Кластеризация профилей пользователей в рекомендательных системах поддержки жизнеобеспечения на основе реальных неявных данных

© С. А. Филиппов © В. Н. Захаров © С. А. Ступников
© Д. Ю. Ковалев
Институт проблем информатики ФИЦ ИУ РАН,
Москва
stanislav@philippov.ru VZakharov@ipiran.ru ssa@ipi.ac.ru
dm.kovalev@gmail.com

Аннотация

Данная работа посвящена описанию решения ключевой задачи в контексте построения рекомендательных систем поддержки жизнеобеспечения. Этой задачей является выявление пользовательских предпочтений и формализация их посредством формирования поведенческих профилей с последующим выявлением групп пользователей со схожими характеристиками. Основным источником информации о пользовательских предпочтениях является массив неявно собираемых данных об их действиях при навигации по страницам Интернет магазинов. Под поддержкой жизнеобеспечения понимается круг задач по обеспечению населения необходимыми для их жизнедеятельности продуктами, включая продукты питания, бытовой химии, косметики и многое другое. Эти задачи, как правило, решают магазины (в том числе и интернет-магазины) оптовой и розничной торговли. Авторами работы представлен подход к построению рекомендательной системы, основанный на решении задачи коллаборативной фильтрации с использованием методов кластерного анализа данных для выявления групп пользователей со схожими предпочтениями. Достоинства решения продемонстрированы на примере тестового массива данных, полученного из действующего интернет-магазина Thaisoap. Работа выполнена при поддержке Министерства образования и науки РФ, уникальный идентификатор проекта RFMEFI60414X0139.

Введение

Современная электронная коммерция, сопровождающая и поддерживающая процессы жизнеобеспечения, активно использует рекомендательные системы для решения задач адресного продвижения товаров и услуг с учетом конкретных пользовательских предпочтений. Основным источником информации о пользовательских предпочтениях являются данные об активности пользователей при посещении конкретного интернет-ресурса. Эти данные собираются в основном неявным образом (протоколирование действий пользователей) и обладают следующими основными свойствами: значительный объем и быстрое изменение (или обновление) данных во времени. При этом адаптация под конкретного пользователя – весьма сложная задача, поскольку для ее решения необходимо принимать во внимание как присущие человеку неопределенность и спонтанность в рамках конкретного интернет-ресурса, так и множество неопределенностей, связанных с особенностями функционирования Интернет.

Одним из простейших подходов к выработке рекомендаций является использование статистических метрик для выявления, например, наиболее популярных, дешевых (дорогих), близких по заданным характеристикам объектов и предложение их пользователям без учета их персональных предпочтений. Более сложные алгоритмы выявляют предпочтения пользователей посредством формирования поведенческого профиля пользователя, который, в свою очередь, определяется на основании анализа его активности при выборе товаров и услуг. Также в настоящее время развиваются подходы, основанные на использовании нечеткой логики, позволяющей учитывать различные типы неопределенностей и кластеризовать пользовательские профили [1].

Самым распространенным подходом при реализации рекомендательных систем в электронной коммерции является метод коллаборативной фильтрации (collaborative filtering). Данный подход позволяет вырабатывать рекомендации, основанные на модели предшествующего поведения

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

пользователей с учетом поведения других пользователей со сходными характеристиками [2]. Для выделения групп пользователей со сходными характеристиками, как правило, используются различные методы извлечения данных (data mining), которые, в свою очередь, используют алгоритмы кластеризации. В частности, в работе [3] указывается, что решение вопросов идентификации групп пользователей по своей природе опирается на использование методов кластеризации. Кластеризация данных может быть также использована для генерации профилей пользователей на основе информации о действиях каждого пользователя, а затем для формирования групп пользователей на основе их профилей.

