УДК 004.89

Д.В. Михнюк, А.А. Егошина
Донецкий национальный технический университет, г. Донецк
кафедра систем искусственного интеллекта

МЕТРИКИ ОЦЕНКИ БЛИЗОСТИ ПОЛЬЗОВАТЕЛЕЙ В КОЛЛАБОРАТИВНЫХ МЕТОДАХ ФОРМИРОВАНИЯ РЕКОМЕНДАЦИЙ

Аннотация

Михнюк Д.В., Егошина А.А. Метрики оценки близости пользователей в коллаборативных методах формирования рекомендаций*. Рассмотрены методы коллаборативной фильтрации, методы, анализирующие содержимое объектов и методы, использующие базы знаний. Проведен сравнительный анализ наиболее популярных метрик оценки похожести пользователей, таких как коэффициент корреляции Пирсона, Евклидово расстояние, Манхэттенское расстояние, косинусная мера и коэффициент Жаккара.*

Ключевые слова: система рекомендаций, коллаборативная фильтрация, мера оценки близости пользователей.

Постановка задачи. В современном мире всеобщей информатизации интернет является огромной всемирной библиотекой, предоставляющей терабайты разнообразной информации, среди которой отыскать то, что необходимо не так то и просто. В этой ситуации использовать стандартные методы предоставления информации пользователю – неэффективно. В связи с этой проблемой стали появляться системы рекомендаций. Они предназначены для предоставления информации, наиболее удовлетворяющей интересам пользователей и наиболее точно соответствующей их запросам.

Системы рекомендаций – это программы, созданные предсказывать, какие объекты (книги, музыка, фильмы, веб-сайты) удовлетворят запросу пользователя, если будет известна информация о его профиле.

С момента появления первых работ по коллаборативной фильтрации в середине 1990-х годов, рекомендательные системы стали объектом пристального научного внимания. В течение последнего десятилетия была проделана большая работа как теоретического, так и прикладного характера, посвященная развитию рекомендательных систем. В настоящее время проблема рекомендательных систем сохраняет к себе большой интерес, так как в этой области остается много задач, решение которых сулит множество возможностей практического применения, что должно помочь пользователям справляться с громадным объемом информации, а также снабдить их инструментами выработки персонифицированных рекомендаций.

Основной этап при построении рекомендательной системы – это оценка близости (похожести) оцениваемых объектов. В настоящее время используют

такие меры как коэффициент корреляции Пирсона, евклидово расстояние, манхэттенское расстояние, коэффициент Жаккара и косинусная мера. Результаты исследований показывают, что выбор метрики оказывает влияние на результаты ранжирования, поэтому целью данной работы является анализ мер близости пользователей, используемых в современных рекомендательных системах.

Обзор методов формирования рекомендаций. 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 предполагают, что 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. 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 (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 пятибальной 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 неоцененным 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.

Множество коллаборативных систем было разработано в бизнесе и в академической науке. Считается, что первой была система Grundy, предложившая использовать стереотипы поведения для построения моделей клиентов, основываясь на небольшом количестве информации о каждом клиенте. Используя стереотипы поведения клиентов, Grundy создавала клиентские профили и использовала их для рекомендации подходящих книг каждому клиенту. Позже система Tapestry использовала индивидуальный анализ для ручного поиска клиентов, обладающих похожими вкусами. GroupLens, Video Recommender, and Ringo были первыми системами, использовавшими алгоритмы коллаборативной фильтрации для автоматического предсказывания. Среди других рекомендательных систем, использующих коллаборативную фильтрацию, можно назвать систему, рекомендующую книги на Amazon.com, систему PHOAKS, помогающую находить нужную информацию на WWW.

Анализ мер близости (похожести) пользователей.

Рассмотрим наиболее популярные меры близости (похожести) пользователей, используемых в современных рекомендательных системах.

Коэффициент корреляции Пирсона. Применяется для исследования взаимосвязи двух переменных, измеренных в метрических шкалах на одной и той же выборке. Он позволяет определить, насколько пропорциональна изменчивость двух переменных. Коэффициент корреляции Пирсона характеризует существование линейной связи между двумя величинами. Если связь криволинейная, то он не будет работать.

У коэффициента корреляции Пирсона есть одно важное свойство, которое можно наблюдать – он корректирует обесценивание оценок. Если один критик склонен выставлять более высокие оценки, чем другой, то идеальная корреляция, все равно, возможна при условии, что разница в оценках постоянна

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

В этой формуле x_i и y_i – оценки критиков, n – количество оценённых объектов, а r – коэффициент подобия.

