УДК 681.3

О МЕТОДАХ СОЗДАНИЯ РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ

© 2019 Ю. П. Преображенский, В. М. Коновалов

Воронежский институт высоких технологий (Воронеж, Россия)
ЗАО «Лаборатория Касперского»

В статье проводится рассмотрение методов, на базе которых создаются рекомендательные системы. Показаны преимущества и недостатки контент-ориентированных методов.

Ключевые слова: рекомендательная система, метод, контент-ориентированный подход.

Существует два основных подхода к созданию рекомендательных систем - контент-ориентированная и коллаборативная фильтрация. Далее мы рассмотрим каждый из них, и когда они могут применяться.

Суть контент-ориентированного подхода заключается в том, что мы сопоставляем пользователей с тем контентом или товарами, которые им нравились или были ими куплены [1, 2].

Здесь важны атрибуты пользователей и продуктов [3, 4]. Например, для рекомендаций к фильмам мы используем такие признаки, как режиссер, актеры, продолжительность фильма, жанр и т.д., чтобы найти сходство между фильмами.

Кроме того, мы можем извлечь такие характеристики, как оценка настроений и оценки TF-IDF из описаний фильмов и обзоров. (Оценка TF-IDF отражает, насколько важно слово для документа в наборе документов).

Цель контент-ориентированных методов – создать «профиль» для каждого пользователя и каждого предмета.

Рассмотрим пример рекомендации новостных статей пользователям. Допустим, у нас есть 100 статей и словарь размера N. Сначала мы вычисляем оценку TF-IDF для каждого слова в каждой статье. Затем мы строим 2 вектора:

1. **Вектор предмета:** это вектор длины N. Он содержит значения 1 для слов, которые имеют высокую оценку TF-IDF в этой статье, в противном случае значение 0.

2. **Вектор пользователя:** снова вектор размерностью **1xN**. Для каждого слова мы храним вероятность появления слова в статьях, которые употребил пользователь. Обратите внимание, что вектор пользователя основан на атрибутах элемента (в данном случае это оценка слов TF-IDF).

Построив эти **«**профили**»**, мы вычисляем сходства между пользователями и предметами.

Предметы должны быть рекомендованы пользователю, если: 1) они имеют наибольшее сходство с пользователем или 2) имеют большое сходство с другими элементами, прочитанными пользователем. Есть несколько способов сделать это. Давайте посмотрим на 2 распространенных метода:

1. **Косинусное сходство:**

Чтобы рекомендовать предметы, которые наиболее похожи на те, которыми интересовался пользователь, мы вычисляем косинусное сходство между статьями, которые пользователь прочитал, и другими статьями. Наиболее схожие будут рекомендованы.

Косинусное сходство лучше всего подходит, когда ваши признаки высокоразмерны, особенно в области поиска информации и анализа текста.

Чтобы вычислить сходство между пользователем и предметом, мы просто берем косинусное сходство между вектором пользователя и вектором предмета.

2. Сходство Жаккара

Также известное как **пересечение над объединением.**

Сходство Жакара (пересечение над объединением) используется для подобия предмет-предмет. Мы сравниваем векторы элементов друг с другом и возвращаем наиболее похожие предметы.

Сходство Жакара полезно только тогда, когда векторы содержат бинарные значения. Если у них есть ранжирование или рейтинги, которые могут принимать более двух воз-

Преображенский Андрей Петрович – Воронежский институт высоких технологий, к. т. н., профессор, petrovichyur@yandex.ru.
Коновалов Виталий Михайлович – ЗАО «Лаборатория Касперского», специалист,
dtenis0t62vrwet@yandex.ru

можных значений, сходство Жаккара не применимо.

В дополнение к контент-ориентированным методам мы можем рассматривать рекомендацию как простую задачу машинного обучения. Здесь пригодятся обычные алгоритмы машинного обучения, такие как случайный лес, XGBoost и т. д.