Основной задачей, которую авторы данной работы ставили перед собой, является разработка комбинированного подхода к построению рекомендательных систем, обеспечивающего наиболее полное использование всех данных о посетителях интернет-магазинов с целью выработки рекомендаций наиболее адекватно отражающих их ожидания (пертинентность предложения). Научно практическая новизна работы заключается в идее комбинированного использования методов Item-Item CF и User-User CF, что позволяет минимизировать недостатки каждого из них и добиться более высокого качества работы рекомендательной системы в целом. Данная статья входит в серию статей по данной проблематике и посвящена решению задачи коллаборативной фильтрации User-User CF с использованием методов кластеризации для выявления групп пользователей со схожими предпочтениями. Предполагается, что при наличии качественных данных о пользовательской активности метод User-User CF дает наиболее адекватные прогнозы. Использование методов кластеризации для решения задачи выявления групп пользователей со схожими характеристиками позволяет добиться хорошего быстродействия и качества работы алгоритма.

1 Персонализация контента

Практически все современные интернет-ресурсы, ориентированные на работу с большим количеством пользователей, собирают информацию об их активности и анализируют (обрабатывают) ее с целью персонализации своего контента для каждого конкретного пользователя. В качестве наиболее характерных примеров можно выделить:

1. поисковые машины, которые собирают и систематизируют информацию о страницах в сети Интернет заинтересовавших конкретных пользователей;
2. интернет-магазины, которые собирают и систематизируют сведения о предпочтениях своих пользователей в части товаров и услуг;
3. форумы, интернет-дневники и социальные сети, которые собирают информацию о том, в каких тематических разделах и группах принимает участие каждый пользователь и насколько активно.

Другими достаточно распространенными примерами протоколирования действий пользователей являются счетчики посещений страниц, прокси-серверы и интернет-провайдеры.

Собранные данные о пользовательской активности характеризуются большим объемом и разнородностью. Традиционные базы данных малоприменимы для работы с этими данными по причине больших объемов данных и повышенных требований к производительности [4]. Как правило используются так называемые NoSQL системы управления данными (HBase, Cassandra). Их характерными особенностями являются отказ от транзакций, практически линейная масштабируемость, высокая скорость обработки запросов, отсутствие жесткой схемы данных.

В контексте проблемы персонализации контента (а также прогнозирования, выявления предпочтений и групп схожих ресурсов) встает задача обработки этих данных и выявления определенных закономерностей, позволяющих сделать выводы о конкретных предпочтениях пользователей. Таким образом, основной целью обработки данных о пользовательской активности является извлечение полезной информации, которая может, в свою очередь, использоваться для решения следующих задач [5]:

1. Кластеризация ресурсов. Группирование схожих по множеству посетителей ресурсов в несколько кластеров (групп) ресурсов. Кластеризация позволяет строить каталоги ресурсов, а также выявлять недостатки существующих тематических каталогов.
2. Кластеризация пользователей. Группирование схожих пользователей в кластеры аналогично кластеризации ресурсов. Позволяет выявлять группы пользователей со схожими интересами.
3. Построение устойчивых поведенческих профилей пользователей в виде перечня групп ресурсов, посещаемых как данным пользователем, так и схожими с ним пользователями.
4. Построение расширенных профилей пользователей, включающих социально-демографические данные (анкеты), описательные статистики и поведенческие профили. Расширенные профили позволяют классифицировать новых пользователей, выявлять зависимости между пользовательским поведением и социально-демографическими характеристиками.
5. Сегментация клиентской базы на основе расширенных профилей позволяет выделять сегменты, как по анкетным данным клиентов, так и по их поведению. Эта

информация используется при маркетинговых исследованиях.

6. Прямой маркетинг. Предоставление рекламы и маркетинговых предложений конкретному пользователю на основе его поведенческого профиля.
7. Персонализация контента. Представление каждому пользователю сайта наиболее интересной для него информации в наиболее удобном для него виде. Знание информационных предпочтений пользователя позволяет динамически перестраивать контент сайта.
8. Построение карт сходства ресурсов и пользователей. Позволяет отображать множества наиболее посещаемых ресурсов и наиболее активных пользователей в виде точечного графика. Схожим ресурсам (пользователям) соответствуют близкие точки на карте. Карту сходства можно использовать как графическое средство навигации.

Существуют различные методы и подходы, используемые на практике при решении перечисленных выше задач. Весь класс этих методов принято называть методами коллаборативной фильтрации.

