Санкт-Петербургский государственный университет

Математическое обеспечение и администрирование информационных систем
Системное программирование

Ершов Александр Владимирович

Алгоритм рекомендации научных конференций для учёных

Бакалаврская работа

Научный руководитель:
к. ф.-м. н., доцент Бугайченко Д. Ю.

Рецензент:
инженер-аналитик Дзюба А. А.

Санкт-Петербург
2016

SAINT-PETERSBURG STATE UNIVERSITY

Information Systems Administration and Mathematical Support
Software Engineering

Ershov Aleksandr

Algorithm for a recommendation of scientific conferences for scientists

Graduation Thesis

Scientific supervisor:
assistant professor D. Y. Bugaychenko

Reviewer:
engineer analyst A.A. Dzyuba

Saint-Petersburg
2016

Оглавление

1. Введение

Ученые во время проведения своих исследований сталкиваются с проблемой поиска научных конференций для нахождения релевантных статей и публикации собственных исследований. При этом количество статей, конференций и публикаций постоянно растет. Решением этой проблемы может выступать рекомендательная система, которая для конкретного исследователя будет рекомендовать релевантные для него конференции.

1.1. Рекомендательные системы

Рекомендательные системы предсказывают, какие объекты могут быть интересны пользователю, исходя из его профиля и информации об объектах. Три самых популярных метода реализации — это коллаборативная фильтрация, контентно-ориентированный подход и объединение этих двух методов — гибридный подход.

Коллаборативная фильтрация Основная идея коллаборативной фильтрации состоит в том, чтобы вычислять рекомендации на основе оценок объектов пользователями. При это существуют две разновидности этой техники.

Первая — memory-based. В данной технике строиться матрица "пользователи - объекты", где M_{ui} обозначает рейтинг пользователя u, данный объекту i. Потом для каждого объекта считается рейтинг по формуле

$$r_{ui} = \frac{\sum_{v \in N_i} w_{iv} r_{uv}}{\sum_{v \in N_i} |w_{iv}|}$$

где r_{ui} - предсказанный рейтинг объекта i для пользователя u, w_{iv} - вес между объектами i и v, N_i - k ближайших объектов к i по весу. Знаменатель используется для нормализации и является опциональным. Для измерения веса может использоваться, например, косинус или корреляция. Если вес считается между пользователями, то данный подход называется user-based, если между объектами, то item-based.

Вторая – model-based. В этой технике строится модель, используя алгоритмы машинного обучения такие, как SVD [18], нейронные сети [10], генетические алгоритмы [3] и так далее.

Основная проблема, возникающая при коллаборативном подходе – проблема холодного старта. Это ситуация, когда появляется новый объект, у которого нет оценок, или новый пользователь, который не оценил ни один объект.

Контентно ориентированный подход Контентно-ориентированный подход в отличие от коллаборативной фильтрации использует внутренние свойства объектов, а не оценки пользователей. Для фильмов, например, это могут быть жанр, актёры, длительность и так далее. Объект представляется как вектор, представляющий эти свойства. Соответственно в общем данный подход можно разбить на 3 части [6]:

- Content-analyzer — выделяет свойства и приводит их к числовому виду
- Profile-learner — обучается на этих данных, либо строит матрицу весов
- Filtering component — рекомендует объекты, используя прошлый компонент и профиль пользователя

При данном подходе главной проблемой являются выбор свойств и привидение их к численному виду. Для привидения текстовой информации к векторному представлению часто используется TF-IDF [11].

Гибридный подход Гибридный подход совмещает несколько алгоритмов, для того, чтобы избавиться от их недостатков и совместить достоинства. Существует несколько вариантов объединения результатов, например, взвешенная сумма, объединение или пересечение результатов, передача результатов одного алгоритма в другой.

