УДК 004.8

Порошкина В.В.,
студент магистратуры
2 курс, факультет «Бизнес-информатики
и управления комплексными системами», НИЯУ МИФИ
Россия, г. Москва

МЕРЫ ПОДОБИЯ В РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМАХ

Аннотация: *рекомендательные системы могут помочь пользователю выбрать из большого количества товаров, предлагая интересные ему варианты. Они широко применяются в электронной торговле и при поиске информации для того, чтобы исключить несоответствующую каким-либо критериям информацию. Существуют разные подходы к составлению рекомендаций, но наиболее популярным и широко используемым является метод совместной фильтрации, который использует различные меры подобия для расчета сходства. В этой статье представлен обзор нескольких важных мер подобия.*

Ключевые слова: *рекомендательные системы, совместная фильтрация, меры подобия, Евклидово расстояние, коэффициент корреляции Пирсона, косинусное сходство, коэффициент Жаккара*

Annotation: *recommender systems can help users to choose from a large space of products, offering interesting options. They are widely used in e-commerce and when searching for information in order to exclude information that does not meet any criteria. There are different approaches to formulating recommendations, but the most popular and widely used method is collaborative filtering, which uses various similarity measures to calculate the similarity. This article provides an overview of several important similarity measures.*

Keywords: *recommender systems, collaborative filtering, similarity measures, Euclidean Distance, Pearson Correlation Coefficient, Cosine Similarity, Jaccard Coefficient.*

Poroshkina V.V.,

master

2nd year, Faculty of "Business Informatics and Management of Complex Systems", National Research Nuclear University «MEPhI»

Russia, Moscow

Введение

В интернете можно найти и купить миллионы товаров. Среди такого разнообразия клиенту сложно выбрать подходящие именно ему товары. Рекомендательные системы используются для предоставления качественных рекомендаций, которые помогают клиенту принять решение о покупке товаров. Главная цель рекомендательных систем – предоставить клиенту точные и качественные рекомендации. Почти все рекомендательные системы обычно начинают с поиска группы клиентов, которые приобрели или оценили похожие товары и совпадают с купленными или оцененными товарами текущего пользователя [1]. Существует множество реализаций рекомендательных систем, которые основаны на различных факторах и применяются в разных сферах, таких как гиперпараметрическая оптимизация для рекомендательной системы [2] или рекомендательная система на основе семантического сходства.

Совместная фильтрация

Одна из первых и широко используемых технологий рекомендательных систем – совместная фильтрация. Совместная фильтрация подбирает информацию для пользователя на основе группы других пользователей, имеющих схожие интересы. Рекомендательные системы должны хранить информацию о предпочтениях пользователя – его профиль. Можно попросить пользователей оценить, что они приобрели или чем пользовались. Такой профиль заполняется явно: пользователь сам выставляет оценку. Неявный профиль основан на пассивном наблюдении и содержит исторические данные о

поведении пользователя. Поиск сходства между пользователями является наиболее важной задачей, поскольку точность и качество рекомендаций в основном зависят от них. Сходство рассчитывается с помощью оценок, сделанных другими пользователями.

В данной статье приводится сравнение различных мер подобия, которые можно применить для алгоритмов совместной фильтрации. Выбор идеальной меры подобия очень важен для совместной фильтрации и, следовательно, для рекомендательной системы, потому что разные меры подобия дадут разные результаты в разных контекстах информации [3].

Гибридные рекомендательные системы объединяют два или более рекомендательных метода для повышения производительности и уменьшения количества недостатков. Чаще всего совместная фильтрация сочетается с каким-либо другим методом в попытке избежать проблемы масштабируемости [4]. Один из способов – объединить алгоритмы фильтрации на основе контента и совместную фильтрацию таким образом, чтобы они создавали отдельные ранжированные списки рекомендаций, а затем объединяли их, чтобы составить окончательные рекомендации. Алгоритм фильтрации на основе контента основан на поиске информации, поскольку контент, связанный с предпочтениями пользователя, рассматривается как запрос к системе, а нерейтинговые элементы оцениваются аналогичными элементами [5].

Меры подобия

Наиболее важным шагом в алгоритмах совместной фильтрации является поиск похожих продуктов и пользователей. После поиска похожих пользователей и продуктов легко определить сходство между ними и, наконец, выбрать группу пользователей и продуктов, наиболее похожих на целевого пользователя [6]. Ниже приведены некоторые из популярных мер подобия, используемых в совместной фильтрации.