2 Коллаборативная фильтрация

Результатом первого этапа обработки данных о пользовательской активности является построение матрицы активности, которая может нести различную информацию о действиях пользователя. Это может быть бинарная информация о посещении или не посещении заданного ресурса данным пользователем, частота (или число) пользований ресурса r пользователем u, стоимость или рейтинг, проставленный пользователем u для ресурса r и т.д. Для оценки степени схожести пользователей в плане их предпочтений и построения поведенческих профилей могут использоваться различные функции сходства (метрики). Наиболее популярными среди них являются: косинусная мера, коэффициент корреляции Пирсона, евклидово расстояние, коэффициент Танимото, Манхэттенское расстояние и другие [6].

Для решения задачи коллаборативной фильтрации используются три основных подхода: основанный на соседстве (memory based), основанный на модели (model based) и гибридный подход (hybrid). Первый подход исторически появился первым и характеризуется как достаточно простой в плане реализации, а также эффективный с точки зрения производительности. Но рекомендации, вырабатываемые с помощью данного метода, являются наименее точными. Подход, основанный на моделях при выработке рекомендаций, использует такие методы как метод байесовских сетей, кластеризации, латентной семантической модели, сингулярное разложение, вероятностный латентный семантический анализ, скрытое распределение Дирихле и марковской процесс принятия решений на основе моделей. Данный подход имеет целый ряд преимуществ и характеризуется более высоким качеством рекомендаций по сравнению с первым подходом. Гибридный подход сочетает преимущества подходов, основанных на соседстве и моделях. Данный подход является наиболее эффективным с точки зрения качества предсказаний, но при этом наиболее сложный в реализации и наиболее требовательный к производительности аппаратной платформы.

Основными проблемами, связанными с реализацией и практическим использованием алгоритмов коллаборативной фильтрации, являются разреженность данных, проблема холодного старта и масштабируемость. Разреженность данных изначально присуща исходным данным, которые используются для построения тематических профилей пользователей (покупатели просматривают и(или) оценивают только ограниченное число товаров и (или) услуг). Тем самым качество рекомендаций может быть очень низким, особенно на начальных этапах эксплуатации рекомендательной системы (когда еще не накоплено достаточное количество данных о пользовательской активности). Проблема разреженности данных напрямую связана с проблемой холодного старта, когда рекомендательная система должна вырабатывать рекомендации, имея минимальное количество данных о пользовательских предпочтениях. Проблема масштабируемости становится особенно острой для крупных интернет-магазинов, продающих тысячи товаров миллионам покупателей. При таком количестве товаров и покупателей сложность алгоритма резко возрастает, что усугубляется тем фактом, что рекомендательная система должна давать результат в считанные секунды. Дополнительно к перечисленному выше можно добавить проблему ограничения разнообразия предложений. Рекомендательные системы, использующие коллаборативную фильтрацию, склонны предлагать товары, уже пользующиеся популярностью, что препятствует продвижению новых товаров и услуг [7].

3 Кластеризация пользовательских профилей

Одним из способов решения задачи коллаборативной фильтрации, успешно используемых при реализации рекомендательных систем в современных системах электронной коммерции, является кластеризация. В настоящее время кластеризация – объединение в группы схожих объектов – является одной из фундаментальных задач в области анализа данных и Data Mining [8]. Существует большое количество методов кластеризации, которые условно можно разбить на

следующие основные группы: использующие вероятностный подход (K-means, EM-алгоритм), использующие методы искусственного интеллекта (нейронные сети, генетические алгоритмы), использующие теоретико-графовые модели, иерархические алгоритмы. В контексте решения задачи коллаборативной фильтрации наибольшее распространение получили алгоритмы, использующие вероятностный подход и иерархические алгоритмы.

В рамках проводимых исследований по повышению пертинентности информации в рекомендательных системах поддержки жизнеобеспечения авторы работы оценили возможность применения всех приведённых выше методов кластеризации, а также их комбинации. В результате проведенного имитационного моделирования было установлено, что наиболее подходящим с учётом существующих для работы ограничений по скорости и объемам обработки данных для выявления групп пользователей со схожими предпочтениями (User-User CF) был признан метод кластеризации K-средних (K-means). В целом данный статистический метод прост в реализации и является хорошо масштабируемым [9]. Вычислительная сложность алгоритма O(nkl), где n – число объектов, k - число кластеров, l - число итераций. Одним из недостатков данного метода является необходимость заранее задавать число кластеров для разбиения. Для решения этой задачи проводится предварительный анализ исходных данных, позволяющий посредством минимизации суммы внутрикластерных расстояний выявить оптимальное число кластеров для разбиения.