1.2. Измерение результатов

Методы оценки качества рекомендательных алгоритмов можно разделить на два вида: те, которые используют реальных пользователей (Online), и те, которые используют сохраненные статические данные о пользователях и объектах (Offline). [15]

К первому типу можно отнести A/B тесты и опрос пользователей. A/B тестом называется исследование, при котором пользователей разбивают на две группы. Первая группа использует один алгоритм, вторая другой, и производится измерение качества по выбранной метрике на каждой группе.

Ко второму типу относится кросс-валидация. Это тип исследования качества алгоритма, при котором уже сохраненные, статические данные разбиваются на k частей. Одно часть называется тестовой — на ней проводится измерение качества работы алгоритма по метрике. Остальные части называются тренировочными. На них производится обучение алгоритма.

Метрики качества работы Существует несколько метрик для оценки качества работы рекомендательного алгоритма, среди них можно выделить:

- Среднеквадратичная ошибка (MSE) — $E(\hat{\theta} - \theta)^2$, где $\hat{\theta}$ — предсказанное, а θ — истинное значение
- Средняя абсолютная ошибка (MAE) — $E|\hat{\theta} - \theta|$, где $\hat{\theta}$ — предсказанное, а θ — истинное значение
- Precision-recall кривая — график, где на одной оси находится precision, а на другой — recall. Precision — статистическая метрика, вычисляемая по формуле:

$$Precision = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

Recall также является статистической метрикой:

$$Recall = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

Количественной мерой, показывающей качество алгоритма является площадь под графиком (AU-PR). Максимально возможным значением является единица, в таком случае precision равен единице при любом recall. Это обозначает, что сначала рекомендуются все релевантные объекты.

- ROC-кривая — график, показывающий отношение между True positive rate (True positive rate $= \frac{\text{True positive}}{Positive}$) и False positive rate (False positive rate $= \frac{\text{False positive}}{Negative}$). Мерой качества также является площадь под графиком (AUC). У идеального алгоритма она также равна единице.

2. Постановка задачи

Целью данной работы является:

- Реализовать гибридную рекомендательную систему для предсказания релевантных конференций для исследователей, используя контентно-ориентированный и коллаборативный подход
- Данные для данной работы должны быть взяты с DBLP [2] . DBLP — это система, которая индексирует научные статьи, конференции и авторов. Данная система была выбрана по причине наличия большой базы (3.3 млн. публикаций, 1.7 млн. авторов, 4.7 тыс. конференций на май 2016) и возможности скачать эту базу.
- Провести измерение полученных результатов

3. Обзор рекомендательных систем

В данном обзоре представлены некоторые коллаборативные, контентно-ориентированные и гибридные системы.

3.1. Контентно-ориентированные

В контентно-ориентированных системах используются внутренние свойста объекта. Если одно из свойств представляет из себя текст, то его надо преобразовать к векторному виду, для применения какой-либо математической модели.В работе, описывающей рекомендации статей на основе оценок пользователем предыдущих, это преобразование делается с помощью TF-IDF. [4]

Также в контентно-ориентированных системах применяются методы машинного обучения. Например в системе, которая рекомендует книги, используется наивный байесовский классификатор для классификации текста. [7]

3.2. Коллаборативные

В memory-based варианте коллаборативного подхода сначала строится матрица "пользователи-объекты", а уже на основе ее — матрица весов. При этом если размерность матрицы "пользователи-объекты" очень большая, но данная матрица является разряженной, то имеет смысл делать расчет весов только между объектами, между которыми есть хотя бы один пользователь. Такой подход применяется в системе рекомендаций интернет-магазина Amazon. [5]

Другой способ уменьшения размерности — кластеризация. Сначала данные разбиваются на кластеры, например алгоритмом k-means[1]. А уже алгоритм рекомендации применяется отдельно для каждого кластера. [17]

[1] https://en.wikipedia.org/wiki/K-means_clustering

3.3. Гибридные

Гибридные системы объединяют в себя несколько алгоритмов, как правило коллаборативный и контентно-ориентированный. Существуют несколько способов объединения алгоритмов, например как взвешенная сумма рейтингов [1] или как объединение списков рекомендаций от обоих подходов. [16]

4. Коллаборативный подход

Исходя из постановки задачи алгоритм должен состоять из двух частей — коллаборативного и контентно-ориентированного подхода. В данной части будет рассмотрен коллаборативный.