Эта функция возвращает значение от –1 до 1. Значение 1 означает, что два человека выставили каждому предмету в точности одинаковые оценки.

Главным недостатком этого способа определения сходства является то, что корреляция Пирсона не определена для векторов с постоянными

значениями. Действительно, если мы имеем вектор, знаменатель становится равным нулю. Именно поэтому мы можем терять рекомендации.

Евклидово расстояние. В случае вычисления коэффициента подобия с помощью евклидова расстояния предметы оценивания представляются в виде координатных осей. В системе координат располагаются точки, соответствующие предпочтениям пользователей, на основе которых и определяется коэффициент подобия. Расстояние Евклида рассчитывается по формуле:

$$r = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

В данной формуле x_i и y_i – оценки критиков, n – количество критиков, а r – коэффициент подобия. Расстояние, вычисленное по этой формуле, будет тем меньше, чем больше сходства между людьми. В случае вычисления коэффициента подобия с помощью евклидова расстояния будет давать худший результат, если данные плохо нормализированы.

Манхэттенское расстояние. Это расстояние является средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако для этой меры влияние отдельных больших разностей уменьшается (так как они не возводятся в квадрат). Манхэттенское расстояние вычисляется по формуле:

$$D(x, y) = \sum_i |x_i - y_i|$$

Коэффициент Жаккара. Это способ измерения сходства, который первоначально предложен для оценки подобия популяций и ландшафтов. Опыт показал, что его универсальность выходит далеко за пределы исследований по географии и биологии.

Мера Жаккара основана на оценке соотношения общих признаков двух объектов к их совокупному количеству. Более точно, пусть имеются два вектора одинаковой длины $X = (x_1, ...x_n), Y = (y_1, ...y_n)$. Тогда степень их сходства оценивается по формуле:

$$K(X, Y) = \frac{X \bullet Y}{X^2 + Y^2 - X \bullet Y}.$$

Можно показать, что при небольших значениях (K $\to$0) коэффициент Жаккара ведет себя подобно косинусной мере, при K$\to$1 приобретает свойства, близкие к евклидовому расстоянию.

Косинусная мера. Наибольшее распространение в работах по анализу данных получила косинусная мера, которая вычисляется как косинус угла между соответствующими векторами оценок для товара пользователями:

$$sim(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| * |\vec{b}|}$$

Однако в реальных приложениях начинают возникать сложности, связанные, например, с тем, что разные товары оценивает разное число пользователей, а оценки каждого пользователя, зачастую, сдвинуты в одну из сторон. Для того, чтобы обойти эту проблему, часто применяют модифицированную косинусную метрику:

$$sim(\vec{a},\vec{b}) = \frac{\sum_{u \in U}(r_{u,a} - \bar{r}_u)(r_{u,b} - \bar{r}_u)}{\sqrt{\sum_{u \in U}(r_{u,a} - \bar{r}_u)^2}\sqrt{\sum_{u \in U}(r_{u,b} - \bar{r}_u)^2}}$$

Заключение. В результате проведенного анализа можно сделать вывод, что разные рекомендательные системы могут использовать различные подходы для как можно более эффективного вычисления сходности между пользователями и анализа вынесенных оценок. Общепризнанная концепция заключается в предварительном подсчете соответствий между всеми пользователями системы и лишь периодическом их пересчете (такая необходимость не возникает часто, так как сообщество сходных пользователей (пользователей – соседей) довольно постоянно и за короткое время радикально не меняется). Затем, когда бы пользователь ни обратился за рекомендацией, оценки могут быть с успехом пересчитаны на основании предварительного сходства.

Наиболее популярными метриками являются коэффициент корреляции Пирсона или косинус угла между векторами. Дополнительным плюсом для них является их нормированность, так как значения укладываются в [0, 1]. Для item-based методов также хорошо зарекомендовал себя уточненный косинус угла [5], в котором из рейтингов вычитаются средние значения рейтинга для данного пользователя. Это помогает учесть различные подходы к составлению рейтинга у пользователей, одни из которых могут оперировать лишь высокими оценками, а другие выставлять их лишь немногим предметам.

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.* Препринт. Москва: Труды Института системного программирования РАН. 2012. 20 С.
3. P. Melville, V. Sindhwani *Recommender systems.* Encyclopedia of Machine Learning. 2010.
4. X. Su, T.M. Khoshgoftaar *Survey of Collaborative Filtering Techniques.* Advances in Artificial Intelligence. 2009.
5. A. Ansari, S. Essegaier, and R. Kohli, “Internet Recommendations Systems,” J. Marketing Research, pp. 363-375, Aug. 2000.