Рекомендательные информационные системы и их приложения

Б. С. Шевченко, Е. А. Черкашин

Иркутский национальный исследовательский униваерситет,  
Институт динамики систем и теории управления им. В.М.Матросова СО РАН,  
г.Иркутск

# Введение

Рекомендательные системы (РС) — Информационные систем поддержки принятия решений, предназначенные для оценки уровня интереса пользователя к определенному продукту или сервиса, на основе имеющейся информации о пользователе и/или продукте и сервисе. Отрасль разработки РС начала развиваться яктивно при появлении онлайн-сервисов, и в настоящее время РС - одно из активных направлений развития систем поддержки принятия решений, ориентированная прежде всео на коммерческое использование, а также на решение задач повышения продуктивности поиска релевантной информации.

В коммерческой области РС позволяют решать задачи выяснения, что именно представляет ценность для потребителя в виде набора конкретных вещей (например, товаров или услуг), сужение вариантов выбора и предоставление схожых вариантов других вещей, тем самым упрощая выбор, выявление новых характеристик у вещей, например, введение вещи в некоторую классификацию и т.п. Отделы снабжения коммерческих фирм получают механизмы обеспечения уникального сервиса каждому потребителю, увеличивая его доверие и лояльность к поставщику, увеличиваются продаж и конверсии, а также получать больше знаний о потребителях [intro].

Рекомендательные системы появились в интернете достаточно давно, около 20 лет назад. Однако настоящий подъем в этой области случился примерно 5-10 лет назад, когда произошло соревнование Netflix Prize. Компания Netflix тогда давала в прокат не цифровые копии, а рассылала VHS-кассеты и DVD. Для них было очень важно повысить качество рекомендаций. Чем лучше Netflix рекомендует своим пользователям фильмы, тем больше фильмов они берут в прокат. Соответственно, растет и прибыль компании. В 2006 году они запустили соревнование Netflix Prize. Они выложили в открытый доступ собранные данные: около 100 миллионов оценок по пятибалльной шкале с указанием ID проставивших их пользователей. Участники соревнования должны были как можно лучше предугадывать, какую оценку поставит определенному фильму тот или иной пользователь. Качество предсказания измерялось при помощи метрики RMSE (средне-квадратичное отклонение). У Netflix уже был алгоритм, который предсказывал оценки пользователей с качеством 0.9514 по метрике RMSE. Задача была улучшить предсказание хотя бы на 10% — до 0.8563. Победителю был обещан приз в $ 1 000 000. Соревнование длилось примерно три года. За первый год качество улучшили на 7%, дальше все немного замедлилось. Но в конце две команды с разницей в 20 минут прислали свои решения, каждое из которых проходило порог в 10%, качество у них было одинаковое с точностью до четвертого знака. В задаче, над которой множество команд билось три года, все решили каких-то двадцать минут. Опоздавшая команда (как и многие другие, участвовавшие в конкурсе) остались ни с чем, однако сам конкурс очень сильно подстегнул развитие в этой области.

В области рекомендательных систем используется специальная терминология. В частности, *профилем* пользователя (продукта или услуги) являются данные, характеризующие *пользователя*, *продукт или услугу*. *Объектом* обозначается то, что система рекомендует пользователям, например, продукты, услуги, товары, новости, книги, DVD и т.п. Именно эти данные используются в процессе оценивания *релевантности* объекта к желаниям пользователя. Этот процесс называется *фильтрацией*. В результате фильтрации объекты ранжируются в соответствии с полученной оценкой, а пользователю предоставляется некоторое конечное подмножество, элементы которого имеют максимальную релевантность, т.е. оцениваются как наиболее интересные пользователю. Далее под *интересом* будем понимать именно интерес пользователя к объекту. Т.к. РС - это прежде всего информационные системы, то все объекты и пользователи описываются при помощи *Атрибутов*. Именно атрибуты - являются входной информацией во все процедуры оценивания интереса.

Рекомендательные системы полезны не только для информационных ресурсов и порталов электронной коммерции, но и могут также открыть новые возможности в области безопасности, автомобильной промышленности [Пятикоп] и др.

Существует ряд подходов к оценке интереса:

1. на основе *фильтрации содержания* (content-based information filtering), при этом в информационной системе создаются профили пользователей и объектов, включающие социальный статус пользователя, возраст, место проживания, род деятельности, а также характеристики, выражающие интерес пользователя к объекту; профили объектов включают позицию в системе классификации, его потребительские характеристики.
2. на основе *коллаборативной фильтрации* (collaborative filtering), где используется информация о поведении пользователей в прошлом, например, перечень покупок или оценок объектов, сделанных на сайте интернет-магазина в прошлом пользователями из той же группы интересов, при этом аналитическим блоком информационной системы автоматически формируются классификации объект, а также *ранжируются атрибуты по степени значимости в оценке интереса*.
3. **интеллектные РС на основе формализованных знаний.** (knowledge-based) ...
4. смешанные **и гибридные** (hybrid prediction) методы, которые базируются на подходах пп. 1 и 2 одновременно.

