И.А. Олянич

ОБЗОР АЛГОРИТМОВ ПОСТРОЕНИЯ АССОЦИАТИВНЫХ ПРАВИЛ ПО КРУПНОФОРМАТНЫМ ДАННЫМ

(Самарский университет)

В наше время идет значительный рост популярности к анализу данных, так как технологии не стоят на месте и развиваются под нужды людей. Особой популярностью пользуются рекомендательные системы. Они помогают вести диалог с каждым пользователем и предлагать товары, фильмы, музыку и многое другое, что сможет заинтересовать конкретного человека. Если же у клиента поменялись предпочтения, то рекомендательные системы будут изменены в сторону новых пожеланий. Это очень мощный инструмент для крупных компаний, которые делают все возможное для улучшения бизнес-процессов и повышения конверсии.

Для построения рекомендательных систем необходимо собрать данные, чтобы иметь представление о человеке. Это возможно сделать двумя путями.

При явном сборе данных клиент сам передает интересующую информацию для дальнейшей обработки путем заполнения анкет, которые могут включать такие поля как, - имя, пол, возраст, место проживания, интересы и многое другое.

Неявный сбор данных применяется в тех случаях, когда клиент отказывается предоставить необходимую о себе информацию. В этом случае осуществляется слежка за человеком, в ходе которой все его действия запоминаются для их дальнейшей обработки. Сюда можно отнести данные о совершенных покупках, посещении страниц, оставлении комментариев и многое другое.

После того, как данные известны, можно приступать к построению рекомендательных систем. Существует два основных типа рекомендательных систем – это контентная фильтрация (англ. content-based filtering) и коллаборативная фильтрация (англ. collaborative filtering) [1]. Можно также выделить гибридные подходы, которые сочетают в себе и то, и другое, что является большим преимуществом при работе с новыми пользователями и товарами, услугами. Однако сложность повышается в разы.

В случае с коллаборативной фильтрацией можно выделить два базовых метода – это рекомендации, основанные на пользователях (англ. user-based collaborative filtering) и рекомендации, основанные на продуктах (англ. item-based collaborative filtering) [2].

User-based предполагает предложение товаров, которые покупали похожие пользователи. Производится усреднение рейтинга товара, проставленный другими пользователями, с учетом степени похожести пользователей [3]. Пример данного метода представлен на рисунке 1.

Item-based предполагает предложение товаров, которые схожи с теми, что были ранее приобретены пользователем. Производится усреднение рейтинга

уже оцененных товаров с учетом степени похожести на неоцененный товар. Пример данного метода представлен на рисунке 2.

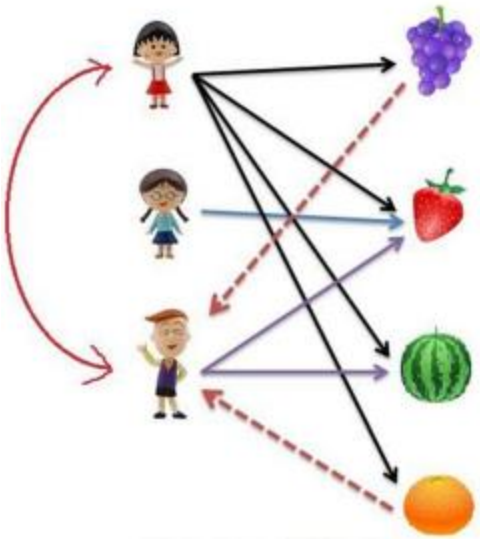

Рис. 1. User-based подход

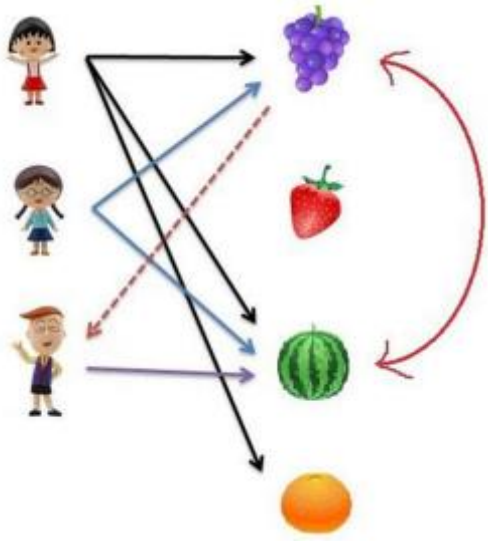

Рис. 2. Item-based подход

Стоит отметить, что подготовка данных для их последующего анализа тоже является одной из ключевых задач и порой занимает достаточно много времени.