Эти методы полезны, когда у нас есть множество «внешних» признаков, таких как погодные условия, рыночные факторы и т. д., которые не являются собственностью пользователя или продукта и могут сильно варьироваться.

Например, цена открытия и закрытия предыдущего дня играет важную роль в определении прибыльности инвестирования в конкретную акцию.

Это относится к классу supervised примеров задач, когда имеются метки, которые могут обозначать, понравился ли пользователю продукт, кликнул ли он на него (0/1), или рейтинг, который пользователь указал этому продукту, или количество единиц, купленных пользователем [5, 6].

Контент-ориентированные рекомендации

Преимущества	Недостатки
Независимость от данных других пользователей	Когда появляется новый пользователь с недостатком данных о его транзакциях, мы не можем качественно делать рекомендации.
Нет проблемы "холодного старта" для новых предметов, т.к. используя признаки предметов, мы можем легко находить похожие предметы	Формирование четких групп похожих продуктов может ограничить рекомендации других продуктов. Мы можем снова и снова рекомендовать лишь малое подмножество из всех продуктов
Результаты рекомендаций интерпретируемы	Если информация о продуктах ограниченна, трудно различать предметы и группировать их. В результате качество рекомендаций будет низким

Рисунок 1 – Преимущества и недостатки контент-ориентированных методов

Перейдём к коллаборативной фильтрации.

Основополагающее предположение подхода коллаборативной фильтрации заключается в том, что если A и B покупают аналогичные продукты, A, скорее всего, купит продукт, который купил B, чем продукт, который купил случайный человек.

В отличие от контентно-ориентированного подхода, здесь нет признаков, соответствующих пользователям или предметам. Все, что у нас есть – это матрица полезности.

Для подхода на основе памяти запоминается матрица полезности, и рекомендации составляются путем запроса данного пользователя к остальной части матрицы полезности.

Кластеризация

Кластеризация обычно используется, когда задача рекомендательной системы становится задачей без учителя.

Если вы только начинаете заниматься бизнесом и у вас очень мало исторических/размеченных данных, вы можете кластеризовать наблюдения на основе набора признаков, а затем назначить рекомендации для кластеров на основе меток, которые имеются у объектов в этом кластере.

Это решение, конечно, не дает лучших результатов сразу, но является хорошей отправной точкой для таких случаев, пока не будет получено достаточно данных.

Кластеризация также может быть использована для создания мета-признаков для объектов. Например, после кластеризации можно назначить значения от 1-k в качестве

нового элемента «кластер» для каждого наблюдения, а затем обучить основную модель всем функциям. Это может быть сделано на уровне пользователя или продукта.

Оценивающие метрики

Основное препятствие при разработке систем рекомендаций – выбор метрик для оптимизации.

Это может быть сложно, потому что во многих случаях цель – НЕ рекомендовать все те же продукты, которые пользователь купил ранее. Так как же узнать, хорошо ли работает ваша модель, предлагая продукты?

Статистические метрики

Они используются для оценки точности метода фильтрации путем сравнения прогнозируемых рейтингов непосредственно с фактическим рейтингом пользователей.

Средняя абсолютная ошибка (MAE), среднеквадратическая ошибка (RMSE) и корреляция обычно используются в качестве статистических метрик.

MAE является наиболее популярным и широко используемым – это мера отклонения рекомендации от фактической стоимости пользователя.

Чем ниже значения MAE и RMSE, тем точнее механизм рекомендаций прогнозирует пользовательские рейтинги. Эти метрики удобны, когда рекомендации основаны на прогнозировании рейтинга или количества транзакций.

Они дают нам представление о том, насколько точны наши прогнозы и, в свою очередь, насколько точны наши рекомендации [7, 8].

Метрики поддержки принятых решений.

Популярными среди них являются **Precision (точность)** и **Recall (полнота)**. Они помогают пользователям выбирать продукты, которые более похожи среди доступного набора продуктов.

Метрики рассматривают процедуру прогнозирования как бинарную операцию, которая отличает хорошие элементы от тех, которые не являются хорошими.

