*Название рубрики: Вычислительные системы*

УДК 004.584:004.855.5:004.855.6

**Математическое обеспечение рекомендательных информационных систем и его реализации**

**Б.С.Шевченко\*, Е.А.Черкашин\* \*\***

\*Иркутский национальный исследовательский технический университет,   
664074, г. Иркутск, ул. Лермонтова, 83.  
\*\*Институт динамики систем и теории управления СО РАН.

В статье приведен краткий литературный обзор исследований в области рекомендательных информационных систем и примерам областей их применения. В общем виде представлены схемы применения математического обеспечения (методов многомерного анализа данных) на разных этапах решения задач оценивания интереса пользователя к товару или услуге. Выделены основные проблемы, требующие решения, как на этапе разработки рекомендательных систем, так и на этапе их эксплуатации.

Библиогр. 12 назв.

Ключевые слова: *рекомендательные системы; анализ данных; искусственный интеллект; коллаборативная фильтрация; фильтрация содержания; вывод на прецедентах.*

**Борис Сергеевич Шевченко**, магистрант кафедры вычислительной техники,   
тел.: +79247110400, e-mail: servant85@mail.ru

**Евгений Александрович Черкашин**кандидат технических наук, доцент кафедры вычислительной техники  
тел.: +79148706754, e-mail: eugeneai@icc.ru

# Введение

Рекомендательные системы (РС) [1] — Информационные систем поддержки принятия решений, предназначенные для оценки уровня интереса пользователя к определенному продукту или сервису (*объекту*) на основе имеющейся информации о пользователе и/или объекте. Отрасль разработки РС начала активно развиваться при появлении онлайн-сервисов продаж, и в настоящее время РС — одно из активных направлений развития систем поддержки принятия решений, ориентированное прежде всего на коммерческое использование, а также на решение задач повышения продуктивности поиска релевантной информации.

В коммерции РС позволяют решать задачи установления, что именно представляет ценность для потребителя в виде набора конкретных объектов (например, товаров или услуг), сужение вариантов выбора и предоставление схожих вариантов других объектов, тем самым упрощая выбор. РС позволяют также выявлять новые характеристики объектов, например, при помощи ведения классификаций объектов и анализа набора известных признаков. Использование РС позволяет отделам снабжения коммерческих фирм-поставщиков предоставлять уникальный сервис каждому потребителю, увеличивая его доверие и лояльность к поставщику, увеличивая продажи и конверсию, а также получая и накапливая больше знаний о потребителях.

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 NetflixPrize. 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 NetflixPrize. 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 $1000000. 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 [2].

В области рекомендательных систем используется специальная терминология. *Объектом* обозначается то, что система рекомендует пользователям, например, продукты, услуги, товары, новости, книги, DVD и т.п.*Профилем* пользователя или объекта являются данные, характеризующие *пользователя* или объект. 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. *Качество* рекомендации - оценка точности предсказания интереса, сделанного РС, например, в сравнении с имеющимися примерами, т.е. оценками конкретных объектов конкретными пользователями.

Рекомендательные системы полезны не только для информационных ресурсов и порталов электронной коммерции, но и могут также открыть новые возможности в области безопасности, автомобильной промышленности [3], рекламе [4] и др.

Существует ряд подходов к оценке интереса:

1. на основе *фильтрации содержания* (content-basedinformationfiltering), 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* (collaborativefiltering), 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 объектов, производится *ранжирование атрибутов* по степени значимости в оценке интереса.
3. *интеллектные* (knowledge-based), где оценка вычисляется на основе формализованных знаний.
4. *гибридные* (hybridprediction) методы, которые базируются на подходах пп. 1 и 2, включая элементы из 3, что призвано повышать эффективность 1 и/или 2.

Например, в MusicGenomeProject 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 РС (фильтрации содержания) — это работоспособность системы на начальном этапе ее эксплуатации, так называемый "*холодный старт*": 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;
* Сбор информации из социальных сетей, например, как в [5-7].

Второй тип РС, коллаборативная фильтрация, сравнивает однотипные данные, полученные от разных людей и вычисляют список рекомендаций для конкретного пользователя. Для вычисления рекомендаций используется, например, граф интересов.Таким образом, РС представляют собой информационные системы, дополненные алгоритмами, позволяющими обнаружить в хранилище объекты, которые не имеют непосредственного отношения к запросу пользователя. Любопытно, что рекомендательные системы часто используют как поисковые машины для индексации необычных данных.

В данной статье произведен краткий обзор ряда последних публикаций по методам оценки интереса пользователя с целью выбора конкретного математического аппарата для решения задачи построения РС в области продажи недвижимости.

# Рекомендательные системы поиска текстовых документов

В обзоре [8] рассмотрены РС в области предоставления пользователям текстовых документов, в частности, научных статей. Больше половины (55%, 34 из 62) систем построены на основе фильтрации содержания. Алгоритмы коллаборативной фильтрации использованы только в 18% (11 из 62) случаев. Представлены подходы, основывающиеся на стереотипировании и гибридных методах. Авторы исследования пришли к выводу, что в 81% случаев моделирование пользователя на основе автоматического сбора информации не приносит значимых результатов по сравнению с явным указанием набора ключевых слов.

