3. Введение в основные рекомендательные алгоритмы

3.1. Системы коллаборативной фильтрации

Основная идея таких систем состоит в том, что если пользователи имели одинаковые интересы в прошлом, то в будущем их предпочтения также будут совпадать. В качестве примера рассмотрим книжный интернет-магазин. Пусть в прошлом истории покупок пользователей А и Б в данном магазине сильно пересекались. При появлении пользователя А на сайте мы хотим предложить ему новую книгу, которую он еще не читал, а пользователь Б, как раз, недавно приобрел книгу, которую А еще не видел. В такой ситуации будет разумно предложить пользователю А прочитать ее.

Выше описан принцип работы класса алгоритмов коллаборативной фильтрации, которые в англоязычной литературе называются user-based, то есть основанные на статистике о пользователях. Как видно из примера, мы сравниваем похожесть между пользователями, основываясь на рейтингах оцененных объектов. В то же время, почему бы нам не использовать ту же статистику для сравнения продуктов между собой, а потом сопоставить результаты двух подходов.

Такой метод, основанный на сравнение схожести объектов, называется item-based. Здесь основная идея состоит в том, что если пользователям, которые оценили два продукта, понравились оба, то пользователям, которые попробовали только один, можно предлагать второй, который вероятнее всего, им понравится. То есть, если в примере с книжным магазином мы заметили, что пользователи, которые покупали книгу А, также покупали книгу Б, то тем потребителям, которые уже купили А, но еще не обратили внимание на Б разумно будет ее предложить.

Первой исследовательской работой по рекомендательным системам считается работа Lee Giles "An autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications"1998 года. Однако первой печатной правильнее назвать работу 1992 года David Goldberg, David Nichols "Using collaborative filtering to weave an information Tapestry"

В данной работе описан принцип работы экспериментальной почтовой системы Taperstry. Разработчики Tapestry первыми использовали термин "коллаборативная фильтрация"как метод сбора качественных данных. Данная система была разработана в Xerox PARC как способ

обработки большого количества сообщений электронной почты и сообщений, отправляемых в группы новостей. Особенностью данной системы было то, что система собирала и анализировала данные о реакции людей на прочитанные ими документы, в следствие чего процесс фильтрации стал более эффективным.

Одновременно с Taperstry развивались и другие рекомендательные системы на основе коллаборативной фильтрации:

- В 1995-1996 годах были разработана сразу три системы для рекомендации музыки: Helpful Online Music Recommendations, Ringo, Firefly.
- Также в то время активно развивались системы рекомендаций наиболее интересных и популярных страниц в интернете: Point's Top 5%, PHOAKS (People Helping One Another Know Stuff), Webdoggie, Alexa Internet.

Метод Item-based был изобретен и использован Amazon.com в 1998 году. Впервые представлен публике на научной конференции в 2001, а его авторы в 2016 получили награду Test of Time.

Данный алгоритм помог справиться с некоторыми из проблем, имевшимися у методов, основанных на схожести пользователей:

- системы работали плохо, когда у них было много продуктов, но сравнительно немного оценок
- трудоемко вычислить сходства между всеми парами пользователей
- профили пользователей быстро менялись, и всю модель необходимо было пересчитывать

Но вопросов, с которыми приходится столкнуться разработчику в процессе создания такой системы остается еще много. Вот одни из них:

1) Как для данного пользователя, для которого мы хотим сделать рекомендацию, определить пользователей, которые имеют схожие предпочтения?
2) Как измерять схожесть между пользователями?
3) Что делать если у нас имеется мало данных о рейтингах?

Преимущества метода:

1) Является достаточно универсальным подходом, поэтому часто дает высокие результаты.

2) Для работы данного метода не нужна детальная информация о продуктах. В примере с книжным магазином - автор, жанр, описание книги. Вместо этого используется как история оценок самого пользователя, так и других пользователей.

Недостатки метода:

1) Как работать с новыми пользователями, для которых еще нет истории покупок (задача холодного старта).

2) Что делать с новыми объектами, которые еще никто не оценил.

3) Ресурсоемкость вычислений, которая замедляет время работы системы.

4) Необходим большой объем данных для высокой точности предсказаний. Ниже в процессе более детального рассмотрения этого типа алгоритмов я приведу ответы на все эти вопросы, а также расскажу, как обходить проблемы данного метода.

3.2. Системы фильтрации на основе содержания

Этот тип систем основан на наличие информации об описании и профиле, состоящем из набора характеристик элемента. Если снова рассмотреть пример книжного магазина, то в качестве характеристик можно взять жанр, тему или автора книги. Затем для каждого пользователя создается профиль путем присвоения характеристик сходных с характеристиками элементов, исходя из анализа его поведения в прошлом, либо явно спрашивая о его предпочтениях. Далее пользователю рекомендуются объекты, похожие на те, которые этот пользователь уже употребил, либо указал как предпочтительные. Похожести оцениваются по признакам содержимого объектов.

