

# Анализ современных тенденций использования коллаборативной фильтрации в веб-приложениях

**Автор:** Д.В. Михнюк, А.А. Егошина  
Донецкий Национальный Технический Университет  
Кафедра систем искусственного интеллекта

**Источник:** Материалы IV международной научно-технической конференции Донецк, 24 - 25 апреля 2013.- Донецк, ДонНТУ 2013, Том 2, с. 332-336.

## **Аннотация:**

*Михнюк Д.В., Егошина А.А. Анализ современных тенденций использования коллаборативной фильтрации в веб-приложениях. Проведен обзор основных алгоритмов, использующихся в интеллектуальных системах формирования рекомендаций. Детально рассмотрен метод коллаборативной фильтрации и показана целесообразность его использования в интернет-магазинах. Предложено модифицировать классический алгоритм коллаборации путем дополнения его временным фактором и параметрами геолокации.*

## **Постановка задачи**

Современное Интернет-пространство предоставляет пользователю огромное количество разнообразной информации, в которой становится все сложнее ориентироваться, поэтому применение классических средств поиска и систематизации не может полностью удовлетворить потребности пользователя: невозможно просмотреть все материалы, чтобы выбрать для себя подходящие. В связи с этим стало появляться все больше так называемых систем рекомендаций, которые ориентированы на предоставление информации, наиболее полно удовлетворяющую интересы пользователя и наиболее полно отвечающую его запросу.

На сегодняшний день практически вся информация, доступная во всемирной паутине не содержит семантики и поэтому ее поиск, релевантный запросам пользователя, а также интеграция в рамках конкретной предметной области затруднены. Для обеспечения эффективного поиска, веб-приложение должно четко понимать семантику документов, представленных в сети. В связи с этим, можно наблюдать бурный рост и развитие технологий Semantic Web, происходящий в настоящее время. Консорциумом W3C была разработана концепция, которая базируется на активном использовании метаданных, языке разметки XML, языке RDF (Resource Definition Framework – Среда Описания Ресурса) и онтологическом подходе. Все предложенные средства позволяют осуществлять обмен данными и их многократное использование.

Системы рекомендаций - это программы, которые пытаются предсказать, какие объекты (книги, фильмы, музыка, веб-сайты) могут понравиться пользователю, имея определенную информацию о его профиле.

Такие программы используются, как правило, в коммерческих целях (в первую очередь, в Интернет-магазинах, либо на специализированных сайтах «по интересам» с целью предложения товаров). С другой стороны, актуальной задачей является интеллектуализация самого процесса поиска в Интернете. Многие пользователи Интернет объективно полагают, что современные возможности поисковых систем не позволят им найти необходимые документы или данные. Для такого мнения пользователей всемирной сети имеются следующие предпосылки: - взрывной рост объемов доступных обществу данных вообще (увеличение числа книг, фильмов, новостей, рекламных сообщений и пр.); - увеличение объема онлайн-данных; - реальный объем информации, окружающей человека, значительно выше того, что он может реально пропустить через себя, чтобы обнаружить необходимую и достаточную, а также понравившуюся ему.

Системы рекомендаций актуально использовать для интернет-магазинов. Это позволит пользователю рекомендовать популярный, качественный товар который может его заинтересовать, либо при отсутствии какого либо товара на складе посоветовать ему аналог запрашиваемой продукции.

В данной работе будут рассмотрены основные подходы и алгоритмы, используемые в различных системах рекомендаций.

### **Анализ методов расчёта рекомендаций.**

В рассмотренных ниже алгоритмах систем рекомендаций используются следующие определения. Объект - это песня, фильм, товар, пользователь (в случае рекомендации друзей). Т.е. то, что потребляют пользователи системы рекомендаций. Это то, что им нужно рекомендовать. Пользователь - это человек, зарегистрированный в системе, он может покупать, слушать, смотреть, оценивать объекты и пользоваться сервисом рекомендации. Рекомендация - это объект или несколько объектов, которые система рекомендации выдает пользователю. Система рекомендаций позволяет человеку обозначить свои вкусы и возвращает результаты, любопытные для него, базируясь на оценках других пользователей и своих предположениях.

