Подходы к повышению пертинентности информационного предложения в медиасервисах на основе обработки больших объемов 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

Сегодня, когда методы релевантного поиска практически достигли своего потолка, всё большее внимание уделяется повышению пертинентности информации. Особенно это справедливо в области розничной торговли товарами и (или) услугами в интернете, где происходит серьезная борьба за интерес покупателя. Все современные крупные Интернет-магазины, социальные сети, новостные и поисковые сервисы в том или ином виде используют рекомендательные системы, которые позволяют индивидуально в том или ином объеме подстроиться к пользовательской активности и предпочтениям. Для обработки таких данных используются специальные алгоритмы и системы управления данными, обеспечивающие высокие показатели масштабируемости и производительности за счет параллельного выполнения большого количества операций. Тем не менее основные успехи пока достигнуты в области торговли цифровыми произведениями: музыкой, видео и текстами. Именно эти сервисы позволяют перешагнуть через сложившиеся классификации и, например, могут помочь найти любителю хип-хопа джазовые композиции, которые ему действительно понравятся. В данной работе рассмотрены основные подходы к выявлению пользовательских предпочтений, а также подробно описаны принципы построения рекомендательной системы известного сервиса просмотра потокового видео Hulu. Работа выполнена при поддержке Министерства образования и науки РФ, уникальный идентификатор проекта RFMEFI60414X0139.

1 Введение

Решение задачи повышения пертинентности предложений товаров сегодня одно из актуальных и перспективных направлений в сфере электронной коммерции, так как позволяет существенно увеличивать размер корзины и среднего чека.

Труды XVII Международной конференции DAMDID/RCDL'2015 «Аналитика и управление данными в областях с интенсивным использованием данных», Обнинск, 13-16 октября 2015

Пертине́нтность (лат. pertineo — касаюсь, отношусь) — соответствие найденных информационно-поисковой системой документов информационным потребностям пользователя, независимо от того, как полно и как точно эта информационная потребность выражена в тексте информационного запроса. Иначе говоря, это соотношение объёма полезной информации к общему объёму полученной информации.

Неотъемлемой частью решения задачи повышения пертинентности является создание рекомендательной системы, обеспечивающей персонифицированное предсказание объектов (например, товаров и услуг), которые в данное время и при данном настроении представляют ценность для пользователя. В своё время рекомендательные системы существенно изменили порядок взаимодействия Интернет сайтов со своими пользователями сделав их более ориентированными на конкретного пользователя. Предсказания уже строятся на выявлении пользовательских предпочтений, которые, в свою очередь, основываются на результатах анализа поведения, как конкретного пользователя, так и целых групп пользователей со схожими предпочтениями при выборе товаров и услуг.

2 Базовые подходы к построению рекомендательных систем

В настоящее время в рекомендательных системах реализуются следующие основные подходы для выявления пользовательских предпочтений [4, 6]:

- не персонализированные рекомендации (Non-Personalized Recommenders);
- контентная фильтрация (Content Filtering);
- коллаборативная фильтрация посредством анализа предпочтений групп пользователей со схожими интересами (User-User Collaborative Filtering);
- использование статистических метрики оценок (Metrics and evaluation);

- коллаборативная фильтрация посредством анализа взаимосвязей между объектами (Item-Item Collaborative Filtering);
- рекомендации на основе алгоритмов факторизации матриц (Matrix factorization recommendation algorithms).

Наибольшей популярностью при реализации рекомендательных систем в сфере Интернет коммерции на сегодняшний день пользуется подход коллаборативная фильтрация в различных его вариациях (User-User CF, Item-Item CF). Соответствие предложений товаров и услуг ожиданиям пользователей (т.е. пертинентность) при использовании данного подхода обеспечивается выработкой рекомендаций на основании анализа поведенческого профиля, как самих пользователей, так и пользователей обладающих похожими вкусами/предпочтениями. В основе данного метода лежат предположения о том, что пользователю будут интересны товары, обладающие сходными характеристиками с теми, что он уже приобрел (или которыми интересовался) и товары, которые пользуются популярностью у пользователей со схожими предпочтениями.

