# Организация больших объемов данных в рекомендательных системах поддержки жизнеобеспечения, входящих в состав глобальных платформ электронной коммерции

© 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**

Основная задача рекомендательных систем – это предсказание объектов, которые будут интересны пользователю с учетом его предпочтений. Рекомендательные системы являются удобной альтернативой традиционным поисковым алгоритмам и активно используются в электронной коммерции. Для хранения данных о пользовательской активности, как правило, не используются традиционные реляционные БД. Основной причиной этого являются повышенные требования к быстродействию и значительные объемы обрабатываемых данных. В данной работе рассмотрены основные подходы, используемые при построении рекомендательных систем, кратко охарактеризованы системы управления данными, используемые при работе с большими данными. Также в работе подробно рассмотрены принципы построения рекомендательных систем в сфере электронной коммерции на примере таких крупных Интернет магазинов, как Amazon, eBay и Ozon, которые сегодня начинают играть всё большую роль в поддержке жизнеобеспечения, предоставляя весь спектр товаров для организации жизнедеятельности человека. Работа выполнена при поддержке Министерства образования и науки РФ, уникальный идентификатор проекта RFMEFI60414X0139.

## 1 Задачи, решаемые рекомендательными системами

Основная задача рекомендательных систем – это предсказание списков объектов, которые будут интересны пользователю с учетом его текущих предпочтений. В настоящее время рекомендательные системы активно используются в электронной коммерции для решения задач

увеличения конверсии, т.е. преобразования посетителей в покупателей. Рекомендательные системы позволяют существенно упростить поиск интересных посетителям объектов, а также организовать адресное продвижения товаров и услуг с учетом конкретных пользовательских предпочтений. Научное исследование, проведенное Häubl and Murray в 2003 году, показало [3], что на сайтах, использующих персонализированные товарные рекомендации, потенциальным покупателям на треть проще найти интересующие их продукты, что увеличивает количество повторных визитов и как следствие продаж.

Простейших подходом 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 предпочтения пользователей выявляют посредством формирования поведенческого профиля пользователя, который, в свою очередь, определяется на основании анализа его активности при выборе товаров и услуг. Данные о пользовательской активности обладают следующими основными свойствами: значительный объем и быстрое изменение/обновление данных во времени.

## 2 Алгоритмы, использующиеся в рекомендательных системах

Эффективность работы рекомендательной системы в значительной степени зависит от объема и качества данных о пользовательской активности. Базовыми подходами при построении рекомендательных систем являются

**Proceedings of the XVII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2015), Obninsk, Russia, October 13 - 16, 2015**

коллаборативная фильтрация (collaborative filtering) и контентная фильтрация (content-based filtering) [2, 5]. Коллаборативная фильтрация вырабатывает рекомендации, основанные на модели предшествующего поведения пользователя. Эта модель может быть построена исключительно на основе поведения данного пользователя или — что более эффективно — с учетом поведения других пользователей со сходными характеристиками.

На практике для выявления пользователей со сходными характеристиками (при коллаборативной фильтрации) часто используется алгоритм корреляции Пирсона, позволяющий довольно точно оценить сходство между двумя пользователями (и их атрибутами, такими как статьи, прочитанные в коллекции блогов) [5]. Этот алгоритм измеряет линейную зависимость между двумя переменными (или пользователями) как функцию их атрибутов. Также, широко используются различные алгоритмы кластеризации, позволяющие выявить структуру в рядах на первый взгляд случайных (или немаркированных) данных. В общем случае такой алгоритм базируется на выявлении сходства между элементами (например, между читателями блога) посредством вычисления их расстояния от других элементов в пространстве признаков (feature space) (признаком в пространстве признаков может, например, быть количество прочитанных статей в наборе блогов).