4.1. Алгоритм

В работе используется item-based memory-based алгоритм, потому что матрица item-item получается более плотная и информативная, чем user-user и сложность по времени и по памяти построения матрицы весов $O(n^2)$ где n — количество объектов для item-based и n — количество пользователей для user-based. А число конференций намного меньше, чем число пользователей.

Алгоритм состоит из 4 частей:

- Вычислить матрицу “пользователи - конференции”
- Вычислить матрицу весов “конференции-конференции” по формуле
$$W_{ab} = \frac{ab}{|a||b|}$$
- косинус между векторами
- Вычислить рейтинг для пользователя по формуле
$$r_{cu} = \frac{\sum_{d \in C_u} w_{cd}}{|C_u|}$$
r_{cu} — рейтинг конференции c для пользователя u. C_u — множество конференций, на которых был пользов атель u. w_{cd} — вес между конференциями c и d. $|C_u|$ - количество конференций, на которых был пользователь u.
- Рекомендовать все объекты, у которых рейтинг выше установленного порога

4.2. Инструменты

Реализация производилась на языке Python, так как он удобен для машинного обучения и научных расчетов из-за существования таких библиотек, как numpy[19] , scipy[13] , scikit-learn[12] .

4.3. Реализация

Матрица “пользователи-конференции” является разреженной (среднее количество пользователей на конференции равно 372, а общее количество пользователей больше миллиона). Поэтому между многими конференциями нет ни одного пользователя, и вес равен нулю. Исходя из этого, вес стоит считать только между теми конференциями, между которыми есть хотя бы один пользователь. А саму матрицу “пользователи-конференции” хранить в виде разреженной матрицы [14] .

Матрица весов “конференции-конференции” была построена по следующему алгоритму:

C - множество всех конференций

U_c - множество пользователей на конференции c

C_u - множество конференций пользователя u

Mod - предпосчитанный заранее массив модулей векторов конференций $Mod_i = |i|$

M - матрица "пользователи - конференции" $M_{cu} = 1$ если пользователь u был на конференции c

W - матрица весов W_{ij} - вес между конференциями i и j

1: **for** $i \in C$ **do**
2: **for** $u \in U_c$ **do**
3: **for** $j \in C_u$ **do**
4: $W_{ij} = \frac{|\{k, M_{ik}=M_{jk}=1\}|}{Mod_i * Mod_j}$
5: **end for**
6: **end for**
7: **end for**

После этого рейтинги для пользователя u могут быть построены по следующему алгоритму:

C - множество всех конференций

C_u - множество конференций пользователя u

W - матрица весов W_{ij} - вес между конференциями i и j

r_u - массив рейтингов конференций для пользователя u

1: **for** $c \in C$ **do**
2: $r_{cu} = \frac{\sum_{d \in C_u} w_{cd}}{|C_u|}$
3: **end for**

5. Контентно-ориентированный подход

5.1. Алгоритм

Для реализации контентно-ориентированного подхода был составлен список из всех слов, входящих в заголовки статей на конференции. Идея состоит в том, чтобы на основе данных слов измерять "похожесть" конференций.

Алгоритм выглядит следующим образом:

- Для каждой конференции составить список слов, входящих в заголовки статей с данной конференции.
- Привести слова к нормальной форме используя стемминг[2].
- Отбросить стоп-слова. В данном случае это слова, которые встречаются больше 50000 раз, это слова, которые используются в общей речи и не несут смысловой нагрузки, например артикли. И слова, которые встречаются один раз, так как они не будут влиять на результат вычисления весов.
- Вычислить матрицу "конференции-слова", где на пересечении слова и конференции стоит TF-IDF слова. TF-IDF — отношение количества употреблений слова в заголовках публикаций данной конференции к количеству употреблений в заголовках всех публикаций.
- Вычислить матрицу "конференции-конференции", аналогично коллаборативному алгоритму, используя косинус между векторами.
- Вычислить рейтинг для пользователя по формуле

$$r_{cu} = \frac{\sum_{d \in C_u} w_{cd}}{|C_u|}$$

r_{cu} — рейтинг конференции c для пользователя u. C_u — множество конференций, на которых был пользователь u. w_{cd} — вес между

[2] https://en.wikipedia.org/wiki/Stemming

конференциями c и d. $|C_u|$ - количество конференций, на которых был пользователь u.