Также при формировании рекомендаций могут учитываться самые разные дополнительные параметры, включая наличие сопутствующих товаров (например, аксессуары для телефонов), общую популярность товаров у потребителей, рекомендации производителей аналогичных товаров и многое другое. Так, например, сервис Яндекс.Музыка (https://music.yandex.ru/) при формировании рекомендаций учитывает позиции музыкальных треков в чартах, новости о релизах исполнителей, которые интересны пользователю, рекомендации друзей пользователя (из социальных сетей) [7]. Таким образом, сервис по истории прослушиваний музыки пользователем устанавливает какие жанры музыки и каких исполнителей он предпочитает. Это базовая информация о пользовательских предпочтениях. Для того чтобы определить сравнительную ценность предпочтений используются дополнительные данные, такие как, например, оценки «Нравится» и «Не нравится», которые можно ставить музыкальным трекам, альбомам, исполнителям и целым музыкальным жанрам. Помимо оценок и прослушиваний, рекомендательная система сервиса учитывает различные действия пользователя при использовании сервиса. Например, пропуски музыкальных треков (в альбоме, подборке или радио) и добавления треков в плейлисты. Все действия пользователей условно разделяются на положительные и отрицательные. Для формирования рекомендаций анализируется профиль пользователя и учитываются взаимосвязи объектов в каталоге «Яндекс.Музыка».

Таким образом, для того чтобы сформировать список рекомендаций, которые действительно могут быть интересны конкретному пользователю рекомендательные системы обрабатывают значительные объемы разнородных данных. При этом новые действия пользователей приводят к переоценке их предпочтений. По мере накопления данных о пользовательской активности повышается пертинентность предложений товаров и услуг, выполненных рекомендательной системой.

В качестве примера можно привести рекомендательный сервис Имхонет (imhonet.ru), программное обеспечение которого позволяет обрабатывать несколько тысяч запросов в секунду без снижения производительности [3].

Тем не менее наиболее посещаемым и больше других медиасервисов продвинувшимся в области повышения пертинентности информации на основе обработки больших объемов данных стоит признать популярный американский сервис Hulu (http://www.hulu.com/) для доступа к просмотру различного видеоконтента (телевизионные шоу, сериалы, фильмы и т.п.), который обрабатывает более 400 миллионов запросов на получение видеоконтента и более 2 миллиардов просмотров рекламных материалов.

3 Обработка больших объемов данных на примере рекомендательной системы сервиса Hulu

Для управления большими массивами данных в Hulu используется система Apache HBase (http://hbase.apache.org/) первоначально разрабатываемая в рамках проекта Hadoop. Структура организации данных в HBase аналогична тому, как это сделано в Google BigTable, позволяя эффективно работать с таблицами, содержащими миллиарды строк и сотни тысяч столбцов (распараллеливая операция доступа к данным). Обращение к строкам происходит по ключу, который представляет собой байтовый массив. В качестве ключа могут использоваться как строки и числа, так сериализованные сложные структуры данных. Данные хранятся в строках, которые объединяются в семейства с одинаковым префиксом (column families). При создании таблицы должны быть определены первичный ключ и семейства столбцов. Столбцы в семействах могут добавляться по мере надобности, что делает структуру БД более гибкой. Кроме сервиса Hulu СУБД HBase используется в таких крупных проектах как Facebook (www.facebook.com), Twitter (twitter.com) и Yahoo (www.yahoo.com).

Для хранения данных о пользовательских предпочтениях используется сервер структур данных Redis (data structure server), обеспечивающий эффективную обработку порядка 7 тысяч запросов в секунду. Redis представляет собой высокопроизводительную реализацию кэша пар ключ значение (key-value cache and store), где в

качестве ключей могут выступать строки, упорядоченные и неупорядоченные списки, битовые карты, хэш теги. Кэш хранится в оперативной памяти (in-memory database), что повышает его быстродействие, но предъявляет повышенные требования к аппаратной платформе. В зависимости от конфигурации, Redis может сохранять весь набор данных (snapshotting) из оперативной памяти на жесткий диск (бинарный файл формата rdb), вести лог изменений (append-only file) или работать вообще без сохранения изменений. Основными функциональными особенностями Redis являются: поддержка транзакций (transactions) без откатов, реализация парадигмы публикация-подписка (Publish/Subscribe messaging paradigm), возможность обработки скриптов на языке Lua (Lua scripting) с использованием встроенного интерпретатора, использование ключей с ограниченным временем жизни (ключи автоматически удаляются из кеша по истечении времени их жизни), реализация различных алгоритмов/политик замещения информации в кэше (LRU eviction of keys), а также автоматическое восстановление функциональности при отказах (Automatic failover). Как уже было сказано ранее, применительно к сервису потокового видео Hulu в кэше Redis хранится около 4-х миллиардов записей и обеспечивается пиковая производительность порядка семи тысяч запросов в секунду.