Контентная фильтрация также использует данные о поведении пользователя для формирования рекомендаций, но учитывает его интересы посредством анализа содержимого тех ресурсов, которые привлекли внимание пользователя. Например, этот подход может использовать ретроспективную информацию о просмотрах (какие блоги читает пользователь и характеристики этих блогов). Если пользователь регулярно читает (или комментирует) разделы посвященные, например, психологии, то контентная фильтрация может использовать эту ретроспективную информацию для выявления подобного контента и предложения его пользователю в качестве рекомендованного. Необходимо отметить, что контентная фильтрация является более трудоемким подходом в плане реализации по сравнению с коллаборативной фильтрацией.

Наиболее распространенным на сегодняшний день подходом в сфере электронной коммерции является коллаборативная фильтрация (в различных вариациях). Не вдаваясь в подробности реализации конкретных подходов/алгоритмов необходимо отметить, что, как правило, они работают с огромными массивами данных и требуют выполнения «на лету» значительного количества операций для выявления наиболее потенциально привлекательных для конкретного пользователя объектов.

## 3 Организация данных

Реляционные СУБД малоприменимы для работы с данными о пользовательской активности, которые используются для формирования поведенческого профиля в рекомендательных системах. Основными причинами этого являются большой объем данных и повышенные требования к производительности (рекомендации должны формироваться на лету в процессе просмотра страниц интернет-магазина пользователем). В настоящее время активно развиваются NoSQL (HBase, Cassandra) системы управления данными. Их характерными особенностями являются отказ от транзакций, практически линейная масштабируемость, высокая скорость обработки запросов, отсутствие жесткой схемы данных. Выделяются следующие основные типы NoSQL баз данных: данные хранятся в виде пар ключ/значение (key/value based), данные хранятся в виде столбцов (column based), основным элементом хранения данных является документ (document based), данные организованы в виде графа (graph based). По состоянию на сегодняшний день, в соответствии с данными сайта db-engines.com (http://db-engines.com/en/ranking) в десятке наиболее популярных СУБД находятся семь реляционных баз и три NoSQL СУБД (MongoDB, Cassandra, Redis). Следующим этапом в развитии систем хранения данных становятся NewSQL (SAP Hana, MemSQL) СУБД органично сочетающие преимущества реляционных и NoSQL систем.

Так, например, популярный сервис для доступа к просмотру различного видеоконтента (телевизионные шоу, сериалы, фильмы и т.п.) Hulu (www.hulu.com) использует систему управления данными HBase, а крупнейший Интернет-магазин eBay (ebay.com) систему управления данными Cassandra (Data Stax Enterprise). Системы управления данными HBase и Cassandra поддерживают модель данных BigTable и обладают следующими основными характеристиками [6]: высокая масштабируемость, надежность, высокая пропускная способность (операции чтения и записи), гибкая схема данных, репликации, поддержка SQL подобного языка запросов к данным.

Использование СУБД Cassandra обеспечивает высокий уровень производительности рекомендательной системы Интернет магазина eBay, выполняя в течение суток обработку около 6 миллиардов операций записи и более 5 миллиардов операций чтения данных. Рекомендательная система eBay работает с графом, узлами которого являются пользователи (около 200 миллионов) и товары (около 2-х миллиардов), а ребра (около 40 миллиардов) описывают связи между ними [4]. Похожий подход описан в работе [7], где авторы предлагают использовать модель, основанную на графе, описывающем информацию типа «потребитель-продукт» (user-product information). В

вершинах такого графа содержатся сущности типа «потребитель» и «продукт», а связи между ними описывают транзакции (transactions) и схожести (similarities). Данная модель позволяет реализовать различные подходы к генерации рекомендаций, в том числе direct retrieval (рекомендуются продукты аналогичные выбранному покупателем) и association mining (рекомендация строится на основе анализа поведенческого профиля покупателя).