Следующим шагом стала проверка метода кластеризации на тестовом массиве данных, предоставленных интернет-магазином Thaisoap. Магазин ориентирован на продажу натуральной тайской косметики и кокосового масла. Каталог товаров магазина содержит более
1 500 наименований товаров, которые разбиты на 180 классов (44 корневых классов, 136 подклассов). Ежедневно магазин посещают в среднем около
1 500 посетителей и проводят на нем (в среднем) порядка 11 минут каждый (на каждого посетителя приходится в среднем 28 переходов по ссылкам).

Таким образом исходными данными для построения матрицы активности пользователей стали log-файлы (размер файла с записями за неделю превышает 100 МБ). При этом из всех предоставленных неявных данных для публикации было решено выбрать вычисления по наиболее универсальному агрегированному показателю – число обращений к конкретной категории товара. Кластеризация по данному показателю позволяет сформировать рекомендацию постоянному пользователю сразу в двух случаях: и когда все товары в категории, к которой тяготеет пользователь, просмотрены (User-User переход между категориями), и когда нет (внутрикатегориальная персональная фильтрация).

В результате обработки поступивших данных для каждого пользователя в матрице активности было определено число обращений к конкретной категории товаров. Таким образом, каждая строка матрицы активности пользователей представляет собой вектор оценок, соответствующих различным категориям товаров (тематический профиль пользователя). Профиль пользователя характеризует степень его интереса к каждой группе товаров.

На рисунке 1 представлен фрагмент матрицы активности пользователей, построенный в результате обработки log-файла интернет-магазина Thaisoap за недельный период времени. Число строк таблицы превысило пять тысяч (количество уникальных посетителей за данный период времени), число столбцов 180 (категории товаров).

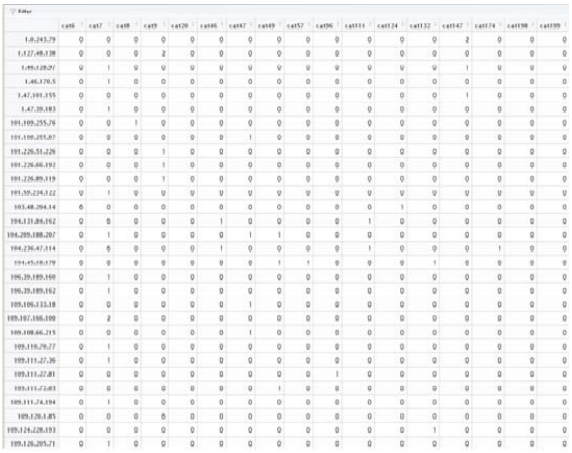

Рисунок 1 Матрица активности пользователей интернет-магазина Thaisoap

Предварительным этапом перед выявлением групп пользователей со схожими характеристиками методами кластеризации является вычисление попарных расстояний между элементами матрицы активности пользователей. В качестве метрики расстояния (функция сходства) было использовано расстояние Евклида (геометрическое расстояние в многомерном пространстве), которое является одной из наиболее простых для реализации и часто используемых на практике метрик на сегодняшний день.

На рисунке 2 представлена гистограмма расстояний, полученная в результате обработки матрицы с попарными расстояниями между объектами. Для построения гистограммы использовался статистический пакет R. Предварительное рассмотрение результатов позволяет сделать вывод, что предпочтения (вектора активности) пользователей интернет-магазина Thaisoap не сильно отличаются (т.е. скорее всего большинство пользователей интересуются схожими категориями товаров). Данный вывод можно сделать исходя из того, что около 90% данных на

гистограмме сосредоточено на отрезке расстояний от 0 до 5.

