следует более подробное описание функций указанных компонент.

Механизм *выделения пользовательских сообществ* основан на кластеризации двудольного графа «пользователь-точка» (см. рис. 1), в котором две точки считаются равными друг другу, если они принадлежат к одному узлу графа в дереве категорий точек геоинформационной социальной сети Foursquare/Swarm (например, обе точки являются ресторанами)[8].

Этот двудольный граф является ненаправленным и содержит два возможных типа вершин «точка» и «пользователь», а также один тип рёбер «пользователь-точка». Каждое ребро характеризуется весом w и представляет пользовательский чекин в какой-либо точке C_k с количеством посещений w. Каждый пользователь представлен вектором целых чисел длиной n (обозначающих число возможных категорий вершин) $[c_1, c_2, c_3, ..., c_n]$, где каждое значение отражает количество чекинов в определенной категории точек c_k для конкретного пользователя u_i. Если соединить векторы-строки характеристик для всех пользователей вместе и построить двумерный вектор, будет получена матрица смежности. Значения в этой матрице идентифицируют количество посещений точки конкретным пользователем. Мы определяем *сообщества пользователей* как *сильно связанные компоненты* указанного графа.

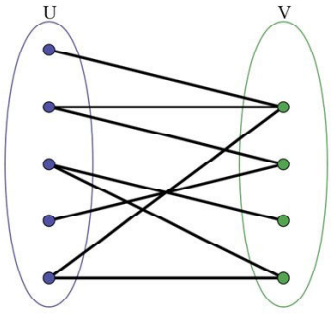

Рис. 1. Двудольный граф «пользователь-точка»

Для *выделения пользовательских сообществ был* запущен процесс *кластеризации пользователей*, который предполагает объединение схожих пользователей в кластеры на основе опубликованных ими данных. Для кластеризации были исполь-

[8] https://developer.foursquare.com/categorytree.

зованы алгоритмы, в которых отсутствует необходимость указания количества кластеров, на которые должны быть распределены данные: X-Means [7], EM-алгоритм [6], алгоритм DBScan (Density-based clustering) [3]. Кластеризация выполнялась на матрице смежности «пользователь-точка». В качестве основных алгоритмов, выбранных для решения задачи кластеризации и показавших наиболее качественный результат, были использованы алгоритмы X-means и EM-алгоритм (Expectation maximization algorithm).

В результате проведения кластеризации формируются пользовательские сообщества, описываемые с помощью списков наиболее популярных категорий точек среди пользователей этих сообществ, что позволяет отразить *интересы* участников каждого сообщества.

С целью *определения ключевых пользователей сообществ* была сформулирована функция, позволяющая вычислить значимость пользователя в сообществе, к которому он принадлежит. Показатель влиятельности вычисляется для всех участников каждого сообщества (кластера), после этого формируются списки, отсортированные по убыванию, и первые k пользователей из них рассматриваются как ключевые (в зависимости от требований к скорости вычисления k может принимать значения от 1 до m, где m — мощность множества пользователей, принадлежащих к данному сообществу). Формула вычисления значимости пользователя в сообществе, к которому он принадлежит A_{u_i}, выглядит следующим образом:

$$A_{u_i} = \alpha \frac{c_{u_i}}{\sum C_{u_i}} + \beta \frac{f_{u_i} + 1}{\sum f_{u_i}},$$

где c_{u_i} — количество чекинов, релевантных сообществу в котором состоит пользователь u_i, f_{u_i} — количество друзей пользователя u_i, находящихся в одном с ним сообществе, α, β — эмпирически вычисленные весовые коэффициенты ($\alpha + \beta = 1$).

Для обеспечения фильтрации нерепрезентативной информации необходима предобработка данных, цель которой — исключить пользователей с чрезмерно большим и чрезмерно низким показателем посещаемости различных точек. Во всех вычислениях фигурировала информация о пользователях, которые совершили более 3 чекинов за отведенный период времени. Данные о пользователях с чрезмерно низким показателем посещаемости не отражают реальных интересов этих пользователей, а их наличие в исследуемых данных может негативно сказаться на результатах работы алгоритмов кластеризации и последующей рекомендации.

Процесс предобработки данных включает в себя также усреднение показателя посещаемости для каждой категории точек и реализуется с помощью вычисления отношения количества чекинов для каждой категории к общему числу чекинов, которые были совершены пользователем. В результате матрица «пользователь-точка» преобразуется в двумерный массив, значения в котором представляют собой отношения количеств посещений пользователя к общему количеству пользовательских чекинов.