## 4 Программная платформа интернет магазина eBay

Интернет магазин eBay имеет развитую программную платформу, объединяющую целый ряд программных интерфейсов (API) и сервисов, направленных на более эффективное взаимодействие конечных пользователей с каталогом товаров eBay. Использование их позволяет улучшить качество работы рекомендательной системы, прямо или косвенно влияя на повышение пертинентности предложений товаров и услуг:

- Выполнение поисковых запросов с учетом семантики запроса, что позволяет обеспечивать лучшее соответствие результатов поиска ожиданиям пользователей. Для учета семантики запросов каждый уровень таксономического графа (каталог товаров и услуг) ассоциируется с набором ключевых слов, которые обеспечивают максимально точное отображение возможных терминов в поисковом запросе на соответствующие разделы каталога. Подход, основанный на использовании ключевых слов, является достаточно ограниченным по сравнению с использованием онтологий, но в контексте Интернет магазина выглядит вполне оправданным.
- Обработка откликов покупателей с целью формирования рейтинга продавца (Feedback API). Товары, предлагаемые продавцами с «хорошим» рейтингом, таким образом, будут иметь больший приоритет при формировании предложений для потенциальных покупателей.
- Формирование групп взаимосвязанных товаров, которые могут предлагаться одним «пакетом» (Related Items Management API). Таким образом, продавец может явно указывать какие товары еще могут быть полезны покупателю в контексте его поиска.
- Создание связей, описывающих какие комплектующие или аксессуары с какими продуктами, могут использоваться (Product Services). Так покупатель при поиске переносного компьютера может, например, получить сразу список сумок, подходящих для его модели ноутбука.

## 5 Метрики «эффективности» продвижения товаров в каталоге eBay

Отдельного рассмотрения в контексте задачи повышения пертинентности предложений товаров и услуг заслуживает сервис eBay Listing Analytics, который позволяет выполнить анализ «эффективности» предлагаемых конкретным продавцом товаров и услуг. Для оценки «эффективности» (в данном контексте, успешности рассматриваемой группы товаров или услуг у целевой аудитории) используются следующие ключевые метрики:

- Rank. Данная метрика определяет положение конкретных товаров и услуг в общем рейтинге предпочтений покупателей. Так, например, товар с рейтингом 5 будет предлагаться покупателю раньше, чем товар с рейтингом 15. При работе с сервисом eBay Listing Analytics поиск товаров с целью оценки их рейтинга выполняется по заданным пользователем ключевым словам. Таким образом, конкретное место товара или услуги в списке предложений зависит от ранга и ключевых слов, которые используются при формировании поискового запроса.
- Format. Данная метрика описывает формат, в котором предлагается товар или услуга конечному пользователю (auction-style listing, fixed price listing).
- Impressions. Данная метрика характеризует количество появлений оцениваемой группы товаров и услуг в качестве предложений для потенциальных покупателей в процессе выполнения ими поиска.
- Clicks. Данная метрика оценивает количество обращений покупателей к описаниям конкретных товаров и услуг при получении их в списке предложений после выполнения поискового запроса.
- Click through. Данная метрика это значение метрики Clicks деленное на значение метрики Impressions. Чем больше значение данной метрики, тем лучше, так как это означает, что покупатели чаще выбирают (т.е. заходят на страницу описания) рассматриваемые товары и услуги по сравнению с остальными аналогичными.
- Sold items. Данная метрика представляет собой количество раз, когда покупатели приобретали рассматриваемые товары и услуги.
- Sell through. Данная метрика представляет собой количество проданных товаров или услуг деленное на количество их просмотров при поиске. Чем больше значение метрики, тем лучше, так как это означает, что покупатели чаще после просмотра описаний данных предложений принимают решение о покупке.

- Watchers. Общее число просмотров конкретных товаров и услуг.
- Sales. Количество проданных товаров и услуг в денежном эквиваленте (т.е. общая сумма покупок каждого из интересующего списка товаров и услуг).