В отличие от поисковых систем, для получения от системы ответа не требуется четкого задания запроса. Вместо этого пользователю предлагается оценить некоторые объекты из коллекции, и на основании этих его оценок и сравнения их с оценками предыдущих пользователей строятся предположения о вкусах пользователя и возвращаются наиболее близкие к ним результаты, формируя для него персонализированную выдачу.

В качестве набора оцениваемых объектов могут, к примеру, выступать: каталог ссылок на веб-сайты, лента новостей, товары в электронном магазине, коллекция книг в библиотеке и т.п. В сферу применения подобных систем входят и ситуации, когда пользователь не ищет информацию по конкретному ключевому слову, а, к примеру, хочет получить список современных статей, похожих по тематике на те, которые он просматривал до этого. В зависимости от того, какие данные используются для расчета рекомендаций, системы делятся на три больших класса:

- Методы коллаборативной фильтрации
- Методы, анализирующие содержимое объектов
- Методы, основанные на знаниях

Методы коллаборативной фильтрации каждого пользователя системы просят высказать свое мнение, выраженное в определенном численном значении на некоторой шкале градации относительно предъявляемого ему ряда объектов. Этими объектами могут быть различные потребительские товары, фотографии, статьи, музыкальные произведения, кинофильмы, телепередачи, компьютерные игры и так далее. По мере того как в базе системы коллаборативной фильтрации набирается все больше и больше собранных оценок, происходят следующие важные вещи:

- система начинает реально понимать, как выглядят собственные предпочтения каждого отдельного пользователя этой системы; - система начинает объединять пользователей в группы по схожести их интересов и делится персональным составом групп с самими пользователями, входящими в эти группы;

- система становится способной дать персональную рекомендацию каждому конкретному пользователю в отношении объектов, с которыми он пока не сталкивался.

Это происходит на основании логики вида Если вы оцениваете это, это и это так-то, так-то и так-то, но не знаете пока чего-то нового, а люди, очень похожие на вас по своим оценкам, оценили это новое вот так, то я предложу вам это новое, потому что я уверен, что вы с высокой степенью вероятности оцените это новое для себя так же, как и те, чьи предыдущие предпочтения совпадают с вашими. что интересы пользователей представлены оценками, которые они дают объектам после просмотра, покупки и т.д. Основная идея данных методов заключается в сравнении между собой интересов различных пользователей или объектов на основе этих оценок. При этом никакой дополнительной информации о самих пользователях и объектах не используется. Методы второго класса, наоборот, используют содержимое объектов для получения рекомендаций. Эти методы работают в тех случаях, когда содержимое объектов представлено в виде текстов. Они хорошо подходят для рекомендации книг. Также их можно использовать для сравнения названий, описаний и другой текстовой информации, доступной у фильмов, песен, товаров и т.д. Методы, основанные на знаниях, требуют от пользователя описать свои требования к нужным ему объектам. А затем ищут с использованием своей базы знаний объекты, удовлетворяющие поставленным требованиям.

### **Коллаборативная фильтрация.**

Для решения задач поставленных в работе предлагается использовать метод коллаборативной фильтрации, поэтому рассмотрим этот метод более детально. Коллаборативная фильтрация (Collaborative filtering) - это метод рекомендации, при котором анализируется только реакция пользователей на объекты. Пользователи оставляют в системе оценки объектов. Причем, оценки могут быть как явные (например, оценка по пятибалльной шкале), так и неявные (например, количество просмотров одного ролика). Конечной целью метода является как можно более точное предсказание оценки, которую поставил бы текущий пользователь системы ранее не оцененным им объектам. Чем больше оценок собирается, тем точнее получаются рекомендации. Получается, пользователи помогают друг другу в фильтрации объектов. Поэтому такой метод называется также совместной фильтрацией. Пусть в системе есть пользователи и объекты. Пусть некоторые пользователи оценили некоторые объекты. И пусть оценка - это натуральное число от 1 до 5. Тогда все оценки можно изобразить в виде матрицы (см. рис. 1). Столбцы в матрице это пользователи, а строки – объекты. Пусть имеется пользователь  $a$ . Наша задача - предсказать, какую оценку поставил бы пользователь  $a$  объекту  $i$ . Будем рассматривать только пользователя  $a$  и тех пользователей, которые оценили объект  $i$ . Алгоритм включает в себя 3 шага: Для каждого пользователя и вычислим, насколько его