Основой многих мер подобия является **Евклидово расстояние**. Расстояние между векторами x и y определяется следующим образом:

$$d(x,y) = \sqrt{\sum_{i}^{n}(x_i - y_i)^2}$$

Где x_i и y_i – оценка элемента, заданная двумя разными пользователями для одного и того же элемента, n – количество оцениваемых элементов. Другими словами, Евклидово расстояние – квадратный корень из суммы квадратов разностей соответствующих элементов двух векторов. Стоит учитывать, что для значений x и y корректировка по шкале данных не производится. Евклидово расстояние подходит только для данных, измеряемых в одинаковом масштабе.

В отличие от оценки Евклидового расстояния (которая масштабируется от 0 до 1), **коэффициент корреляции Пирсона** измеряет, как сильно связаны две переменные и измеряется от -1 до +1. Как и модифицированное Евклидово расстояние, коэффициент корреляции Пирсона 1 указывает на то, что объекты данных идеально коррелированы, но в этом случае оценка -1 означает, что объекты данных не коррелированы.

Другими словами, оценка корреляции Пирсона количественно определяет линейную зависимость двух объектов.

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

Косинусное сходство (Коэффициент Отиаи) обычно используется для оценки сходства между двумя экземплярами a и b. Оба объекта приводятся к векторам xa и xb, после чего вычисляется косинусный вектор (векторное пространство) – расстояние между этими векторами, которое указывает на сходство между ними.

$$K = \frac{\sum_{i}^{n} A_i \times B_i}{\sqrt{\sum_{i}^{n}(A_i)^2} \times \sqrt{\sum_{i}^{n}(B_i)^2}} \ ,$$

где A_i – рейтинг пользователя A, а B_i – рейтинг пользователя B для одного и того же элемента, n – количество обычно оцениваемых элементов.

В контексте рекомендации по элементу для вычисления сходства пользователей, пользователь u указывает вектор x_u, где $x_{ui} = r_{ui}$, если пользователь u оценил элемент i, а для нерейтингового элемента – 0. Сходство между u и v двух пользователей будет рассчитываться как:

$$CV(u,v) = \cos(x_a, x_b) = \frac{\sum_i^n r_{ui} r_{vi}}{\sqrt{\sum_i^n (r_{ui})^2}\sqrt{\sum_i^n (r_{vi})^2}}$$

Где r_{uv} – элементы, оцененные пользователями u и v.

Недостаток этой меры в том, что она не учитывает дисперсию оценок, проставленных пользователями u и v.

Коэффициент Жаккара измеряет сходство как пересечение, разделенное объединением объектов. Для текстового документа коэффициент Жаккара сравнивает вес суммы общих терминов с весом суммы терминов, которые присутствуют в любом из двух документов, но не являются общими терминами. Формальное определение:

$$SIM_j(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a} \cdot \overrightarrow{t_b}}{\left|\overrightarrow{t_a}\right|^2 + \left|\overrightarrow{t_b}\right|^2 - \overrightarrow{t_a} \cdot \overrightarrow{t_b}}$$

Где t_a и t_b – рейтинги пользователя A и пользователя B для объединения элементов между пользователями.

Коэффициент Жаккара является мерой сходства и находится в диапазоне от 0 до 1. Коэффициент 1 означает, что два объекта t_a и t_b одинаковы, а 0 означает, что они совершенно разные. Соответствующая мера расстояния $DJ = 1 - SIM_J$.

Исследование состоит из четырех важных мер подобия. Данные для анализа – статистика рейтинга книг, оцененная разными пользователями. Все меры подобия дают одинаковые или разные результаты в зависимости от различных факторов, таких как свойства информации и контекст. На рисунке 1 приведены результаты исследования с помощью всех четырех мер.