Использование аналитических средств, таких как eBay Listing Analytics, позволяет проанализировать успешность различных групп товаров и услуг у покупателей, определить более эффективную стратегию продвижения товаров и скорректировать ассортимент, а в конечно итоге повысить качество обслуживания покупателей за счет предложения им тех товаров и услуг, которые максимально близки их ожиданиям. Стоит напомнить, что указанные метрики как раз обеспечивают срез с тех больших объемов данных, что собирает каталог

## 6 Рекомендательная система интернет магазина Amazon

Другой крупнейший Интернет магазин Amazon (amazon.com) разработал собственное высокопроизводительное хранилище пар «ключ-значение» (Highly Available Key-value Store) Dynamo, которое используется рекомендательной системой Amazon. Dynamo использует синтез хорошо известных техник для достижения масштабируемости и высокой доступности: данные секционируются (partitioning) и реплицируются, используя согласованное хеширование (consistent hashing), а непротиворечивость данных обеспечивается с помощью версий объектов [1]. Для 99% запросов СУБД обеспечивает время отклика на запрос не более 300 мс. Для извлечения информации из хранилища достаточно знать ключевое значение. В период пиковых нагрузок система обеспечивает обработку нескольких миллионов запросов в день.

Рекомендательная система Интернет магазина Amazon реализует различные подходы к формированию рекомендаций, основной целью которых является максимальный учет интересов покупателей посредством вовлечения их в процесс "оценивания" товаров, а также неявного анализа их поведения на сайте:

- Customers who Bought. Данный подход к формированию рекомендаций использует информацию о популярности товаров у покупателей со схожими интересами. Так при выборе конкретной книги, пользователю будет предложен список книг пользующихся популярностью у людей уже купивших данную книгу, а также список авторов книг, чьи работы приобретают покупатели книг, автором которых является автор выбранной книги. Для реализации данного механизма сервис использует специальный алгоритм «Item to Item Correlation», запатентованный компанией Amazon [2]. Основной идеей алгоритма является сопоставление товаров, купленных пользователем товарам с аналогичными по ключевым параметрам характеристиками и формировании рекомендаций с учетом их (товаров) рейтинга.
- Eyes. Данный сервис позволяет пользователям получать оповещения по электронной почте о добавлении новых товаров в каталог Amazon. Пользователи сами могут задавать запросы, на основании которых будет готовиться выборка для оповещения. Запросы можно формировать неявно, посредством использования уже имеющейся выборки в качестве шаблона для поиска.
- Amazon.com Delivers. Данный сервис близок по функциональности к сервису Eyes. Пользователь имеет возможность задать категории каталога (например, книги по кулинарии и ведению домашнего хозяйства) и оформить подписку на получение оповещений о рекомендуемых Amazon товарах из выбранных категорий.
- Book Matcher. Данный сервис дает возможность покупателям оставлять отклики непосредственно о купленных и прочитанных ими книгах. Для прочитанных книг покупатель формирует рейтинг по пяти бальной шкале (from “hated it” to “loved it”). На основе предпочтений пользователя сервис формирует рекомендации для него. При этом рекомендованные книги в свою очередь могут быть оценены (посредством задания рейтинга) покупателем (функция “rate these books”), что позволит в следующий раз более точно учесть его пожелания при формировании рекомендаций.
- Customer Comments. Данный сервис дает возможность получать рекомендации, основанные на мнениях других покупателей. Так, например, для каждой книги в каталоге приводится читательский рейтинг в виде списка из одной или более (до пяти) звездочек, который сопровождается текстовыми комментариями покупателей. Это дает возможность получить более точное представление о книге перед ее покупкой.

Когда посетитель выбирает для покупки какой-либо товар, Amazon на основе этого исходного товара рекомендует посетителю другие товары, приобретенные другими пользователями (с помощью матрицы покупки следующего товара на основе его схожести с предыдущей покупкой).

## 7 Персонализация и товарные рекомендации в Ozon.ru

Ozon.ru находится на рынке уже более 17 лет и также как и зарубежные системы имеет сервисы «Другие пользователи сейчас смотрят», «Бестселлеры», «С этим товаром часто покупают», «Рекомендуем также» и «Аналоги».