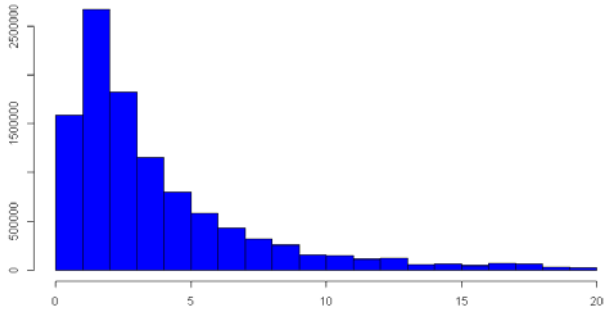

Рисунок 2 Гистограмма расстояний.

Следующим шагом было определение оптимального числа кластеров, требуемого для применения метода кластеризации K-средних. На рисунке 3 приведены результаты расчёта внутрикластерной суммы квадратов расстояний по методу локтя (Elbow method). Данный метод дал число в 30 кластеров.

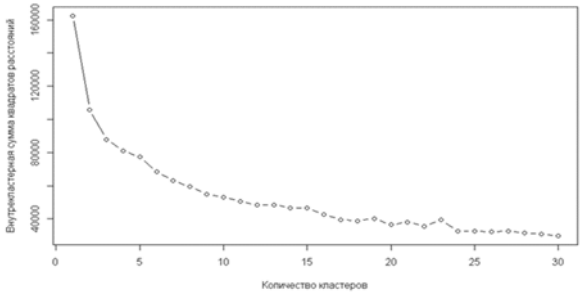

Рисунок 3 Анализ количества кластеров

Таким образом были получены все необходимые параметры и проведена кластеризация данных по матрице активности пользователей. На рисунке 4 в графическом виде представлен результат работы алгоритма кластеризации. Анализ полученных результатов позволяет выделить несколько наиболее крупных кластеров пользователей с номерами 1, 9, 20, 21 и 22. Пользователи из кластера под номером 1 (размер кластера 1 384) демонстрируют слабое предпочтение ко всем категориям товаров. Возможно, что в кластер с номером 1 попали посетители сайта, которые пришли без конкретной цели, например, просто ознакомиться с предлагаемым ассортиментом товаров. Пользователи из кластера номер 9 (размер 180) демонстрируют явное предпочтение к категории cat9 ("Лицо"). Пользователи из кластера номер 20 (размер 1 366) демонстрируют предпочтение к категории cat7 ("Кокосовое Масло"). Сумма квадратов расстояний относительно других кластеров мала, что говорит о небольшом различии объектов внутри кластера. Пользователи из кластера номер 21 (размер 622) демонстрирую также предпочтение к категории cat7 ("Кокосовое Масло"). Пользователи из кластера номер 22 (размер 180) демонстрирую предпочтение к категориям cat47 ("MUST HAVE Зима 2016") и cat49 ("ХИТЫ нашего магазина"). Проведение расчётов на основании данных за другие недели дало схожие результаты.

Таким образом использование метода K-средних на реальных данных с одной стороны подтвердило результаты имитационного моделирования, а с другой стороны выявило недостатки анализа только предпочтений групп пользователей со схожими интересами только по одному агрегированному неявному показателю (User-User CF): для большинства посетителей отсутствует возможность сформировать сколь-либо персональное информационное предложение и требуется как расширенный анализ оставшихся неявных данных, так и иных методов, например, коллаборативной фильтрации посредством анализа взаимосвязей между объектами (Item-Item CF).