Анализ пользователей социальной сети Twitter осуществлялся на основе их сходства с другими пользователями этой же социальной сети, которые уже принадлежат к определенному сообществу. Для оценки сходства пользователей использовалось более 400 признаков (фич), полученных на основе анализа текстовых данных. Алгоритм их получения состоит из следующих шагов:

1. Для каждого объекта категории c_j социальной сети Foursquare/Swarm выбирается k-список наиболее часто встречаемых терминов (термины категории обозначены w_j) из словаря d_j, которые представляют данную категорию (см. рис. 2). Словарь d_j сформирован на основе данных из Wikipedia, Google, Foursquare/Swarm, собранных ранее.

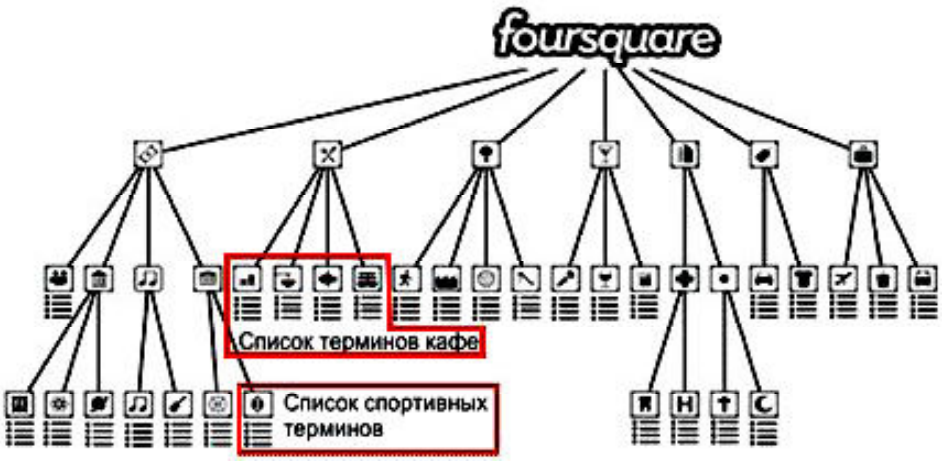

Рис. 2. Пример выбора терминов для категорий в Foursquare/Swarm

2. Для каждого пользователя u_i социальной сети Twitter подсчитывается количество терминов в опубликованных ими сообщениях, которые содержатся в множестве терминов каждой категории.

3. Для каждого пользователя социальной сети Twitter u_i строится вектор $W_1 = [w_1, w_2, ..., w_n]$, в котором каждое значение — целое число, соответствующее количеству вхождений термина из множества категорий w_j в сообщения, опубликованные пользователем u_i.

В качестве меры сходства объектов был выбран коэффициент Охайи. Главными его преимуществами для данной задачи являются простота и эффективность. Приведем формулу для его вычисления:

$$S_{cosine} = \frac{|W_i \cap W_j|}{\sqrt{|W_i||W_j|}},$$

где W_i и W_j — векторы, сформированные для пользователей u_i и u_j соответственно и содержащие количества вхождений терминов, полученных на шаге 3 и описанных выше.

Мера сходства вычисляется между конкретным пользователем u_q и ключевыми пользователями из всех выделенных сообществ пользователей, информация о которых содержится в базе данных и определяет пользователя с наибольшим показателем S_{cosine} как наиболее похожего на данного пользователя (см. рис. 3).

Далее рекомендация осуществляется на основе профиля, составленного для сообщества, к которому принадлежит наиболее похожий на ключевой пользователь. Данная идея

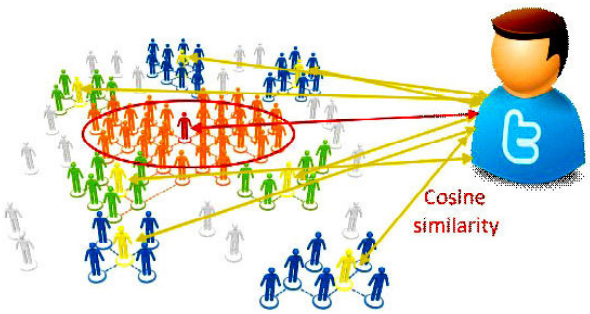

Рис. 3. Сравнение пользователя с ключевыми пользователями выделенных сообществ

основана на предположении о том, что выбранные ключевые пользователи могут объективно представлять определенные сообщества. Такой подход нацелен на максимальное увеличение скорости работы рекомендательной системы без значительного снижения точности выдаваемых рекомендаций.

