Метод коллаборативной фильтрации для масштабируемых рекомендательных систем.

Автор: Кольчугина Е.А., Макарь В.А.
Пензенский государственный университет, к.т.н., доцент
Пензенский государственный университет, магистрант 2-го курса

Источник: Кольчугина Е.А., Макарь В.А. Метод коллаборативной фильтрации для масштабируемых рекомендательных систем // Современные научные исследования и инновации. – Июнь 2012. - № 6 [Электронный ресурс].

Аннотация:
Статья посвящена методу коллаборативной фильтрации для масштабируемых рекомендательных систем.

Постановка задачи

Объем информации во всемирной паутине постоянно увеличивается. В связи с этим пользователи сталкиваются со сложной задачей поиска информации, которая бы соответствовала их предпочтениям. Для ее решения, все более важным становится разработка информационных систем, позволяющих автоматизировать процесс рекомендации объектов потребления. Создание качественной рекомендательной системы, с одной стороны, может привести к увеличению прибыли поставщиков услуг (товаров), а с другой ? экономии времени потребителя на поиски необходимой вещи. Таким образом, подобные системы очень важны как для экономики, так и для информационного общества в целом.

При создании рекомендательных систем чаще всего используют методы коллаборативной фильтрации [1]. Эти методы определяют возможные предпочтения пользователя, основываясь на известных предпочтениях других пользователей. В их основе лежит идея, что если пользователи А и Б оценили несколько объектов приблизительно одинаково, то и последующие объекты они будут оценивать так же.

Примерами успешных рекомендательных систем, которые основаны на коллаборативной фильтрации, являются системы компаний Amazon и Netflix. Однако, несмотря на активные научные исследования в области коллаборативной фильтрации с 1994 года и успешное использование коллаборативной фильтрации в системах электронной коммерции и цифровой дистрибуции, эта область продолжает оставаться сферой научных интересов. Это связано с рядом проблем, для которых не было найдено удовлетворительных решений [1]. Одной из таких проблем является проблема масштабирования систем коллаборативной фильтрации, которая тесно связана с проблемами анализа данных большой размерности. При анализе данных большой размерности возникает ряд трудностей, получивших название «проклятие размерности». Эффективными методами для решения задач коллаборативной фильтрации данных большой размерности являются методы поиска приближенных ближайших соседей [2],[3] и методы кластерного анализа [5]. Применительно к

коллаборативной фильтрации исследователями предлагалось использовать эти методы по отдельности. Однако нахождение приближенных ближайших соседей позволяет существенно ускорить процесс кластеризации.

Метод коллаборативной фильтрации LSHClust

На основе идеи использования приближенных ближайших соседей для ускорения кластеризации был разработан метод коллаборативной фильтрации LSHClust, который состоит из двух этапов: создание модели (выполняется в фоновом режиме) и формирование списка рекомендаций или прогнозирование оценок (выполняется в режиме реального времени).

Этап создания модели состоит из следующих шагов:

Выбор параметров для алгоритма кластеризации. Кластеризация пользователей коллаборативной системы с помощью алгоритма кластеризации LSH-CLC, который разработан для реализации нового метода. Создание модели, которая представляет собой s векторов {c1 , c2 , ... , cs}, каждый из которых имеет длину m и является центроидом соответствующего кластера. Формирования списка рекомендаций осуществляется для активного пользователя, который характеризуется вектором предпочтения (r1, r2,...,rm), где ri -- оценка, поставленная i-му объекту. Для того чтобы сформировать список рекомендаций необходимо выполнить следующие шаги: Вычислить коэффициент сходства между вектором предпочтения активного пользователя и центроидами s кластеров. В качестве коэффициента сходства может использоваться коэффициент корреляции Пирсона. Выбрать центроиды, для которых коэффициент сходства превышает порог d. Вычислить прогнозируемые значения оценок с помощью формулы:

Алгоритм иерархической кластеризации LSH-CLC.

Основой метода LSHClust является алгоритм кластеризации пользователей LSH-CLC. Данный алгоритм основан на методе локально-чувствительного хеширования [7], который решает задачу поиска приближенных ближайших соседей. На основе найденных ближайших соседей (схожих пользователей) впоследствии формируются кластеры (группы по интересам). Шаги алгоритма LSH-CLC: Для каждого пользователя u вычислить l сигнатур g1(Iu), g2(Iu), ..., gl(Iu), где Iu - множество объектов, которые оценил пользователь. Затем добавить пользователя u в хеш-таблицу Hi так, чтобы сигнатура gi(Iu) была ключом. Таким образом, ячейка хеш-таблицы с ключом gi(Iu) будет представлять собой множество пользователей (приближенных ближенных ближайших соседей). В качестве семейства локально-чувствительных функций используется семейство, образуемое методом MinHash [8]. Найти кластеры, которые удовлетворяют следующим условиям: кластеры содержат пользователей, для которых совпадают сигнатуры по крайней мере в одной хеш-таблице; размер обоих кластеров не превышает N; сходство между центроидами кластеров не меньше t. Объединить кластеры, найденные на шаге 2. Если есть кластеры, разрешенные для объединения, и t > tmin, уменьшить t и перейти на шаг 2. Иначе прервать работу алгоритма.

Для проведения экспериментов использовался популярный набор данных, который был собран в результате работы проекта MovieLens и был опубликован исследовательской лабораторией GroupLens Research. Набор данных MovieLens содержит оценки различным фильмам. Он включает 6040 пользователей, 3900

фильмов и 1000209 оценок. Для экспериментов история оценок каждого пользователя была разделена на два непересекающихся множества: обучающее и тестовое. Тестовое множество было сформировано путем отбора 20% случайных оценок из всего множества оценок пользователя. Таким образом, обучающее множество включало 80% оценок от начальных данных. Используя обучающее множество, создавалась модель коллаборативной фильтрации. После создания модели, она оценивалась с помощью тестового множества. На рисунке 1 показана зависимость качества прогнозирования от параметра tmin. Точность прогнозирование оценивается с помощью среднего абсолютного отклонения значений тестовых оценок от прогнозируемых оценок. Чем меньше значение MAE, тем лучше.

Заключение.

В статье был представлен метод коллаборативной фильтрации LSHClust, который достаточно прост для реализации, обладает хорошими показателями масштабируемости. При этом разработанный метод позволяет достичь точности прогнозирования оценок, сравнимой с одним из самых точных методов (основанным на SVD). Большое количество параметров позволяет гибко настроить метод под различные рекомендательные системы.

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. Al, Mamunur Rashid et al. "ClustKNN: a highly scalable hybrid model- memory-based CF algorithm." KDD Workshop on Web Mining and Web Usage Analysis. ACM, 2006.

4. Ungar, Lyle H, and Dean P Foster. "Clustering Methods for Collaborative Filtering." AAAI Workshop on Recommendation Systems pp.1 (1998) : 1-16.

5. Han, S et al. "RecTree : An Efficient Collaborative Filtering Method." Matrix (2001) : 141–151.