В данной работе для построения рекомендательных систем, используется данные о транзакциях в различных магазинах по всему миру. Их ценность в том, что это покупки реальных людей. Тем самым, построение рекомендательных систем уже будет показывать реальную картину в каждой из стран. Исходные данные включают в себя следующую информацию:

1. InvoiceNo – номер чека, с его помощью можно проследить историю покупок конкретного человека в определенном магазине;
2. StockCode – уникальный код магазина;
3. Description – название товара;
4. Quantity – количество товара;
5. InvoiceDate – дата покупки;
6. UnitPrice – стоимость;
7. CustomerID – уникальный идентификатор клиента, к которому привязываются чеки. Стоит отметить, что если клиент не использует банковскую карту, либо карту лояльности, то значение является пустым, так как не предоставляется возможным идентификация такого клиента;
8. Country – страна.

Для обработки большего количества данных был выбран Spark совместно с R. Были реализована алгоритмы ALS и SVD и применены попытки их оптимизации, о которых кратко будет сказано в заключении. Если говорить об анализе имеющийся информации, то удалось выявить, что страны, находящиеся рядом друг с другом не всегда схожи по потребительской корзине. В частности,

будет ошибочно предполагать, что рекомендации для Франции будут такими же, как для Германии.

На рисунке 3 представлен пример топ 5 рекомендация для жителей Франции на основе всех магазинов.

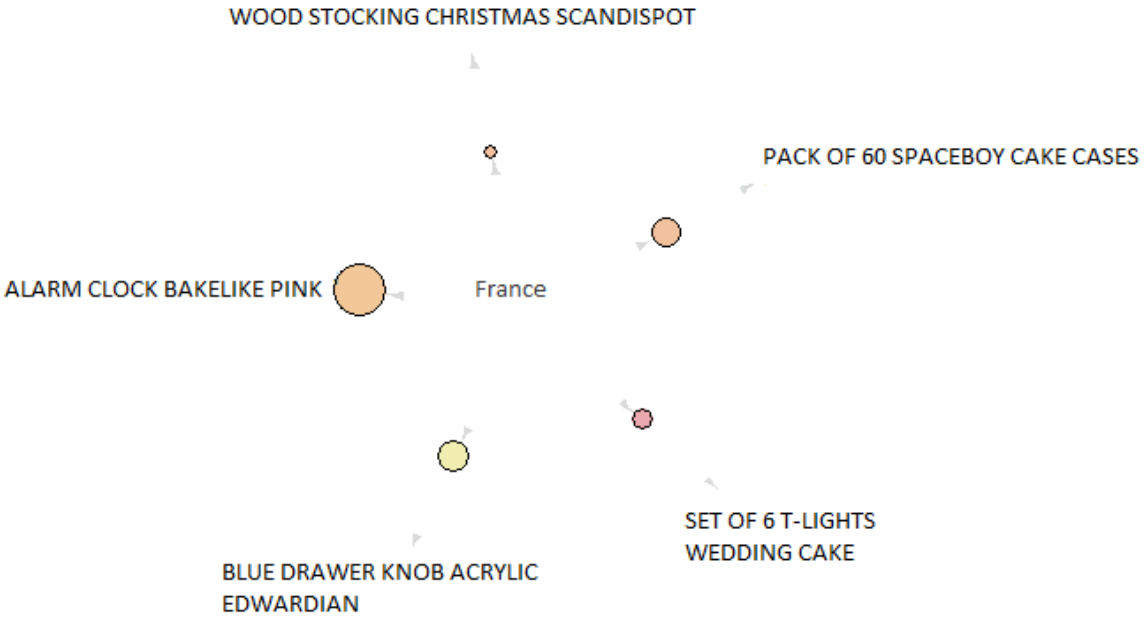

Рис. 3. Наиболее популярные рекомендации во Франции

Говоря о самих алгоритмах, ALS представляет собой метод градиентного спуска. В нем необходимо попеременно находить минимумы по разным координатам. Достаточно важным моментом является, что каждый шаг может был распараллелен, что в нашем случае является важным аспектом, говоря про оптимизацию.

SVD представляет собой сингулярное разложение матриц и достаточно плох в распараллеливании.

Проведя эксперимент на выше указанных данных, удалось получить следующие результаты, представленные на рисунке 4.

Как видно из диаграммы, отличие не критичное, но все же есть. Алгоритм SVD позволил достичь наилучших показателей, а вот алгоритм ALS по своему принципу работы не может показать такой же результат при любых своих модификациях.