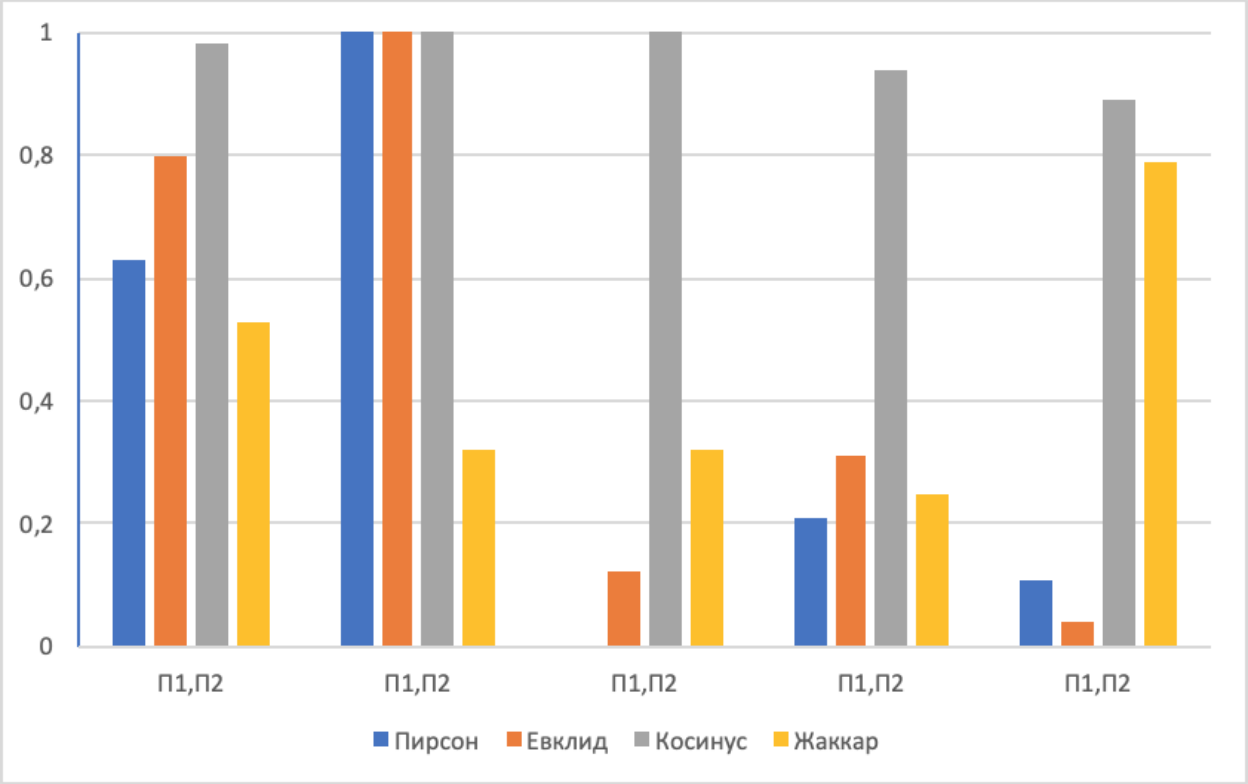

Рисунок 1. Результаты исследования различных мер подобия

Алгоритмы измерения сходства, используемые в исследовании – корреляция Пирсона, Евклидово расстояние, косинусное сходство и коэффициент Жаккара. Алгоритмы, используемые в этой статье, ведут себя по-разному в разных контекстах. Большинство алгоритмов показали одинаковый результат в поиске сходства между пользователями. Полученные значения масштабируются в диапазоне от 0 до 1 для евклидова расстояния, косинусного сходства и коэффициента Жаккара, тогда как значения для корреляции Пирсона находятся в пределах от - 1 до 1. Значение 1 во всех четырех алгоритмах представляет собой полную схожесть предпочтений, а значение 0 показывает, что предпочтения пользователей различны. Значение - 1 в корреляции Пирсона представляет отрицательное сходство между сущностями.

Существует несколько ограничений при выборе наилучшего алгоритма для измерения подобия, например, алгоритм коэффициента Пирсона требует, чтобы минимальное количество объектов (элементов) было больше двух. Корреляция Пирсона, евклидово расстояние и косинусное сходство рассматривают только общие элементы, которые были оценены для измерения

подобия, в то время как коэффициент Жаккара учитывает как общие элементы, так и различные.

В первом эксперименте Пользователь1 оценивал шесть элементов, а Пользователь2 – четыре. Оба пользователя оценили по четыре общих элемента, три из которых они оценили одинаково. Оценка четвертого элемента отличалась в 3 балла. Функция косинусного сходства показала значение 0.982, которое близка к 1, а корреляция Пирсона и Евклидово расстояние показали значения 0.63 и 0.8 соответственно. Корреляция Пирсона и Евклидово расстояние показали приемлемое значение, так как оценки пользователей не были полностью похожи, но почти похожи. Следовательно функция косинусного сходства не дает хороший результат, когда различается только один из оцененных элементов. Её значение стремится к 1.0, но это неверно, т. к. предпочтения пользователей не настолько схожи. Коэффициент Жаккара показал значение 0,53 – частично схожи. Коэффициент Жаккара учитывает как общие оцененные элементы, так и различные. Поскольку Пользователь2 дал оценку только для четырех элементов по сравнению с шестью элементами Пользователя1, коэффициент Жаккара показал меньшее сходство, чем остальные три алгоритма. Поэтому коэффициент Жаккара не подходит, когда мы хотим рассмотреть только общие оценки товара.