Качество предсказаний рекомендательной системы Hulu обеспечивается использованием различных механизмов получения прямых (выбор конкретного продукта для просмотра, выставление оценок, комментарии) и косвенных (сведения о предпочтениях пользователей со сходными интересами, время пребывания на конкретных страницах сайта, предпочитаемые жанры, актеры, студии) сведений о пользовательских предпочтениях. Для каждого пользователя формируется поведенческий профиль, определяются группы пользователей со сходными предпочтениями. Поведенческий профиль пользователя динамично меняется в зависимости от его активности на сайте и в социальных сетях.

Рекомендательная система сервиса Hulu обеспечивает решение двух основных задач: помощь пользователям в поиске интересующей информации (тысячи видео фильмов) и помощь владельцам контента в его продвижении. Исходными данными для ее работы являются видео ролики, которые объединяются в программы (show), а также явные и не явные данные о пользовательской активности. Явные данные представляют собой отклики пользователей о просмотренных роликах, а также их рейтинг (формируемый пользователями). Неявные данные (составляют подавляющее большинство собранных данных) представляют собой информацию о пользовательской активности на сайте (посещенные страницы, просмотренные ролики и т.д.). Для реализации рекомендательной системы используется подход коллаборативной фильтрации (авторы назвали его ItemCF, так как в основе лежит предположение о предпочтительности для пользователя продуктов аналогичных тому, что он выбирал ранее). Архитектура системы разделяется на две основных составляющих: On-line Architecture (компоненты обрабатывающие запросы «на лету») и Off-line Architecture (компоненты выполняющие обработку данных в фоновом режиме).

On-line Architecture включает в себя следующие основные компоненты (см. рис. 1):

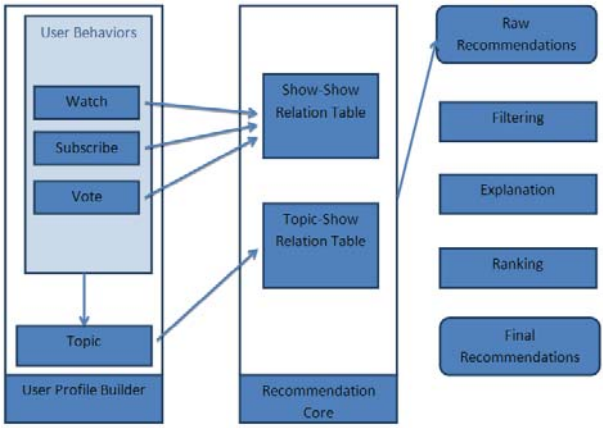

Рис. 1 HULU - On-line Architecture

- User profile builder. Для каждого пользователя в системе формируется индивидуальный профиль, который используется для сохранения сведений о его активности на сайте. Собранные данные используются как для формирования тематических разделов (Topic Model), так и для генерации рекомендаций. Данных накапливается огромное количество, поэтому для их эффективной обработки используется Hadoop кластер.
- Recommendation Core. Анализ данных о пользовательской активности позволяет определить предпочтения каждого пользователя сервиса Hulu. Предпочтения выявляются для тематических разделов (topics) и конкретных программ (show). Предпочтения пользователей учитываются при формировании системой рекомендаций для них.
- Filtering. Фильтрация используется для обработки первичного списка рекомендаций, полученного от Recommendation Core, с целью исключения уже просмотренных конкретным пользователем передач.
- Ranking. Ранжирование используется для лучшего учета предпочтений конкретного пользователя. Список рекомендаций, полученный от Recommendation Core, сначала фильтруется, а затем ранжируется в таком порядке, чтобы пользователь первыми увидел

наиболее интересные ему и еще не просмотренные передачи.

