

Рекомендательные системы: обзор основных постановок и результатов

Кутянин А.Р.

В работе описываются основные алгоритмы рекомендательных систем. Приводится содержательная постановка задачи нахождения N лучших рекомендаций, описываются преимущества и недостатки каждого из названных методов. Приводится обзор работ и результатов, начиная с возникновения первых рекомендательных систем и до настоящего времени.

Ключевые слова: Рекомендательная система, рекомендации, коллаборативная фильтрация, item-based, user-based, фильтрация на основе содержания, фильтрация на основе знаний, гибридные рекомендательные системы.

1. Введение

Мы живем в мире материального и информационного изобилия. Поэтому каждый день мы делаем выбор и задаемся следующими вопросами:

Какую камеру, книгу, футболку мне бы купить? Какое образование выбрать для своих детей? Куда поехать в отпуск? Какое кино посмотреть? И т.д.

Каждый из нас с легкостью может дополнить этот список. Во время принятия решения о покупке того или иного продукта, либо о том, куда сходить сегодня вечером, мы часто обращаемся за помощью: спрашиваем совета у друзей, ищем информацию в интернете, строим деревья решений, нанимаем команду экспертов, доверяемся инстинктам или просто следуем за толпой. При этом каждый сталкивался с ситуацией, когда советы друзей оказывались бесполезными, либо, просидев несколько часов в интернете, мы не могли найти нужную информацию. В подобных ситуациях нам трудно было заполучить нужный совет, либо на это требовалось много сил и времени.

Как результат, все мы, наверняка, задумывались о такой программе, которая могла бы давать нам индивидуальные рекомендации, способные помочь в принятии решений.

Создание алгоритма для такой программы является главной задачей области, которая занимается разработкой рекомендательных систем. А уже цель самой системы - это обеспечить большую группу пользователей доступной информацией и высококачественными рекомендациями.

В связи с повсеместным внедрением интернета и наличием большого объема статистических данных о предпочтениях пользователей ученые всего мира стали очень активно заниматься разработкой таких алгоритмов в последние двадцать-тридцать лет. А вследствие широкой применимости такие программы оказались очень действенными. Например, интернет-магазин Amazon.com делает рекомендации даже относительно покупки автомобильных двигателей.

Конечно, можно возразить, что такие системы могут быть полезны только тем, кто не может или не хочет позволить себе высококачественные советы экспертов в данной области. Но это верно лишь частично и только в некоторых областях, таких как финансы или медицина. В то же время в других сферах может и не быть возможности обратиться к квалифицированному эксперту, либо мнение других пользователей, воспользовавшихся данными продуктом или услугой, может больше влиять на наше решение. Таким образом, возникает вопрос, можем ли мы, имея большую базу данных о предпочтениях пользователей, разработать программу, которая будет давать более качественные рекомендации, чем человек.

Вопрос предоставления доступных, индивидуальных, а главное высококачественных рекомендаций является основным в области рекомендательных систем и ставит перед нами много интересных задач, как с технической точки зрения, так и с точки зрения психологии. С технической точки зрения мы пытаемся разработать алгоритм, который сможет наиболее точно и логично спрогнозировать пожелания пользователей на основе имеющихся данных. А психологические факторы мы должны учесть, когда будем продумывать то, каким образом программа будет взаимодействовать с конечным потребителем. Коммуникация между потребителем и системой должна быть настроена таким образом, чтобы потребитель доверял рекомендациям, которые она дает, а также своим решениям, принятым на основе этих рекомендаций. Также сложность заключается в том, что пользователи не всегда точно знают, что они хотят. Поэтому при создании рекомендательного алгоритма важно

учитывать то, каким образом программа будет собирать данные о потребностях пользователей, в каком виде представлять обратную связь, а также как все это будет влиять на решения пользователей.

Несмотря на все эти трудности, рекомендательные системы активно развиваются с момента их первого появления двадцать-тридцать лет назад. Область применения таких алгоритмов очень широкая. На сегодняшний день их можно встретить в областях занимающихся недвижимостью, финансами, электроникой, фильмами, книгами, музыкой новостями и т.д. Каждый день мы все чаще начинаем встречать образцы таких программ, например, на сайтах Amazon.com, Last.fm, Ozon.ru и других. По статистике на Amazon.com за год совершается около 5 млрд. транзакций, а около 30% из них идет от рекомендаций. Таким образом, приблизительно 125 млн. покупок в месяц совершается через рекомендации, а сколько их выдается системой и представить сложно.