интересы совпадают с интересами пользователя  $a$ ; После этого выберем множество пользователей, наиболее близких к  $a$ ; Предскажем оценку на основе оценок объекта  $i$  "соседями" из предыдущего шага. Первый шаг. Каждому пользователю в матрице  $R$  соответствует одна строка. Поэтому будем вычислять близость векторов-строк пользователей. Существует множество способов подсчета близости векторов. Один из самых простых - посчитать косинус между этими векторами:

Здесь  $\text{sim}(u,a)$  - мера близости (похожести) пользователей  $a$  и  $u$ .  $r_{u,i}$  - значение матрицы  $R$ :  $u$  строка,  $i$  столбец.  $\text{sim}(u,a)$  принимает значения из отрезка  $[0,1]$ . Если пользователь не указал оценку для какого-то объекта, соответствующее значение матрицы равно 0. Также часто используется коэффициент корреляции Пирсона, мера близости Дайса, а также другие меры близости. Второй шаг. Теперь нужно выбрать множество  $K$  наиболее похожих на  $a$  пользователей. Есть несколько способов выбора. Чаще всего фиксируется целая константа  $k$ . Затем все пользователи сортируются по убыванию меры близости. И во множество  $K$  входят первые  $k$  пользователей, наиболее близких к  $a$ . Третий шаг. Имея множество  $K$  близких пользователей, нужно вычислить оценку, которую поставил бы пользователь  $a$  объекту  $i$ . Напомним, что рассматриваются только те пользователи, которые оценили объект  $i$ . Нужная оценка вычисляется по формуле:

$r_{a,i}$  - это предсказываемая оценка пользователя  $a$  для объекта  $i$ . Она представляет собой среднее по всем пользователям из множества  $K$ . Используются веса: чем ближе пользователь  $u$  к пользователю  $a$  (согласно мере близости  $\text{sim}(a,u)$ , вычисленной на первом шаге), тем сильнее его вклад в предсказание оценки. Таким образом, описанный алгоритм предсказывает оценки для объектов, которые текущий пользователь еще не оценил. Для того, чтобы сделать рекомендацию для данного пользователя, достаточно предсказать оценки для всех не оцененных объектов и выбрать объекты с наибольшей предсказанной оценкой. Для решения поставленной задачи матрица оценок будет не двумерная а четырехмерная. В ней кроме объектов и пользователей будет ещё 2 параметра: время и геолокация. Время будет показывать какой промежуток времени пользователи тратили на просмотр того или иного объекта, а геолокация будет отвечать за целесообразно заказывать товар (если нет товара на складе в твое городе стоит ли заказывать его с другого конца страны).

## **Выводы.**

В данной работе были кратко изложены основные алгоритмы, используемые в системах рекомендаций. Различные методы используют различные данные о пользователях и об объектах. Каждый подход имеет свои достоинства и недостатки. Например, метод коллаборативной фильтрации рекомендует объекты, не имея никакого представления о том, что они собой представляют. Проблему рекомендации новых объектов решают методы анализа содержимого. Но для их хорошей работы требуются текстовые данные об объектах. Если информации о пользователях, объектах и оценках недостаточно для этих алгоритмов, применяются методы, использующие базы знаний. При этом интерактивно выявляются требования пользователя. Чем больше доступно данных, тем базее точную систему рекомендации можно разработать.

## **СПИСОК ЛИТЕРАТУРЫ:**

1. D. Jannach, M. Zanker, A. Felfernig, G. Friedrich Recommender Systems. An Introduction. New York: Cambridge University Press 32 Avenue of the Americas, 2011. 352 P.;

2. А. Гомзин, А. Коршунов Системы рекомендаций: обзор современных подходов. Пре-принт. Москва: Труды Института системного программирования РАН. 2012. 20 С.;
3. P. Melville, V. Sindhvani Recommender systems. Encyclopedia of Machine Learning. 2010.