- Explanation. Данный модуль генерирует обоснование (explanation) для каждого из предлагаемых рекомендательной системой пользователю вариантов. Считается, что модуль обоснований является важным элементом рекомендательной системы, так как он позволяет сделать более понятными для пользователя причины появления тех или иных предложений в списке рекомендаций (например, «вам предложен фильм А потому, что Вы посмотрели фильм Б»).

Off-line Architecture включает в себя следующие основные компоненты (см. рис. 2):

- Data Center. Используется для хранения данных о пользовательской активности в системе. Для хранения данных используются реляционные БД, а также Hadoop кластер.
- Related Table Generator. В системе используется два типа таблиц описывающих ресурсы интересные пользователям. Одна таблица содержит результаты коллаборативной фильтрации другая результаты, полученные в результате анализа контента.
- Topic Model. Тематический раздел (Topic) содержит передачи, имеющие похожую направленность. Модель основана на реализации латентного размещения Дирихле (LDA) для формирования разделов содержащих похожий контент.
- Feedback Analyzer. Анализ откликов пользователей позволяет более точно определять рейтинг передач. Так если некоторый ролик рекомендуется большому количеству пользователей, но никто из них этот ролик не стал смотреть, то рейтинг ролика автоматически будет снижен.
- Report Generator. Генератор отчетов позволяет формировать отчеты, используя различные метрики, что необходимо для анализа качества работы рекомендательной системы.

Необходимо отметить, что в основе платформы Hulu лежит использование модели распределенных вычислений Map Reduce, обеспечивающей эффективное решение различных задач в области генерации и анализа данных, извлечения информации, машинного обучения и сортировки [1]. В качестве программной платформы реализующей модель MapReduce используется Apache Hadoop – бесплатная реализация MapReduce с открытым исходным кодом на языке Java.

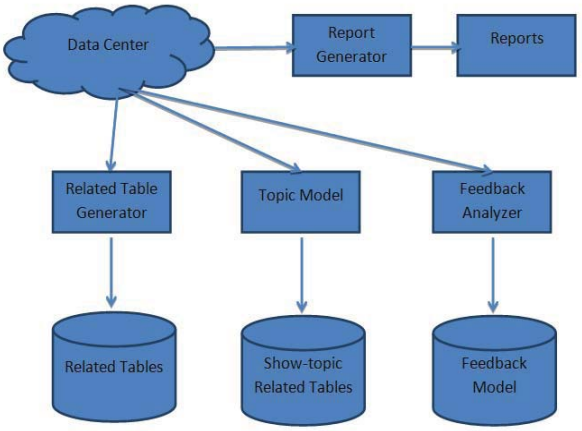

Рис 2. HULU - Off-line Architecture

4 Анализ данных

Пользователи сервиса Hulu генерируют огромное количество данных (как было сказано ранее, более 400 миллионов только просмотров видео в месяц), обработка и анализ которых является критически важной задачей с точки зрения планирования развития бизнеса. Анализ данных позволяет сделать выводы как о выборе потенциального видео контента для инвестирования и размещения его на сайте, так и о комплексе мер, которые необходимо предпринять для повышения привлекательности сервиса Hulu в качестве рекламной площадки для сторонних инвесторов. Таким образом, анализ собранных рекомендательной системой данных является важным инструментом развития бизнеса применительно к сервису потокового видео Hulu.

Анализ данных преследует цели выявления различных зависимостей и трендов (краткосрочных и долгосрочных) в поведении пользователей с тем, чтобы можно было оперативно на них отреагировать. По результатам анализа данных, например, выявляются видео программы, которые пользуются наибольшей популярностью, а также программы, которые не пользуются спросом. На основании этих данных и с учетом поведенческих трендов пользователей вырабатываются рекомендации по инвестированию в те или иные продукты (например, приобретение телевизионных шоу или трансляций спортивных матчей). Также, на основании этих данных могут вырабатываться рекомендации к телевизионным или киностудиям в части целесообразности реализации тех или иных проектов. Еще одной важной задачей анализа собранных системой данных становится выявление пользовательских предпочтений в части используемых аппаратно-программных платформ. Так переход значительного числа пользователей сервиса на использование мобильных устройств требует оптимизации дизайна и компоновки страниц под работу на экранах с небольшой диагональю, ограниченными вычислительными ресурсами и пропускной способностью сети.

Использование пользователями определенных типов браузеров требует соответственно оптимизации сайта для более эффективной работы именно в этих браузерах.