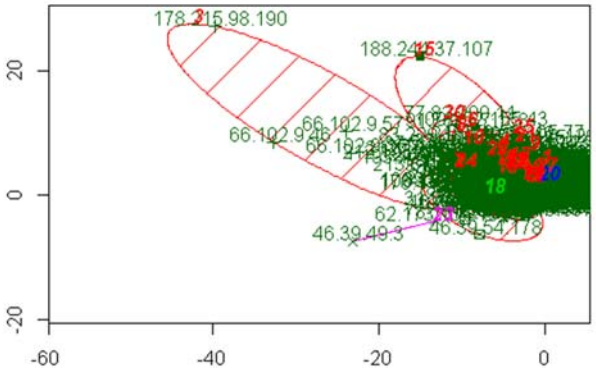

Рисунок 4 Результат работы алгоритма кластеризации

Заключение

В заключение необходимо отметить, что использование различных алгоритмов кластеризации для решения задачи коллаборативной фильтрации в настоящее время является одним из перспективных направлений. Большие объемы данных о пользовательской активности и высокие требования к быстродействию рекомендательных систем накладывают определенные ограничения на используемые алгоритмы. Поэтому наибольшее применение в этой области получили алгоритмы, основанные на оптимизации некоторой целевой функции, определяющей оптимальное (в контексте задачи) разбиение множества объектов на кластеры. В частности, большой популярностью пользуются алгоритмы семейства K-средних (K-means, fuzzy C-means, Густафсон-Кесселя), которые в качестве целевой функции используют сумму квадратов взвешенных отклонений координат объектов от центров искомых кластеров.

В статье на примере данных интернет-магазина Thaisoap показаны достоинства и недостатки кластеризации по простому агрегированному неявному показателю – числу обращений к конкретной категории товара с использованием алгоритма кластеризации K-средних (K-means). В том числе была подтверждена слабая применимость

метода в условиях холодного старта (для новых или малоактивных пользователей). В данном случае требуется применение иных неявных показателей или принципиально других методов, например, Item-Item CF, которые соответственно дают больше информации о пользователе за единицу времени или лучше работают в ситуациях, когда данные о пользовательской активности минимальны. При этом по мере накопления данных о предпочтениях пользователей при выработке рекомендаций рассмотренный метод начинает давать всё более уместные предложения и рекомендуется к использованию как основной.

Литература

[1] А.Н.Алфимцев, В.В. Девятков, С.А.Сакулин Персонализация в гипертекстовых сетях на основе распознавания действий пользователей и нечеткого агрегирования // Вестник МГТУ им.Баумана, Сер. «Приборостроение», 2012, №3.

[2] М. Тим Джонс Рекомендательные системы: Часть 1. Введение в подходы и алгоритмы // Статья в сети Интернет, URL: http://www.ibm.com/developerworks/ru/library/os-recommender1/, 2013.

[3] Марманис Х., Бабенко Д. Алгоритмы интеллектуального Интернета // СПб.-М.: Символ, 2011. – 466 с.

[4] С.А.Филиппов, В.Н.Захаров, С.А.Ступников, Д.Ю.Ковалев Организация больших объемов данных в рекомендательных системах поддержки жизнеобеспечения, входящих в состав глобальных платформ электронной коммерции // XVII международная конференция «Аналитика и управление данными в областях с интенсивным использованием данных» DAMDID/RCDL'2015. Обнинск, 2015.

[5] В.А. Лексин Технология персонализации на основе выявления тематических профилей пользователей и ресурсов Интернет // ВКР Магистра, Вычислительный Центр им. А.А. Дородницина РАН, 2007.

[6] Xiaoyuan Su, Taghi M. Khoshgoftaar A survey of collaborative filtering techniques // Advances in Artificial Intelligence, Volume 2009 (2009), Article ID 421425, 19p.

[7] Fleder D., Hosanagar K. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity // Management Science, Vol. 55, No. 5, May 2009, pp. 697-712.

[8] Барсегян и др. Методы и модели анализа данных: OLAP и Data Mining. – СПб., 2004.

[9] Adam Coates and Andrew Y. Ng. Learning Feature Representations with K-means // Stanford University, 2012, Статья в сети Интернет, URL: http://www.cs.stanford.edu/~acoates/papers/coatesng_nntot2012.pdf.

Clustering of user profiles based on real implicit data in e-commerce recommender systems

Stanislav A. Philippov, Victor N. Zakharov, Sergey A. Stupnikov, Dmitriy Yu. Kovalev

This work is devoted to description of key tasks in the context of building the online store information systems. The main objective is to identify user preferences and their formalization through the formation of the users' behavioral profiles followed by the identification of user groups with similar characteristics. The main source of information about user preferences is implicitly an array of data collected about their actions when navigating through the pages of online shopping. The authors present an approach to building a recommendation system based on collaborative filtering problem solving using cluster analysis techniques to identify groups of users with similar preferences. Advantages of the solution are demonstrated on the example of test data set obtained from the current online store Thaisoap. A unique identifier of the project supported by the Ministry of education and science of the RF is RFMEFI60414X0139.