Например, в Music Genome Project музыкальный аналитик оценивает каждую композицию по сотням различных музыкальных характеристик, при помощи которых выявляются музыкальные предпочтения пользователя. Перечень оценок формирует *профиль музыкального произведения*. Основная проблема первого типа рекомендательных систем — это работоспособность системы на начальном этапе ее эксплуатации, так называемый "*холодный старт*". Например, для новых пользователей в системе нет необходимой информации в профиле для принятия решения о том, какие объекты следует предлагать. В связи с этим в современных рекомендательных системах реализуется механизм сбора и анализа данных о пользователях с применением *явных* и *неявных методов*.

Явные методы сбора данных выполняют следующие действия:

* запрос у пользователя оценки объекта по некоторой шкале;
* запрос у пользователя ранжировки группы объектов от наилучшего к наихудшему;
* предъявление пользователю двух объектов с вопросом о том, какой из них лучше;
* предложение создать список объектов, характеризующих предпочтения пользователя.

Примерами неявного сбора данных выступают:

* наблюдение за тем, что просматривает пользователь в интернет-магазине или базе данных;
* **ведение записей о поведении пользователя онлайн**;
* **отслеживание содержимого компьютера пользователя**.

Рекомендательные системы сравнивают однотипные данные, полученные от разных людей и вычисляют список рекомендаций для конкретного пользователя. Примеры коммерческого и некоммерческого использования РС приведены в статье [???]. Для вычисления рекомендаций используется граф интересов [3]. Рекомендательные системы — **удобная** альтернатива поисковым алгоритмам, так как позволяют обнаружить объекты, которые не имеют непосредственного отношения к поисковому запросу. Любопытно, что рекомендательные системы часто используют поисковые машины для индексации необычных данных.

# Методы фильтрации содержания

В статье [Амелькин2013] рассматриваются контентые рекомендательные системы, решающие задачу top-N. Предлагается математическая модель контентной рекомендательной системы, основанная на нечетких множествах, критерий оценки качества рекомендаций и алгоритм решения задачи.

# Методы коллаборативной фильтрации

Подходы, основанный на коллаборативной фильтрации, в настоящее время более популярны, чем подходы на основе фильтрации содержимого, вероятно из-за того, что представляет собой отражение практического опыта: большинство коммерческих РС вынуждены решать проблему недостатка информации, "холодный старт", а также адаптируемости существующих сообществ пользователей к новым объектам.

...

Математические обозначения элементов модели сравнения состоит из набора пользователей и набора объектов . В виде - множество элементов, оцененных пользователем , - множество пользователей, которые оценили объект , - оценка пользователя для объекта , - вектор всех оценок пользователя , - вектор всех оценок объекта , и - средние значения оценок пользователя и объекта соответственно. Сравнительная оценка обозначается . Для задания этой оценки сначала задается мера близости объекта к объекту . Рассмотрим несколько популярных вариантов оценки близости.

Коэффициент Пирсона [Пятикоп]:

где - множество пользователей, которые оценили объекты и .

Косинус угла между двумя векторами и :

....

Затем множество объектов , наиболее близких к объекту . Например, можно выбрать 30 таких объектов. Вычисление рейтинга объекта делается по формуле:

Популярный подход к формированию множества рекомендаций - это упорядочивавшие всех объектов по критерию схожести и выборке некоторого фиксированного количества объектов с максимальным рейтингом [Нефедова]. В качестве меры similarity двух объектов выступает cos угла между -мерными векторами.

В [Нефедова, Пятикоп] так же представлен обзор способов использования вышеупомянутых методов вычисления оценок, которые разделены на два класса - *анамнестические*, т.е. основывающиеся на одновременной обработке всех имеющихся данных, и *модельные*, где производится предварительная обработка данных, выполняемая, например, раз в сутки. Второй класс позволяет быстрее вычислять оценки интереса, однако не обеспечивает актуальности данных. В классе аналитических способов, как правило используются методы многомерного анализа данных на основе "ближайшего соседства" (Neighbourhood-based), в то время как в модельных методах используется методы анализа скрытых факторов (Latenet Factors). **Кроме того, представителей классов тоже гибридизируют**.

# Гибридные методы

В статье [Дьяконов] рассмотрена задача разработки алгоритмов оценивания лекционного материала. Авторами предложен алгоритм вычисления близости лекций (объектов), где каждая лекция характеризуется подмножество некоторого набора значений (например, подмножеством авторов лекций относительно множества всех авторов). Базовый алгоритм реализует подход фильтрации содержания. Для алгоритма подобранны коэффициенты, при помощи которых можно объединять оценки различных атрибутов в одну общую оценку лекции. Наиболее значимыми атрибутами оказались «категории», «авторы», «языки», «название» и «описание». Цель - синтезировать набор лекций фиксированной длины, рекомендованных для просмотра заданному пользователю, из фиксированного множества "новых", не использованных в построении профилей пользователя и объекта.