Летом 2007 года на OZON.ru был запущен сервис «Персональные рекомендации». Каждый зарегистрированный пользователь интернет магазина получил свою уникальную страницу с подборкой товаров, сделанной на основе автоматического анализа поведения пользователя на сайте. Сервис увеличил количество товаров, которые были заказаны благодаря рекомендациям, на существенные 18 процентов.

Персонализация в Ozon.ru основана на трех ключевых направлениях:

- Поведенческий таргетинг: Размещение маркетинговых/сервисных сообщений для клиентов, чье поведение говорит о интересе к конкретному продукту/услуге;
- Сервис рекомендаций: Размещение товаров, которые могут заинтересовать клиента. За основу взята история посещений сайта;
- Кастомизация контента сайта: Динамически изменяемый контент сайта в зависимости от сегмента клиента и изменения в его поведении.

Персональные рекомендации учитывают:

- отметки «Мне нравится» и «Мне не нравится»;
- добавление товара в корзину;
- покупка товара;
- покупка схожих товаров;
- оценка товара;
- отзыв или мнение об отзыве;
- добавление товаров в список «Хочу в подарок»;
- размещение меток;
- подписка на отзывы или на товар.

Кроме того, данный сервис используется для прицельного рекламирования новинок, т.е. рекомендации отправляются только тем людям, которые потенциально заинтересованы в данных товарах, что позволяет существенно изменить картину старта продаж на более эффективную.

Для сервиса были реализованы решения, позволяющие оставить приемлемым время доступа к сайту с учетом увеличения нагрузки по формированию страниц сайта. К сожалению используемые технологии не раскрываются владельцами бизнеса.

### Заключение

В заключение необходимо отметить, что эффективная обработка больших объемов данных в современных рекомендательных системах достигается за счет использования специальных систем управления данными, называемых NoSQL СУБД. Подобные СУБД имеют высокие показатели отказоустойчивости, масштабируемости и производительности за счет параллельной обработки большого количества запросов. В данной работе представлены различные подходы к организации хранения и доступа к большим массивам данных применительно к рекомендательным системам в сфере электронной коммерции. В качестве примеров эффективной реализации структур хранения больших объемов данных в рекомендательных системах рассмотрены крупнейшие Интернет магазины eBay и Amazon, ежедневно обслуживающие запросы сотен тысяч пользователей по всему миру.

## Литература


[1] Giuseppe De Candia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash, Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall and Werner Vogels Dynamo: Amazon's Highly Available Key-value Store // Статья в сети Интернет, URL: http://db.cs.pitt.edu/courses/cs3551/11-1/handouts/10-1.1.1.115.1568.pdf, 2007.

[2] Greg Linden, Brent Smith and Jeremy York Amazon.com recommendations: Item-to-Item Collaborative Filtering // Industry Report, IEEE INTERNET COMPUTING, 2003.

[3] James Doman-Pipe Personalization Reduces Online Shopping Effort by 32% // Статья в сети Интернет, URL: http://www.smartfocus.com/blog/personalization-reduces-online-shopping-effort-32#sthash.vHdvU3aB.dpuf

[4] Jonathan Gottfried Graph Based Recommendation Systems at eBay // Статья в сети Интернет, URL: http://www.slideshare.net/planetcassandra/e-bay-nyc, 2013.

[5] M. Tim Jones Recommender systems, Part 2: Introducing open source engines // Статья в сети Интернет, URL: http://www.ibm.com/developerworks/library/os-recommender2/index.html/, 2013.

[6] Srinath Perera Consider the Apache Cassandra database // Статья в сети Интернет, URL: http://www.ibm.com/developerworks/opensource/library/os-apache-cassandra/index.html?S_TACT=105AGX99&S_CMP=CP, 2012.

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

[8] 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.

# Organization of Big Data in the Global e-Commerce Platforms

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

This paper discusses the main approaches used in the architecture of recommender systems in e-commerce, i.e. Amazon, eBay and Ozon.ru. This work was supported by the Ministry of Education and Science of the Russian Federation. A unique number of work is RFMEFI60414X0139.