В основе характеристик объектов, научных статей, в исследованных РС используют просто ключевые слова, содержащиеся в документах, реже N-граммы, а также нетекстовые элементы, такие как ссылки на другие статьи и фамилии авторов. Самая популярная модель для хранения представления статей - модель векторного пространства. Моделирование пользователя осуществляется при помощи графов и списков тем, назначенных пользователям в результате машинного обучения. Темы объединяются в иерархические справочники, например, на основе классификаторов АСМ. В рассмотренных подходах тексты извлекаются из заглавий, аннотаций, заголовков, введения, предисловия, предоставленных автором ключевых слов, библиографии, основного текста, социальных тегов и цитирований контекста.

В РС, где применялась коллаборативная фильтрация, и ни в одном из проектов не удалось успешно использовать явные рейтинги: пользователи были слишком ленивы, чтобы самостоятельно задавать рейтинг статьям. Неявные рейтинги получены из данных по количеству страниц, прочитанных пользователем, взаимодействию пользователей с документами (загрузка, редактирование, представление) и цитирования. Главная проблема коллаборативной фильтрации для научных работ - это дефицит информации, например, для РС научных статей Mendeley по сравнению с Netflix (он-лайн фильмы) дефицит составляет три порядка.

Неявные рейтинги объектов получаются из анализа одновременной загрузки статьи (со-загрузка) разными пользователями одной группы, совместного просмотра (со-просмотр), совместное цитирование статьями (со-цитирование) одних и тех же источников. Оказалось, что со-цитирование, будучи эффективным в начале появления статьи на сервисе РС, через два года начинает уступать со-загрузке. Популярным подходом представления результата такого анализа являются графы. Вершины графа - это статьи, представленные наборами атрибутов, а дуги - со-отношения между статьями.

Основными проблемами в области РС являются

* отсутствие общего базиса оценивания качества систем (по предметным областям), включая объективную информацию о реальных оценках реальных пользователей, нестабильность методов оценивания и высокая их зависимость от "шума";
* неиспользованный потенциал научных исследований: новые научные результаты не внедряются в практические приложения (большинство работающих РС базируются на простых методах), данные существующих практических реализаций РС научно не исследуются, нет тесного взаимодействия со снежными областями анализа данных, а так же друг с другом, низкий научный интерес к РС;
* в оценке удовлетворенности не учитывается факторы конфиденциальность, безопасности данных, разнообразие, разметка и презентация информации; в значимом количестве РС моделирование пользователя было крайне примитивно - набор ключевых слов, собственная статья или просто фрагмент текста, представляющих научные интересы пользователя.

Среди открытых проектов выделяются MyMediaLite, LensKit, Mahout, Duine, RecLabCore, easyrec и Recommender.

# Методы фильтрации содержания

В статье [5] решается задача анализа профиля пользователя в социальной сети ВКонтакте для решения проблемы холодного старта в решении задачи рекомендации жанров и произведений музыки и фильмов. Авторами разработана РС «EZSurf» автоматизирующая процесс веб-сёрфинга и фильтрации контента, используя профиль пользователя в социальной сети "ВКонтакте", а также API сервисов last.fm, TheMovieDB для получения сведений о схожих объектах (музыкальных произведений). Такой подход существенно упрощает хранилище данных РС, поскольку не требует создания собственной системы классификаций и базы объектов.

В статье [9] рассматривается задача выделения объектов с наивысшими оценками интереса, задача top-N, при применении фильтрации контента. Предлагается математическая модель контентной рекомендательной системы, основанная на нечетких множествах, критерий оценки качества рекомендаций и алгоритм решения задачи. Математическая модель и алгоритм протестированы на данных сайта last.fm.



# 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 [3]:



где - множество пользователей, которые оценили объекты  и.



Косинус угла между двумя векторами  и:



Затем производится формирование конечного множества объектов наиболее близких к объекту  Вычисление рейтинга объекта делается по формуле:





Популярный подход к формированиюмножества 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 [3,10] 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 (LatenetFactors). Существуют гибридные методы, объединяющие оба предыдущих класса.

# Гибридные методы

В статье [11] рассмотрена задача разработки алгоритмов оценивания лекционного материала. Авторами предложен алгоритм вычисления близости лекций (объектов), где каждая лекция характеризуется подмножеством 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.

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.

# Приложения рекомендательных систем на новостных сайтах

В [12] 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 [13] поднята проблема сравнительной оценки различных подходов к построению РС. Классической оценкой точности предсказания интереса является оценкRMSE, среднеквадратическое отклонение, выполняется при помощи формулы:



где  - множество пар пользователей  и объектов – оценка интереса объекта пользователем , – оценка интереса, сделанная РС.