На основании анализа собранных данных могут делаться выводы о эффективности проведения рекламных компаний конкретных товаров в частности и эффективности используемых подходов к рекламированию товаров на сайте вообще. Также может выявляться степень доверия к различным брендам и интереса к деятельности, предлагаемым потребителям продуктам и рекламным акциям конкретных компаний. Не менее интересным для развития сайта является выявление наиболее популярных/эффективных методов рекомендации товаров конечным пользователям. Наиболее распространенными из них являются адресные e-mail рассылки, формирование рейтинга товаров (например, посредством выставления баллов или лайков) и возможность оставлять текстовые комментарии относительно качества предлагаемых товаров.

Заключение

В заключение необходимо отметить, что рекомендательные системы являются важным элементом современных медиасервисов. При этом повышение пертинентности информационного предложения является одной из важнейших задач всех рассмотренных систем: Яндекс.Музыка (аудиоконтент), Имхонет, Hulu (видеоконтент), т.к. напрямую влияет на коммерческую успешность, серьезно увеличивая количество обращений к контенту.

В данной работе представлен краткий обзор основных подходов к построению рекомендательных систем как важного механизма повышения пертинентности (т.е. соответствия ожидания пользователей) предложений товаров и услуг в сфере электронной коммерции. На примере рекомендательной системы сервиса Hulu рассмотрены вопросы хранения и оперативной обработки запросов к большим массивам данных о пользовательских предпочтениях. За рамками статьи остались такие медиасервисы как Spotify, iTunes, Pandora, YouTube и Netflix. Однако это сделано осознанно, т.к. в них применяются те же самые методы Item-based Collaborative Filtering, использующие отклики пользователей для корректировки силы связей между информационными единицами, и модель распределенных вычислений Map Reduce [8-9].

Литература

[1] J. Urbani, S. Kotoulas, E. Oren, F. van Harmelen. Scalable Distributed Reasoning using Map Reduce // Proceedings of the International Semantic Web Conference (2009) Volume: 5823, Publisher: Springer, Pages: 293-309

[2] Konstantin V. Shvachko Apache Hadoop The Scalability Update [Электронный ресурс]. Режим доступа: URL: https://www.usenix.org/ system/files/login/articles/ 105470-Shvachko.pdf, 2011.

[3] Liang Xiang Hulu's Recomendation System [Электронный ресурс]. Режим доступа: URL: http://tech.hulu.com/blog/2011/09/19/recommendation-system/, 2011.

[4] Zan Huang, Wingyan Chung, and Hsinchun Chen. A Graph Model for E-Commerce Recommender Systems. // Journal of the American society for information science and technology, 55(3):259-274, 2004.

[5] Дмитрий Исайкин Хранение и обработка больших массивов данных в рекомендательном движке сайта Имхонет // Профессиональная конференция разработчиков высоконагруженных систем Highload++, 2008.

[6] М. Тим Джонс Рекомендательные системы: Часть 1. Введение в подходы и алгоритмы [Электронный ресурс]. Режим доступа: URL: http://www.ibm.com/developerworks/ru/library/os-recommender1/, 2013.

[7] Панфилов Константин Как работает рекомендательная система «Яндекс.Музыки» [Электронный ресурс]. Режим доступа: URL: http://siliconrus.com/2015/03/yandex-music/, 2015.

[8] Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T. The YouTube video recommendation system // (2010) RecSys'10 - Proceedings of the 4th ACM Conference on Recommender Systems, pp. 293-296.

[9] Benhardson E. Implementing a Scalable Music Recommender System [Электронный ресурс]. Режим доступа: URL: https://www.nada.kth.se/ utbildning/grukth/exjobb/ra pportlistor/2009/ rapporter09/bernhardsson_erik_090 71.pdf – Яз. англ. Дата обращения: 10.05.2015.

Approaches to Improve the Pertinence of Information in the Media Services on the Basis of Big Data Processing

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

Today, when relevant search methods have almost reached its ceiling, increasing attention is paid to improving pertinence information. In this paper, the basic approaches to identifying user personal interests, as well as detailed principles of the recommendation system known as streaming video service Hulu are considered.

This work was supported by the Ministry of Education and Science of the Russian Federation. A unique number of work is RFMEFI60414X0139.