**Формулы**

Далее алгоритм дополняется предсказателем последовательностей лекций: заданы примеры последовательностей из трех лекций, требуется для последовательностей из двух предложить третью, четвертую и т.д. Последовательности лекций приобретены системой неявно, т.е. фиксируя просмотренные пользователем лекции. Алгоритм занял первое место в соревновании, причем со значительным отрывом от второго места. Производится внедрение результатов исследований в области анализа сигналов и предсказания последовательностей событий.

**(Прочесть статью, понять, что это первый подход, используется (какой?) явный/неявный метод сбора информации о видеолекциях). Как решена задача? Как внедрено, и было ли? Какие недостатки?**

# Приложения рекомендательных систем в на новостных сайтах

**Алгоритмы MinHash Covisitation**

1. Das A. S., Datar M., Garg A., Rajaram Sh. Google news personalization: Scalable online collaborative filtering // 16th Conference (International) on World Wide Web Proceedings, 2007. P. 271–280.
2. Strauch C. NoSQL databases. http://www.christof-strauch.de/nosqldbs.pdf. (?)

В [Клеменков] создан рекомендательный сервис новостей посетителям сайта, время пересчета рекомендаций в котором на каждую тысячу новых записей в журнале WEB-сервера составляет 1.5–2 с., что авторами заявлено как ресурс, функционирующий в режиме, близком к реальному времени. Для проекта Рамблер-новости подобный результат является удовлетворительным, так как 1000 новых запросов к сайту делается за чуть большее время. В исследовании использован адаптированный алгоритм MinHash для идентификации записей журнала и неточного их сравнения. Целью работы было показать целесообразность применения NoSQL-технологий для создания сервисов указанного качества. Важным свойством приведенной реализации является то, что задачи хранения и анализа данных удалось объединить с задачей предоставления доступа к результатам в единой системе, избежав накладных расходов на перемещение данных из одного источника в другой, что улучшило общую производительность сервиса. Кроме того, предложенный подход упрощает решение повседневных задач сбора статистики о взаимодействии пользователя с веб-приложением путем анализа структурированных логов мощным языком запросов СУБД MongoDB. В результате продемонстрировано, что применение NoSQL к решению подобного класса задач является весьма перспективным.

# Оценка эффективности *работы* рекомендательных систем

В [Амелькин] поднята проблема сравнительной оценки различных подходов к построению РС. Разработчиками рекомендательных систем для измерения эффективности работы рекомендательных систем введен ряд различных показателей (*Табл. 1*). *Аббревиатуры показателей эффективности будут расшифрованы ниже*. Такое разнообразие показателей показывает общее состояние разработок в этой области, когда на основе немногочисленных признанных достижений работают параллельно много независимых групп, создающих свою терминологию, методы верификации и проверки полученных результатов. К автором не сделано окончательного заключения или рекомендации по использованию того или иного критерия в заданном классе РС. **дочитать**.

В настоящее время развиваются методы на основе анализа структуры онтологии предметной области. ….. Подходы на онтологии позволяют решать проблему Холодного старта в той или иной мере. (Черкашин)

# Предметная область

Например в Иркутске можно зяняться следующими задачами, где разработка Рсистемы позволит автоматизировать….

# Литература

1. [intro] Jannach D., Zanker M., Felfernig A., Friedrich G. *Recommender Systems: An Introduction*. Cambridge University Press (2010).
2. Melville P., Mooney R., Nagarajan R. *Content-Boosted Collaborative Filtering for Improved Recommendations (англ.)*// University of Texas, USA : Материалы конф. / AAAI-02, Austin, TX, USA, 2002. — 2002. — P. 187-192.
3. Жернакова О. *Системы рекомендаций и поиска видеоконтента* // Телемультимедиа, 2012.
4. Nadim Hossain. *Why the Interest Graph Is a Marketer’s Best Friend* (англ.). Mashable. Проверено 7 декабря 2013.
5. [Дьяконов] А.Г. Дьяконов. *Алгоритмы для рекомендательной системы: технология Lenkor* // Бизнес-информатика No1(19) – 2012 г. с. 32-39.
6. [Клеменков] П. А. Клеменков. *Построение новостного рекомендательного сервиса реального времени с использованием NoSQL СУБД* // Информатика и еë применения. 2013, Т.7, вып.3, с.14–21.
7. [Нефедова] Ю. С. Нефедова. *Архитектура гибридной рекомендательной системы GEFEST (Generation–Expansion–Filtering–Sorting–Truncation)* // Системы и средства информатики. 2012, Т.22, вып.2, с.176–196.
8. [Амелькин] С. А. Амелькин. Оценка эффективности рекомендательных систем. // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.
9. [Амелькин2013] С. А. Амелькин, Д. М. Понизовкин. Математическая модель задачи top-N для контентных рекомендательных систем. // Известия МГТУ «МАМИ» No 3(17), 2013, т.2. с. 26-31
10. [Пятикоп] Е.Е. Пятикоп. *Исследование метода коллаборативной фильтрации на основе сходства элементов* // Наукові праці ДонНТУ Серія “Інформатика, кібернетика та обчислювальна техніка”, вып.2(18), 2013. с.109-114.
11. wikipedia