Предложение для исследования заключается в поиске контента с истинной информацией, которая была получена не из исторических данных пользователей. Затем этот контент сравнивается с неисторическим содержанием в рекомендациях с использованием любой из стандартных метрик, описанных выше.

Это дает нам представление о том, насколько хороша наша модель в рекомендации продуктов, которые напрямую не связаны с прошлыми транзакциями.

При разработке системы рекомендаций, особенно для рекомендаций, основанных на содержании, важно помнить, что нужно оптимизировать НЕ только одну метрику.

То есть для рекомендации новостной статьи не рекомендуется отдавать статью с очень высоким значением привлекательности, потому что это означает, что мы рекомендуем пользователям контент, который они, скорее всего, потребляли бы и без нашей рекомендации [9, 10].

Так мы в любом случае не улучшим бизнес.

Мы должны обеспечить достойный precision/recall как показатель того, что наша модель способна изучать предпочтения пользователя, но нет цели пытаться максимально улучшить их. Кроме того, мы не хотим терять вовлечение пользователей в долгосрочной перспективе, рекомендуя одни и те же типы продуктов снова и снова.

ЛИТЕРАТУРА

1. Гуськова, Л. Б. О построении автоматизированного рабочего места менеджера / Л. Б. Гуськова // Успехи современного естествознания. –2012. –№ 6. – С. 106.

2. Завьялов, Д. В. О применении информационных технологий / Д. В. Завьялов // Современные наукоемкие технологии. – 2013. – № 8-1. – С. 71-72.

3. Черников, С. Ю. Использование системного анализа при управлении организациями / С. Ю. Черников, Р. В. Корольков // Моделирование, оптимизация и информационные технологии. –2014. – № 2 (5). – С. 16.

4. Преображенский, Ю. П. О повышении эффективности работы промышленных предприятий // Исследование инновационного потенциала общества и формирование направлений его стратегического развития. Сборник научных статей 8-й Всероссийской научно-практической конференции с международным участием. –2018. – С. 45-48.

5. Будко, Н. А. Применение ИНС в интерфейсах человек-машина / Н. А. Будко, Р. Ю. Будко, А. Ю. Будко // Моделирование, оптимизация и информационные технологии. –2019. –Т. 7. – № 1 (24). – С. 328-340.

6. Преображенский, А. П. Возможности обеспечения развития предприятий / А. П. Преображенский // В мире научных открытий. –2015. – № 10 (70). – С. 196-201.

7. Преображенский, Ю. П. Проблемы управления в производственных организа-

циях / Ю. П. Преображенский // Актуальные проблемы развития хозяйствующих субъектов, территорий и систем регионального и муниципального управления Материалы XIII международной научно-практической конференции. Под редакцией Ю. В. Вертаковой. – 2018. – С. 208-211.

8. Петращук, Г. И. Маркетинг в прикладном менеджменте / Г. И. Петращук // В мире научных открытий. – 2010. – № 4-7 (10). – С. 35-36.

9. Свиридов, В. И. Лингвистическое обеспечение автоматизированных систем управления и взаимодействие пользователя с компьютером / В. И. Свиридов, Е. И. Чопорова, Е. В. Свиридова // Моделирование, оптимизация и информационные технологии. – 2019. – Т. 7. – № 1 (24). – С. 430-438.

10. Кизим, А. В. Программный комплекс поддержки модернизации технических систем / А. В. Кизим, А. В. Матохина, А. Г. Кравец, И. П. Мединцева // Моделирование, оптимизация и информационные технологии. – 2019. – Т. 7. – № 2 (25). – С. 311-324.

ABOUT METHODS FOR CREATING RECOMMENDATION SYSTEMS

© 2019 Yu. P. Preobrazhenskiy, V. M. Konovalov

Voronezh institute of high technologies (Voronezh, Russia)
Kaspersky Lab CJSC

The paper discusses the methods on the basis of which recommendation systems are created. The advantages and disadvantages of content-oriented methods are shown.

Key words: recommendation system, method, content-oriented approach.