# Сопутствующие технические задачи

Одной из важных задач, решаемых при разработке РС, является создания пользовательского интерфейса, адекватно отображающего систему критериев, ко которым необходимо производить подбор объектов для пользователя. Например, в статье [14] представлен модуль естественно-языкового интерфейса к базе данных РС, который реализован на основе математических моделей семантических объектов. При помощи модели решаются задачи определения семантики языковой конструкции, заданной пользователем, включая синонимы, классы, отношения и ограничения. В статье приводятся сведения о программной реализации предложенного метода в среде PHP + SQL и результатах тестирования программы на задаче доступа к базе данных РС автомобильного салона.

В [15] решается проблема обеспечения ограничения доступа к личным данным пользователей в контексте построения РС встраивания рекламных сообщений в информационный поток. При этом необходимо контролируемо предоставлять в одностороннем порядке в РС информацию из БД пользователей. Предложено вместо традиционных средств VPN использовать режим функционирования сети с синхронным изменением IP-адреса сервера и переключение клиента на этот адрес.

# Заключение

В статье приведен краткий литературный обзор исследований в области рекомендательных информационных систем (РС) и примерам областей их применения. В общем виде представлены схемы применения математического обеспечения (методов многомерного анализа данных) на разных этапах решения задач РС и оценки их качества. Выделены основные проблемы, требующие решения как на этапе разработки РС, так и на этапе их эксплуатации:

* пользователи неохотно предоставляют информацию о себе и своих потребностях, либо разработчики РС уделяют мало внимания процессу информационного наполнения профиля пользователя;
* в предметных областях, связанных с большой стоимостью объекта или услуги (где принимается серьезные решения по вложения материальных средств), информационные модели объекта и профиля пользователя сложны по своей структуре и связи компонент структуры, что требует явного представления концептуальной модели предметной области во время выполнения РС как своих основных функций, так и функций предсказания значений атрибутов объекта или профиля пользователя на основе прецедентов;
* для предыдущего пункта важным является также разработка пользовательского интерфейса, позволяющего в удобной для пользователя форме и достаточно гибко задавать запросы к РС, а также визуализировать результаты, предлагаемые РС.

Таким образом РС, как системы поддержки принятия решения, являются типичным представителем систем искусственного интеллекта, ориентированными, прежде всего, на обработку неполной и противоречивой информации, а также использующими системы, основанные на формализованных знаниях (knowledge-basedsystems).

# Литература

1. D. Jannach, M. Zanker, A. Felfernig, G. Friedrich. *Recommender Systems: An Introduction*. Cambridge University Press (2010).
2. Как работают рекомендательные системы. Лекция в Яндексе / Блог компании Яндекс / Хабрахабр. [Электронной ресурс] URL:https://habrahabr.ru/  
   company/yandex/blog/241455/ (дата обращения: 12.12.2016).
3. Е.Е. Пятикоп. Исследование метода коллаборативной фильтрации на основе сходства элементов // НауковіпраціДонНТУСерія “Інформатика, кібернетика та обчислювальнатехніка”, вып.2(18), 2013. с.109-114.
4. Е. В.Бритвина. Сегментирование рекомендательной системы с использованием метода организации соединения «клиент - сервер», основанного на программно-конфигурируемых сетях и применении протокола с быстрым перескоком IP-адреса. // Современные проблемы науки и образования. Электронный научный журнал. URL:https://www.science-education.ru/ru/article/view?id=16875 № 6. 2015. (датаобращения: 12.12.2016)
5. Б. Р. Авхадеев, Л. И. Воронова, Е. П. Охапкина. Разработка рекомендательной системы на основе данных из профиля социальной сети «ВКонтакте» // Вестник Нижневартовского государственного университета. Выпуск № 3. 2014.
6. N. Hossain. Why the Interest Graph Is a Marketer’s Best Friend. URL: http://mashable.com/2012/06/19/interest-graph-marketer/  
   #Hr95qUR\_7Eqa (дата обращения: 12.12.2016)
7. О. Жернакова. Системы рекомендаций и поиска видеоконтента // Телемультимедиа. 2012. URL: [http://www.telemultimedia.ru/art.php?  
   id=464](http://www.telemultimedia.ru/art.php?id=464) (дата обращения: 12.12.2016)
8. J. Beel, B. Gripp, S. Langer, C. Breitinger. Research-paper recommender systems: a literature survey // International Journal on Digital Libraries (2016) 17: 305. doi:10.1007/s00799-015-0156-0. (датаобращения: 12.12.2016)
9. 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. 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.
11. А.Г. Дьяконов. Алгоритмы для рекомендательной системы: технология Lenkor // Бизнес-информатика No1(19) – 2012 г. с. 32-39.
12. 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.
13. 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.
14. А. А. Правиков, В. А. Фомичев Разработка рекомендательной системы с естественно-языковым интерфейсом на основе математических моделей семантических объектов. // Бизнес-информатика № 4(14), 2010. с.3-11.
15. Е. В.Бритвина. Сегментирование рекомендательной системы с использованием метода организации соединения «клиент - сервер», основанного на программно-конфигурируемых сетях и применении протокола с быстрым перескоком IP-адреса. // Современные проблемы науки и образования. Электронный научный журнал. URL:https://www.science-education.ru/ru/article/view?id=16875 № 6. 2015.