Использование описанных выше компонентов и построенных моделей позволяет реализовать рекомендательную систему для сообществ и площадок.

Схема, представленная на рис. 4, реализует описанную выше структуру и может быть разделена по принципу работы на офлайн-часть (то есть часть, где постоянный доступ в сеть Интернет не требуется) и онлайн-часть (для работы которой необходимы постоянное подключение к сети Интернет и доступ к API социальных сетей).

Офлайн-часть решает следующие задачи:

1. Сбор данных.
2. Определение сообществ в социальной сети Foursquare/Swarm.
3. Определение ключевых пользователей.

Онлайн-часть решает следующие задачи:

1. Определение идентичности пользователей социальной сети Twitter.
2. Рекомендация для сообществ и площадок, которая базируется на соответствующем профиле пользователя определенного сообщества (соответствие профиля социальной сети Foursquare/Swarm и Twitter может быть найдено с помощью специальной таблицы).

В реализации системы хранение данных осуществлялось с использованием NoSQL-решения на основе MongoDB. Библиотека машинного обучения Accord.net и пакет Knime[9] с

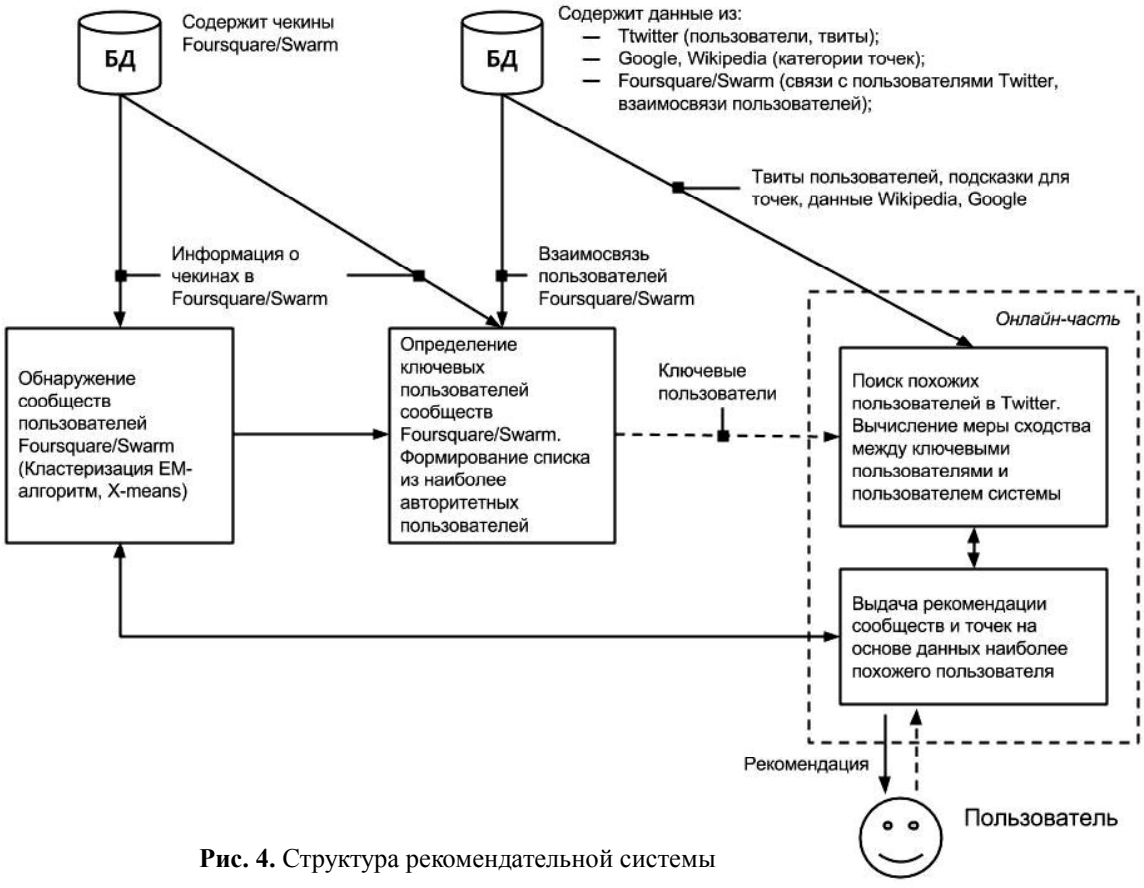

Рис. 4. Структура рекомендательной системы

[9] https://www.knime.org.