В дальнейшем при помощи оптимизированного алгоритма будут построены наиболее точные рекомендации для каждой из стран. Планируется на основе полученных рекомендация выявить наиболее востребованные товары в магазине по его уникальному идентификатору и стране. Эта информация может помочь магазинам закупать ту часть продуктов, которая пользуется наибольшим спросом, минимизируя издержки. Стоит отметить, что в Китае начинают открываться мини-маркеты без продавцов, которые содержат в себе наиболее востребованные товары в конкретной местности.

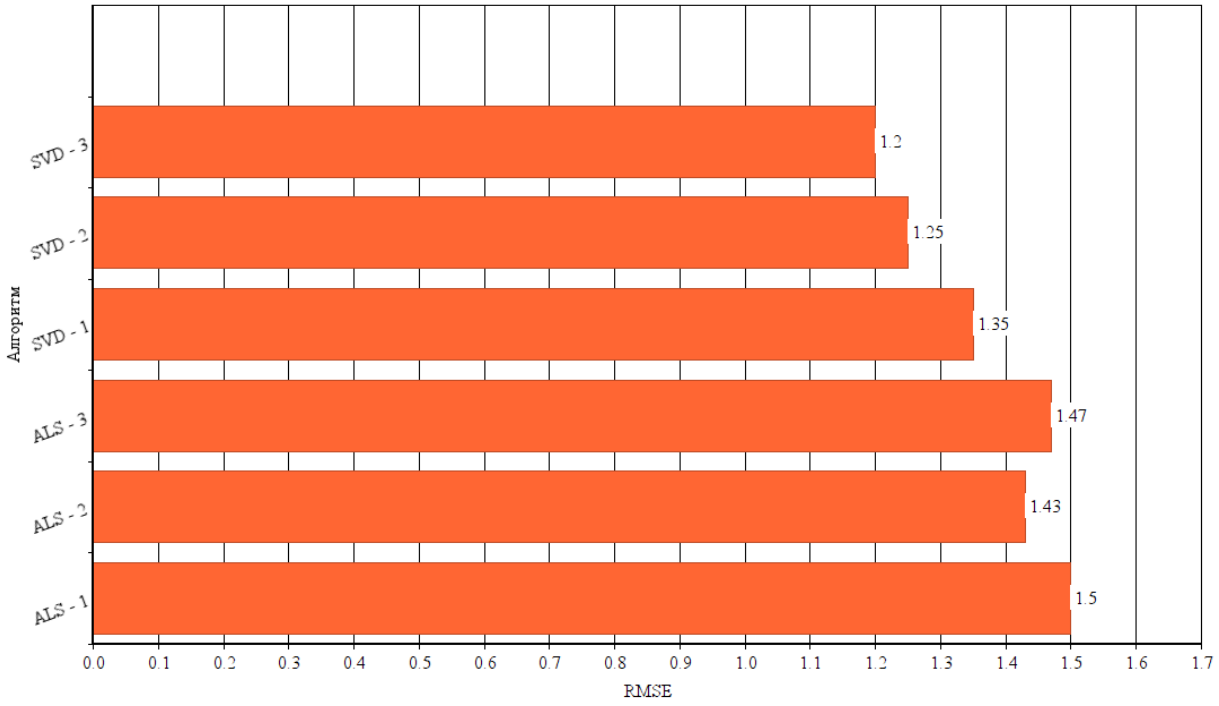

Рис. 4. Результат вычислительного эксперимента

Литература

1 Daniel Nee. Practical Data Science [Electronic resource]. - Access mode: http://danielnee.com/ (15.10.2017)

2 Шалаев С. Эволюция рекомендательных сервисов [Электронный ресурс]. - Режим доступа: http://firrma.ru/data/articles/5006/ (20.10.2017)

3 Recommendation Engine Introduction [Electronic resource]. - Access mode: http://dataaspirant.com/2015/01/24/recommendation-engine-part-1/ (20.10.2017)

С.А. Онисич, О.П. Солдатова

ВЛИЯНИЕ АЛГОРИТМОВ ОБУЧЕНИЯ НА СХОДИМОСТЬ И ПОГРЕШНОСТЬ МНОГОСЛОЙНОГО ПЕРСПЕТРОНА

(Самарский университет)

Целью данной работы является изучение особенностей обучения многослойного персептрона при использовании различных алгоритмов: алгоритма обратного распространения ошибки и генетического алгоритма. В качестве задачи для проверки используется задача ирисов Фишера как классическая задача классификации. Для рассмотрения различных аспектов используемых алгоритмов входные данные разбиты на два множества: обучающее и тестирующее, по 120 и 30 векторов соответственно.

Персептрон – одна из первых попыток создать искусственную нейронную систему. Математическая модель персептрона была предложена Ф. Розенблат-