5.2. Реализация

При реализации также возникли проблемы с памятью и скоростью вычислений. Но в данном случае матрица “конференции-слова” вышла не такой разряженной, как при коллаборативном варианте и поэтому вариант с разреженной матрицей не подошел. Вместо этого была использована библиотека pytables[9] , которая позволяет хранить numpy массивы на диске, а не в оперативной памяти. При этом операция чтения выполняется дольше, чем операция скалярного умножения векторов, поэтому для того что бы уменьшить количество обращений к диску, матрица считывается блоками 500 * 500.

6. Гибридный подход

Гибридный подход был реализован как взвешенная сумма коллаборативного и контентно-ориентированного алгоритма.

Формула:

$r_h = \alpha * r_{cf} + (1 - \alpha) * r_{cb}$, где r_h обозначет рейтинг гибридной системы, r_{cf} — коллаборативной, r_{cb} — контентно-ориентированной. a - выбирается по результатом тестирования, которое описано в следующем разделе.

7. Измерение результатов

7.1. Описание методов

Измерения качества алгоритма проводилось с использованием кросс-валидации. Это техника оценки качества предсказательного алгоритма, при котором данные разбиваются на k частей, на k–1 части производится обучение, а на одной — проверка. . В данной работе данные разбиваются на две части, соответственно половина используется как тренировочная, половина — как тестовая выборка.

В качестве метрик используются precision и recall[3]. Precision, в данном случае, — отношение верно порекомендованных конференций ко всем рекомендованным. Recall — отношение верно порекомендованных ко всем верным.

7.2. Распределение данных

Пользователи по количеству конференций распределены следующим образом:

[3] https://en.wikipedia.org/wiki/Precision_and_recall

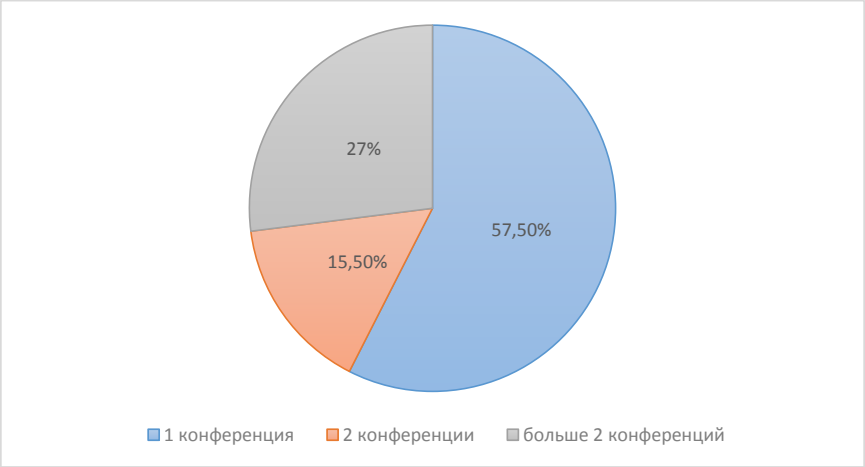

Рис. 1: Распределение пользователей по количеству конференций

Для тех, у кого одна конференция, провести кросс-валидацию по конференциям не возможно. Для пользователей с двумя конференциями был посчитан precision при recall, равном одному. Это значение показывает, сколько в среднем надо порекомендовать конференций, для того чтобы выдать релевантную. И для последней группы была посчитана precision-recall кривая.