В данной статье рассказывается об основных рекомендательных алгоритмах, а также приводятся примеры самых свежих разработок в данной области.

Целью данной работы является обзор основных рекомендательных алгоритмов, а именно рассказать о принципах работы методов коллаборативной фильтрации, о системах фильтрации на основе содержания, а также на основе знаний о пользователях, познакомить читателей с преимуществами и недостатками каждого из методов. Знакомство с ними наиболее важно, так как их чаще всего используют, и, по большому счету, все новые разработки являются усовершенствованием данных подходов

2. Постановка задачи

Формализуем нашу задачу. Пусть у нас есть множество пользователей $U = \{u_1, u_2, \dots, u_n\}$, множество объектов $P = \{p_1, p_2, \dots, p_m\}$ и матрица рейтингов $R = (r_{i,j})$ размера $n \times m$, где $i \in 1 \dots n$, $j \in 1 \dots m$. Возможными значениями рейтингов могут быть числами от 1 (совсем не понравилось) до 5 (очень понравилось), либо значениями понравилось/не понравилось, выраженные 0 и 1, а также другие. Если пользователь i никак не оценил элемент j , то на соответствующем месте $r_{i,j}$ будет стоять пустое значение.

Описанную матрицу рейтингов можно представить в виде таблице, показанной на рисунке. Обозначим через $\hat{r}_{i,j}$ наш прогноз относительно того, какую оценку пользователь i поставит продукту j . Наша задача наилучшим образом предсказать, какие оценки $r_{i,j}$ должны стоять на

Продукты	Продукт 1	Продукт 2	Продукт 3	Продукт 4	Продукт 5	Продукт 6
Пользователь 1	2	4	?	3	4	?
Пользователь 2	1	?	1	4	5	?
Пользователь 3	3	4	?	1	4	?
Пользователь 4	?	?	4	2	?	2
Пользователь 5	?	4	5	3	?	?

месте пропусков в матрице рейтингов, то есть рассчитать $\hat{r}_{i,j}$. Затем для каждого пользователя u на основе спрогнозированных оценок $\hat{r}_{i,j}$, нам нужно сформировать список из N продуктов, которые наиболее точно удовлетворяют предпочтениям пользователя, и, которые еще не были им оценены. Список этих N продуктов обозначим через N -мерный вектор $(p_{i_1}, p_{i_2}, \dots, p_{i_N})$. Таким образом, математическая задача звучит так:

Дано:

$U = \{u_1, u_2, \dots, u_n\}$ - множество пользователей,

$P = \{p_1, p_2, \dots, p_m\}$ - множество продуктов,

$R = (r_{i,j})$ - матрица рейтингов размера $n \times m$, где на месте $r_{i,j}$ будет стоять некоторое число, если пользователь u_i оценил продукт p_j и пусто в противном случае.

N - требуемое число рекомендаций, которые хотим получить от системы.

Требуется найти:

Для данного пользователя u_i найти N -мерный вектор $(p_{i_1}, p_{i_2}, \dots, p_{i_N})$, где продукты $p_{i_k}, k \in N$ еще не оценены этим пользователем, то есть в матрице рейтингов $R = (r_{i,j})$ стоит пусто на месте r_{i,i_k} , а также, чтобы эти продукты наиболее точно удовлетворяли предпочтениям пользователя, то есть прогнозные рейтинги \hat{r}_{i,i_k} были наибольшими.

Алгоритмы, которые это делают, могут быть очень разными, и использовать различные входные данные. Одни из них формируют рекомендации, основываясь только на данных об известных рейтингах. Другие используют дополнительные характеристики продуктов, на основе рейтингов определяют, какие из этих характеристик наиболее точно удовлетворяют предпочтениям пользователя, а затем подбирают продукты с такими характеристиками.

А теперь более подробно поговорим о каждом из этих алгоритмов, проблемах, с которыми сталкивается исследователь в процессе их использования, преимущества и недостатках каждого метода.

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 Tapestry. 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%, PNOAKS (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 пользователь 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 продуктах, 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 применен 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 результаты 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?

- 2) Должны ли разные техники использоваться в разных ситуациях, либо результат каждой должен браться с определенным весом?
- 3) Как результат разных методов должны быть взвешены, чтобы на выходе получить один результат?

