УДК 004.67

А.В. Григорьев, Е.А. Заплетин
Донецкий Национальный Технический Университет
кафедра прикладной математики и информатики
E-mail: zapletin.yevhenii@gmail.com

ОСНОВНЫЕ ПОДХОДЫ ПРИ СОЗДАНИИ РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ

Аннотация

Григорьев А.В., Заплетин Е.А. Основные подходы при создании рекомендательных систем. *Рассмотрена классификация подходов для создания рекомендательных систем. Определено назначение и преимущества каждого из подходов. Определены основные проблемы при реализации данных подходов.*

Общая постановка проблемы

Основная задача рекомендательных алгоритмов: анализировать запросы пользователя и на основе этих данных предугадывать его действия.

Наиболее простой способ рассмотреть основную проблему, которую решают алгоритмы рекомендаций можно на примере. В сервисах которые предлагают большой выбор товаров или контента существует проблема длинного хвоста [1]. Данная проблема подходит под категорию распределение Парето (отношение 20% к 80%). Так например 20% видов товаров в интернет-магазинах делают 80% выручки. Пользователи покупают в основном популярные товары, а про существование оставшихся 80% товаров пользователь может даже не знать. Таким образом появляется проблема, когда пользователь может не знать о существовании товара только потому, что он не популярен среди большинства пользователей. Для решения данной проблемы созданы алгоритмы рекомендаций которые персонализируют предложения исходя из данных о пользователях.

Рис. 1. Эффект длинного хвоста.

Исследования

Существует несколько типов рекомендационных систем[2]:

- Системы основанные на данных о действиях пользователей (collaborative systems);
- Системы основанные на данных о свойствах объектов рекомендаций и свойствах пользователей (content-based systems);
- Системы основанные на знаниях связанных с объектом рекомендаций (knowledge-based systems);
- Гибридные системы (hybrid systems).

1.Системы, основанные на данных о действиях пользователей

Системы используют только информацию о действиях пользователя. Основная идея следующая - если пользователи U1 и U2 купили книгу B1, а пользователь U2 купил еще и книгу B2, то логично предположить что книга B2 будет также интересна пользователю B1.

Стоит заметить, что данные системы не учитывают знания и данные про свойства пользователей и предметов рекомендаций. То есть неважно какого именно жанра будут купленные книги и также не имеет значения что пользователь указал в своем профиле что ему интересны книги определенного автора.

Плюсы:

• Нет необходимости выявлять свойства объектов;

• Нет необходимости обратной связи от пользователя;

• Обучения на множествах данных без участия человека;

• Нет разницы в предсказании любых типов объектов;

• Учитывается автоматическая обратная связь от пользователей(просмотры, установки и т.д.).

Минусы:

• “Холодный старт” для новых пользователей. Алгоритм будет работать неверно для новых пользователей, информации о которых очень мало;

• “Холодный старт” для новых предметов. Алгоритм будет работать неверно для новых предметов, информации о которых очень мало;

• Необходимость обрабатывать наибольшие массивы данных по сравнению с другими системами;

• Интересы пользователей усредняются;

• Рекомендации не основываются на интересах конкретного пользователя;

• Рекомендует в основном популярные объекты;

• Сложно указать причину рекомендации.

2. Системы, основанные на данных о свойствах объектов рекомендаций и свойствах пользователей

Системы данного типа не используют данные о действиях пользователя и только анализируют свойства рекомендуемых предметов и свойства текущего пользователя.

Данные свойства могут быть введены вручную(например пользователь указывает какой жанр книг ему интересен или администратор сайта вводит информацию про книгу: жанр, автор,

год издания, и т.д.), автоматически извлекаться из рекомендуемых предметов(например анализ длины текстов в книгах и разделения их на группы: короткие, средние, длинные) и нахождения новых признаков при помощи алгоритмов кластеризации анализируя исходные данные.

Плюсы:

• Нет “холодного старта” для новых предметов. Новые предметы сразу участвуют в процессе рекомендаций.

• Рекомендует непопулярные объекты;

• Рекомендации основываются на интересах конкретного пользователя;

• Легко указать причину рекомендаций;

• Можно использовать алгоритмы кластеризации для поиска новых свойств объектов.

Минусы:

• “Холодный старт” для новых пользователей. Алгоритм будет работать неверно для новых пользователей, информации о которых очень мало;

• Необходимо участие человека для выявления свойств объектов;

• Интересы пользователей усредняются;

• Желательно явное указание пользователем своих интересов;

• Не используется обратная связь от пользователей(оценки, нравится/не нравится, установки, просмотры и т.д.).