В примере с книжным магазином система могла определить, что автору нравятся детективы и новеллы определенных авторов, и в последствие рекомендовать книги этих жанров или авторов.

В процессе изучения систем фильтрации на основе содержания также возникают интересные вопросы:

1) Как система может автоматически создать профиль пользователя и затем улучшать в процессе обновления данных?

2) Как определить какой элемент соответствует предпочтениям пользователя?

3) Как автоматически извлекать информацию о продукте, чтобы избежать ручного заполнения?

Преимущества метода:

1) Не требует большой группы пользователей для достижения высокой точности рекомендаций.

2) Новые элементы можно рекомендовать сразу, как только у них появляются заполненные характеристики.

Недостатки метода:

1) Сильная зависимость от предметной области, полезность рекомендаций ограничена.

2) Профиль пользователей и элементов должен состоять из одинакового набора характеристик, чтобы их можно было сравнивать.

3.3. Системы, основанные на знаниях

Рекомендации, основанные на знаниях, используются обычно в таких областях, как электроника, где покупатели совершают покупки раз в пару лет, так как в данной области мы не можем положиться на историю покупок, которая используется в качестве входных данных для методов коллаборативной фильтрации и методов, основанных на содержании.

Рассмотрим для примера рекомендательную систему, которая помогает пользователя выбрать фотокамеру. Обычный пользователей покупает новую камеру только один раз в несколько лет. Таким образом, рекомендательная система не может построить профиль пользователя или предложить камеры, которые понравились другим пользователям, так как в противном случае предлагаться будут только бестселлеры. Поэтому алгоритмы, основанные на знаниях, обычно используют дополнительные данные, как о пользователях, так и о самих продукта, для формирования списка рекомендаций. В области фотокамер такая система

может использовать детальную информацию о характеристиках камер, таких как разрешение, вес, цена.

Просто представлять продукт, удовлетворяющий выбранному пользователем набору характеристик, является недостаточным, поскольку в таком случае каждый пользователь получает одинаковые рекомендации с теми, кто выбрал такой же набор характеристик. Таким образом, данные системы должны не просто собирать информацию о желаемых характеристиках, а также формировать некоторый профиль пользователя.

Поэтому важным аспектом построения таких систем является настройка взаимодействия между пользователем и системой. Если вспомнить пример с книжным магазином и алгоритмом коллаборативной фильтрации, то можно заметить, что пользователь может взаимодействовать с программой ограниченным числом способов. Множество приложений допускает только возможность ставить рейтинги от 1 до 5. Возвращаясь к примеру с фотокамерой, когда у нас нет информации об истории покупок пользователя, нам необходимо настроить диалог между пользователем и системой, в процессе которого программа задаст вопрос о требованиях покупателя, таких как максимальная цена, минимальное разрешение и т.д.

Такой подход требует не только детального технического понимания характеристик продукта, но также строит приблизительный сценарий на основе выбранных характеристик. В такой ситуации ограничивающие факторы могут быть использованы для описания контекста, в котором определенные характеристики являются релевантными для покупателя. Например, камера высокого разрешения является более предпочтительной, если пользователь планирует печатать фотографии большого размера.

В целом при рассмотрении систем такого вида возникает достаточно много вопросов:

1) В каких областях может быть применим данный метод?

2) Как получить профиль пользователя в областях, где нет истории его покупок, и как учесть предпочтения пользователя?

3) Как настроить взаимодействие с пользователями?

4) Каким образом можно персонифицировать процесс взаимодействия, чтобы максимизировать точность процесса сбора информации о предпочтениях пользователей?

Преимущества метода:

1) Требования пользователей могут быть определены точнее, благодаря явному взаимодействию.

2) Метод дает хорошие результат в сфере, где нет достаточной информации об истории покупок.

Недостатки метода:

1) От пользователя требуются дополнительные действия, чтобы система могла собрать данные о его предпочтениях.

2) Данные о требованиях пользователя могут быть неправильно интерпретированы системой.

3.4. Гибридные системы

Каждый из вышеописанных методов имеет свои преимущества и недостатки в зависимости от поставленной задачи. Достаточно очевидным решением всех этих проблем является объединение различных подходов для того, чтобы обеспечить большую точность рекомендаций. О том, как измерять точность рекомендаций, мы поговорим в разделе о способах оценки качества рекомендаций.

Если, например, у нас есть данные об описании продуктов, профиль пользователей и история его покупок, то мы можем улучшить рекомендательную систему путем объединения методов коллаборативной фильтрации и алгоритмов фильтрации по содержанию. Таким образом, в случае появления нового пользователя в системе, о котором нет истории покупок, мы сможем использовать рекомендации на основе алгоритмов фильтрации по содержанию, а в случае большого объема статистических данных строить более точный прогноз, используя методы коллаборативной фильтрации.

Несмотря на то, что гибридные системы помогают бороться с недостатками описанных ранее методов, они все же оставляют достаточно вопросов, на которые нужно ответить при проектировании такой системы:

1) Какие подходы могут быть объединены, и какие условия должны выполняться, чтобы это могло быть сделано?