Список литературы

- [1] Jannach D., Zanker M., Felfering A., Friedrich G., Recommender Systems: An Introduction.. — М.: Cambridge University Press, 2011.
- [2] Ricci F., Rokach L., Shapira B., Kantor P. Recommender Systems: Handbook.. — М.: Springer, 2010.
- [3] Bollacker K.D., Lawrence S., Giles C.L., CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications // 2nd international conference on Autonomous agents — 1998. — С. 116-123.
- [4] Goldberg D., Nichols D., Oki B.M., Terry D. Using collaborative filtering to wave an information tapestry // Communication of the ACM — 1992. — 35(12):61-70.
- [5] Linden G., Smith B., York J., Amazon.com recommendations: item-to-item collaborative filtering // Internet Computing — IEEE 7 2003 — С. 76-80.
- [6] Miller B., An open Architecture for collaborative filtering // GroupLens — 1995
- [7] Gomez-Uribe C., Hunt N., The Netflix Recommender System: Algorithms, Business Value, and Innovation // ACM Transactions on Management Information Systems — 2015
- [8] Melville P., Mooney R.J., Nagarajan R. Content-boosted collaborative filtering for improved recommendations // in Proceedings of the National Conference on Artificial Intelligence — 2002 — С. 187-192.
- [9] Shi Y., Larson M., Hanjalic A., Collaborative Filtering Beyond the User-Item Matrix: A Survey of the State of the Art and Future Challenges // ACM Comput. Surv. — 2014 — С. 3:1-3:45.
- [10] Gunawardana A., Shani G., A survey of accuracy evaluation metrics of recommendation tasks // The Journal of Machine Learning Research — 2009 — С. 2935-2962.

- [11] Herlocker, J.L., Konstan J.A., Terveen L.G., Riedl J.T., Evaluating collaborative filtering recommender systems // ACM Transactions on Information Systems (TOIS) — 2004 — C. 5–53.
- [12] Ge, M., Delgado-Battenfeld C., Jannach D., Beyond accuracy: evaluating recommender systems by coverage and serendipity // in Proceedings of the fourth ACM conference on Recommender systems — 2010 — C. 257–260.
- [13] McNee, S., Kapoor N., Konstan J.A., Don't look stupid: avoiding pitfalls when recommending research papers // in Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work — 2006 — C. 171–180.
- [14] McNee, S., Albert I., Cosley D., On the Recommending of Citations for Research Papers // in Proceedings of the ACM Conference on Computer Supported Cooperative Work — 2002 — C. 116–125.
- [15] Geyer-Schulz A., Hahshler M., Comparing two recommender algorithms with the help of recommendations by peers // in Proceedings of the WEBKDD 2002 - Mining Web Data for Discovering Usage Patterns and Profiles — 2003 — C. 137–158.
- [16] Geyer-Schulz A., Hahshler M., Evaluation of recommender algorithms for an internet information broker based on simple association rules and on the repeatbuying theory // in Proceedings of the 4th WebKDD Workshop: Web Mining for Usage Patterns & User Profiles — 2002 — C. 100–114.
- [17] Geyer-Schulz A., Hahshler M., Jahn M., A customer purchase incidence model applied to recommender services // in Proceedings of the Third International Workshop on Mining Web Log Data Across All Customers Touch Points — 2002 — C. 25–47.
- [18] Geyer-Schulz A., Hahshler M., Jahn M., Educational and scientific recommender systems: Designing the information channels of the virtual university // International Journal of Engineering Education — 2001 — C. 153–163.
- [19] Huang Z., Chung W., Ong T.H., Chen H., A graph-based recommender system for digital library // in Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries — 2002 — C. 65–73.
- [20] Balabanovic M., An Adaptive Web Page Recommendation Service // in Proceedings of the First international Conference on Autonomous Agents — 1997

- [21] Menczer F., ARACHNID: Adaptive retrieval agents choosing heuristic neighborhoods for information discovery // In Machine Learning: Proceedings of the fourteenth International Conference — 1997 — C. 227–235.
- [22] Liang Y., Li Q., Qian T., Finding relevant papers based on citation relations // in Proceedings of the 12th international conference on Web-age information management Conference — 2011 — C. 403–414.
- [23] Pennock D.M., Horvitz E., Lawrence S., Giles C.L., Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach // in Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence — 2000 — C. 473–480.

Recommender systems: overview of main statements and results Kutyanin A.R.

Main recommender algorithms are considered in this paper. A meaningful formulation of the problem of finding N best recommendations is given, pros and cons are described for all mentioned approaches. Also reader can find papers and results overview starting from appearance of the first recommendation systems up to the present.

Keywords: Recommender systems, recommendation, collaborative filtering, item-based, user-based, content-based filtering, knowledge-based filtering, hybrid recommender system.