установленным расширением Weka использовались для построения модели и кластеризации данных соответственно. Система была запущена на сервере, расположенном в школе вычислений при Национальном университете Сингапура (NUS School of Computing).

Для оценки работы рекомендательной системы использовалась k-кратная перекрестная проверка (k-fold cross-validation) точности. Чтобы оценить производительность работы рекомендательной системы, была использована усредненная пониженная величина совокупного прироста релевантности (normalized discounted cumulative gain — NDCG)[10], которая характеризует качество рекомендательной системы, основанной на ранжировании релевантности рекомендованных сущностей.

Результаты оценки работы рекомендательной системы с использованием метрики *NDSC@k* (где k — количество рекомендуемых сущностей) приведены ниже (см. рис. 5). Оценивались алгоритмы кластеризации K-means, X-means, EM, DBScan. Из результатов эксперимента можно отметить, что наибольший прирост точности показывают алгоритмы EM и X-Means при значениях k равных 6 и 7. Для $k > 7$, а также при $k < 3$ значения метрик уменьшаются. Данная особенность может быть связана с невысоким разнообразием интересов пользователей (как правило, пользователи имеют сравнительно небольшое количество интересов относительно типов мест, которые они посещают регулярно, в то время как остальные места отображают случайные посещения мест и, следовательно, не могут быть расценены как реальные предпочтения пользователей).

Анализ результатов эксперимента приводит к выводу о том, что для организации кроссплатформенной системы рекомендации наиболее оптимальным будет использование алгоритмов EM и X-Means. При значении $k = 6$ (количество рекомендуемых категорий) достигается наилучшая интерпретируемость списка рекомендаций с приемлемой точностью. С учетом того, что рекомендация осуществляется в режиме онлайн, важным является показатель времени работы алгоритма. Так как время работы EM-алгоритма примерно в 10 раз дольше, чем для X-means алгоритма (10 часов и 1 час соответственно), *наиболее оптимально использование алгоритма X-means*. Представленные результаты также показывают, что использование данных из различных источников позволяет решать задачу рекомендации с

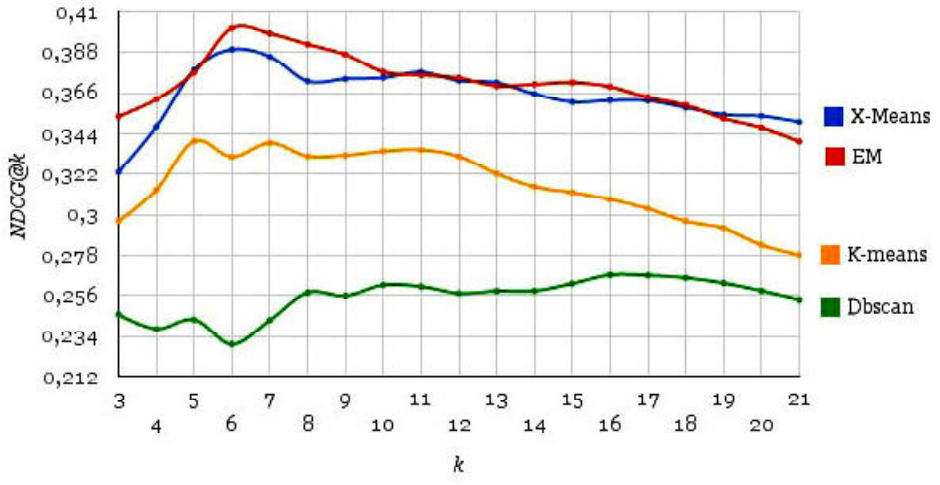

Рис. 5. Точность работы рекомендательной системы, рассчитанная с использованием *NDSC@k*

[10] http://www.kaggle.com/wiki/NormalizedDiscountedCumulativeGain.
[11] https://www.flickr.com
[12] http://instagram.com

повышенной точностью, а значит, должно быть рассмотрено более детально в дальнейших исследованиях.