3. Системы, основанные на знаниях связанных с объектом рекомендаций

Системы из этой группы используют знания про сферу предполагаемых продуктов. Данные системы очень похожи на системы основанные на свойствах пользователя и предметов рекомендаций, и в некоторых источниках данный тип систем является подтипом систем основанных на свойствах.

Основной принцип работы данных систем основан на активном взаимодействии с пользователем(получения обратной

связи) и знаниях о сферах рекомендации. Данный тип систем используют в том случае, когда тип 1 и 2 использовать невозможно из-за небольшого количества данных. Примером может служить система рекомендации фотоаппаратов. Как правило, люди покупают новый фотоаппарат только раз в несколько лет, поэтому у каждого конкретного магазина будет недостаточно данных для рекомендации на основе действий пользователя. Также фотоаппараты имеют большое количество характеристик, которые имеют различную важность для пользователей. Для фотографов которые снимают в студии наиболее важно качество съемки, в тоже время для фотографов которые снимают на улице и ведут активный образ жизни также важен вес фотоаппарата, удобство переноски, размер памяти и время автономной работы.

Система рекомендаций данного типа активно взаимодействует с пользователем, но не просит ввести все параметры пользователя вручную. В примере с фотоаппаратами система будет спрашивать пользователя где будет происходит съемка, портретную или пейзажную съемку предпочитает пользователь. Анализируя ответы пользователя система определить наиболее важные для данного пользователя свойства и их оптимальные значения. Знания про сферу рекомендаций изначальной устанавливает администратор системы.

Плюсы:

• Нет “холодного старта” для новых предметов. Новые предметы сразу участвуют в процессе рекомендаций.

• Нет “холодного старта” для новых пользователей. Новые пользователи сразу получают правильные рекомендации.

• Рекомендует непопулярные объекты;

• Рекомендации основываются на интересах конкретного пользователя;

• Легко указать причину рекомендаций;

• Рекомендации учитывают интересы конкретного пользователя;

• Интересы пользователей не усредняются.

Минусы:

• Очень сложно и трудозатратно выделить знания о предметах;

• Добавления новых объектов требует участия человека;

• Практически нет автоматизации в обучении системы;

• Необходимо много обратной связи от пользователя;

• Желательно явное указание пользователем своих интересов;

4. Гибридные системы

Приведенные ваше системы используют различные подходы к анализу данных и предсказанию результата, тем самым решаю задачу различными способами. На практике, не один из способов в одиночку не может рекомендовать товары или контент с большой эффективностью. Поэтому реальные системы рекомендаций всегда используют несколько подходов к анализу и к рекомендациям.

Гибридные системы объединяют все описанные подходы в единый алгоритм и имеют наибольшую эффективность. Также данные системы наиболее сложны в реализации и проектировании. Сложность заключается в адаптации алгоритмов к конкретной сфере применения рекомендационной системы. Так например в сферах где каждый пользователь сам оценивает предлагаемый контент (каталог фильмов с оценками) наиболее важной частью являются оценки пользователей и анализ их действий. А при использовании рекомендательной системы на сайте интернет магазинов наиболее важными задачами для администратора системы является увеличение прибыли, а значит при создании рекомендации необходимо учитывать фактор цены товаров и необходимости их продать.

Плюсы:

• Нет “холодного старта” для новых предметов. Новые предметы сразу участвуют в процессе рекомендаций.

• Нет “холодного старта” для новых пользователей. Новые пользователи сразу получают правильные рекомендации.

• Рекомендует непопулярные объекты;

• Рекомендации основываются на интересах конкретного пользователя;

• Рекомендации учитывают интересы конкретного пользователя;

• Интересы пользователей не усредняются;

• Наиболее эффективны.

Минусы:

• Большая сложность в реализации;

• Необходимо объединять воедино несколько алгоритмов в один;

• Очень сложно и трудозатратно выделить знания о предметах;

• Сложно указать причину рекомендаций;

• Интересы пользователей усредняются.

• Желательно обратная связь от пользователей;

• Желательно участие человека при добавлении новых товаров.

Вывод

Существует несколько способов реализации алгоритма рекомендаций. Однако не один из способов в одиночку не может дать хороших результатов. Поэтому на практике всегда необходимо использовать гибридную систему, которая объединяет несколько способов.

Список литературы

1. Leskovec J. Mining of Massive Datasets [Текст] / J.Leskovec, A. Rajaraman, J. Ullman // Cambridge University Press. – Cambridge : 2011.

2. Francesco Ricci and Lior Rokach and Bracha Shapira, Introduction to Recommender Systems Handbook, Recommender Systems Handbook, Springe