Во втором эксперименте Пользователь1 оценивал три элемента, а Пользователь2 – четыре. Пользователи оценили два общих элемента одинаково. Все алгоритмы, кроме коэффициента Жаккара, показали значение 1.0, так как все общие элементы были оценены с одинаковой оценкой. Как упоминалось ранее, коэффициент Жаккара учитывает также не только общие элементы при расчете сходства, поэтому он показал значение, отличное от 1.0.

В третьем эксперименте Пользователь1 оценил только один элемент, а Пользователь2 оценил четыре элемента. Элемент, который оценивал Пользователь1, также оценивал Пользователь2. Разница в оценке составила 8 баллов. Поскольку между пользователями был только один общий элемент, корреляция Пирсона показала значение 0.0, так как этот алгоритм не подходит,

если общих элементов меньше двух. С другой стороны, значение сходства Косинуса приближено к 1.0, потому что только один элемент имел разную оценку. Евклидово расстояние показало приемлемое значение 0.1, поскольку пользователи были не очень похожи. Пользователь2 оценил на три элемента больше, и коэффициент Жаккара показал значение 0.3199. Оценки, сделанные Пользователем2 для оставшихся элементов, были очень похожи на оценку, сделанную Пользователем1. Отсюда и значение, направленное в сторону частичного сходства. Исходя из исследования, алгоритм коэффициента Жаккара лучше всего применим в случаях, когда пользователи оценили одинаковое количество элементов и, возможно, но не обязательно, разные. Коэффициент Жаккара может быть полезен при рекомендации элементов пользователям на основе количества раз, когда они оценивали или покупали элементы. Остальные эксперименты также дали положительные результаты.

Заключение

Рекомендательная система помогает любой организации в развитии бизнеса. В этой статье объясняются методы, используемые для рекомендательных систем. Сравнение различных алгоритмов измерения подобия показывает, что каждый алгоритм работает лучше и дает точные результаты в разных сценариях. Все четыре алгоритма показали положительные результаты в большинстве экспериментов.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1. Грег Линден, Брент Смит и Джереми Йорк, «Совместная фильтрация Amazon от элемента к элементу». [Электронный ресурс]: IEEE Computer Society, февраль 2003 г. URL: https://datajobs.com/data-science-repo/Recommender-Systems-%5BAmazon%5D.pdf (дата обращения 10.12.2018)
2. Саймон Чан, Филип Треливен, Лиция Капра, «Непрерывная оптимизация гиперпараметров для крупномасштабных рекомендательных систем». [Электронный ресурс]: Доклад международной конференции IEEE по большим

данным 2013 г. URL: https://ieeexplore.ieee.org/document/6691595 (дата обращения 10.12.2018)

3. Анна Хуан, «Меры сходства для кластеризации текстовых документов». [Электронный ресурс]: NZCSRSC 2008, апрель 2008. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.332.4480&rep=rep1&type=pdf (дата обращения 17.12.2018)
4. Транос Зува, Сандей О. Оджо, Селеман М. Нгвира и Кенильве Сан, «Обзор методических рекомендаций систем, проблем и метрик оценки». [Электронный ресурс]: Международный журнал новейших и передовых технологий, том 2, выпуск 11, ноябрь 2012 URL: https://www.academia.edu/9470993/A_Survey_of_Recommender_Systems_Techniques_Challenges_and_Evaluation_Metrics (дата обращения 23.01.2019)
5. Ричи Ф., Рокаш Л., Шапира Б., Кантор Пол Б. Руководство по рекомендательным системам – США: изд-во Спрингер Нью-Йорк, 2011 г. – 842 с.
6. Карамолла Багери Фард, Мербахш Нилаши, Мохсен Рахмани, Осман Ибрагим, «Система рекомендаций, основанная на семантическом сходстве». [Электронный ресурс]: Международный журнал электротехники и вычислительной техники том 3, выпуск 6, декабрь 2013. URL: https://www.iaescore.com/journals/index.php/IJECE/article/view/5412/4926 (дата обращения 23.01.2019)