В данной работе нами был предложен подход к реализации кроссплатформенной системы, нацеленной на формирование рекомендаций пользователям категорий точек для посещения. В этом подходе были использованы алгоритмы кластеризации и механизмы выборки ключевых пользователей сообществ, в то время как итоговая рекомендация производилась с использованием анализа статистик посещения мест членами сообществ пользователей. Это позволило снизить нагрузку на ресурсы при определении объединяющих факторов сообществ, а следовательно, повысить общую производительность системы. Кроссплатформенное решение позволило осуществлять рекомендацию для пользователей с заранее неизвестными предпочтениями, что даёт вариант решения проблемы холодного старта. Результаты, полученные в ходе испытания реализованной системы, позволяют сделать выводы о качественном решении поставленных перед ней задач с помощью описанных в работе техник.

Дальнейшая работа в этой области может заключаться в реализации механизма сбора и анализа данных из сервисов Flickr [11] и Instagram [12]. Кроме того, планируется учитывать структуру социальных связей пользователя, представляя её в виде графа. Данный подход был описан в исследовании Zhao [11]. Повышение эффективности поиска объединяющих факторов сообществ возможно за счёт использования модели LDA (latent dirichlet allocation) [1] в рамках решения задачи конструирования наборов пользовательских ключевых слов.

Литература

1. *Blei D.M., Ng A.Y., Jordan M.I.* Latent dirichlet allocation // Journal of machine Learning research, 2003. Т. 3. С. 993–1022.
2. *Chua T.S. et al.* NExT: NUS-Tsinghua Center for Extreme Search of User-Generated Content // MultiMedia, IEEE, 2012. Т. 19, №. 3. С. 81–87.
3. *Ester M. et al.* A density-based algorithm for discovering clusters in large spatial databases with noise // Kdd. 1996. Т. 96. С. 226–231.
4. *Lee M.J., Chung C.W.* A user similarity calculation based on the location for social network services // Database Systems for Advanced Applications. Springer Berlin Heidelberg, 2011. С. 38–52.
5. *Leung K.W.T., Lee D.L., Lee W.C.* CLR: a collaborative location recommendation framework based on co-clustering // Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011. С. 305–314.
6. *North B., Blake A.* Learning dynamical models using expectation-maximisation // Computer Vision, 1998. Sixth International Conference on. IEEE, 1998. С. 384–389.
7. *Pelleg D. et al.* X-means: Extending K-means with Efficient Estimation of the Number of Clusters // ICML. 2000. С. 727–734.
8. *Raad E., Chbeir R., Dipanda A.* User profile matching in social networks //Network-Based Information Systems (NBiS), 2010 13th International Conference on. IEEE, 2010. С. 297–304.
9. *Spaeth A., Desmarais M.C.* Combining Collaborative Filtering and Text Similarity for Expert Profile Recommendations in Social Websites // User Modeling, Adaptation, and Personalization. Springer Berlin Heidelberg, 2013. С. 178–189.
10. *Wang H., Terrovitis M., Mamoulis N.* Location recommendation in location-based social networks using user check-in data // Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2013. С. 364–373.
11. *Zhao Y.L. et al.* Detecting profilable and overlapping communities with user-generated multimedia contents in LBSNs // ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP). 2013. Т. 10. №. 1. С. 3.
12. *Zhu X. et al.* Topic hierarchy construction for the organization of multi-source user generated contents // Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013. С. 233–242.

CROSS-PLATFORM ONLINE VENUE AND USER COMMUNITY RECOMMENDATION BASED UPON SOCIAL NETWORKS DATA MINING

Farseev A. I., Zhukov N. N., Gossoudarev I. B., Zarichnyak Yu. P.

Abstract

The article contains:
– the structure and description of the implementation of cross platform data mining;
– the overview and the description of data used for user communities detection;
– the description of the recommendation system for users of different social networks based on community detection.

Keywords: *location based social network, data mining, machine learning, clustering, cross platform data gathering.*

Фарсеев Александр Игоревич,
аспирант кафедры информационных систем и программного обеспечения РГПУ им. А.И. Герцена,
farseev@u.nus.edu

Жуков Николай Николаевич,
ассистент кафедры информационных и коммуникационных технологий РГПУ им. А.И.Герцена,
zhukov@herzen.spb.ru

Государев Илья Борисович,
кандидат педагогических наук,
доцент кафедры информационных и коммуникационных технологий РГПУ им. А.И.Герцена,
gossoudarev@herzen.spb.ru

Заричняк Юрий Петрович,
доктор физико-математических наук,
профессор кафедры компьютерной теплофизики и энергофизического мониторинга СПбНИУ ИТМО,
zarich@grv.ifmo.ru