7.3. Пользователи с двумя конференциями

Была выбрана группа из 1500 пользователей, на которых была проведена кросс-валидация и посчитано значение precision при recall, равном единице.

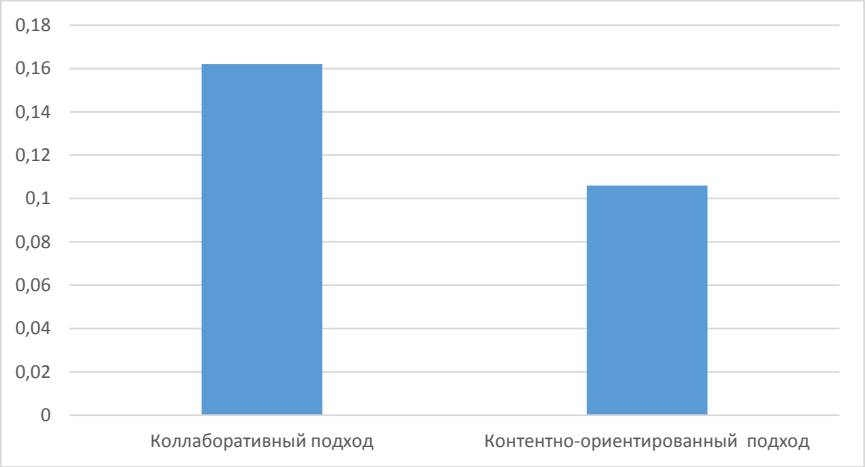

Рис. 2: Значение precision

7.4. Пользователи, у которых больше двух конференций

В данном случае также были выбраны 1500 пользователей, но вместо вычисления precision при recall, равном единице, была построена precision-recall кривая.

Сначала будет рассмотрен пример вычисления precision-recall кривой для рекомендации одному пользователю, а после этого уже усредненное значение для 1500.

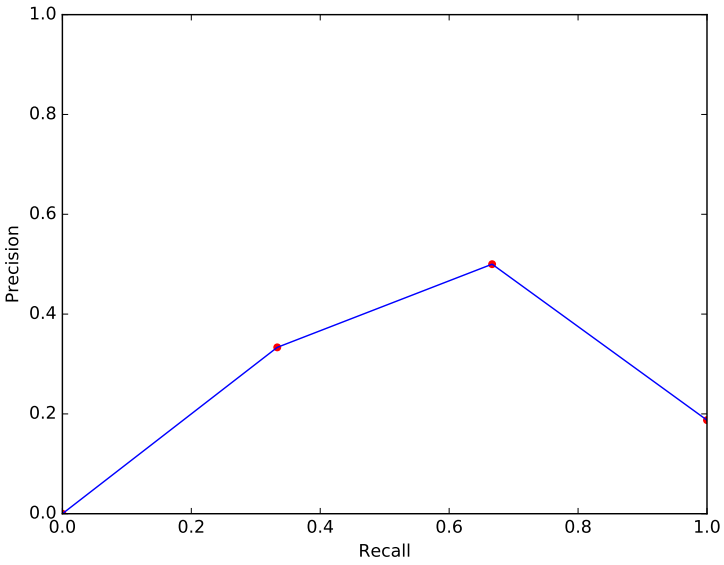

Рис. 3: Precision-recall кривая

Рекомендации	Веса	№	Точки
STACS	0.4415		
ESA	0.4384		
ISTCS	0.4007		
SODA	0.2369	1	(0, 0)
ICALP	0.2243	2	
FOCS	0.2207	3	(1/3, 1/3)
STOC	0.2084	4	(2/3, 1/2)
ICALP (1)	0.1999	5	
SWAT	0.1967	6	
ISAAC	0.1871	7	
MFCS	0.1808	8	
LATIN	0.1655	9	
WAOA	0.1538	10	
WADS	0.1512	11	
FSTTCS	0.1457	12	
WG	0.1452	13	
APPROX-RANDOM	0.1452	14	
COCOON	0.1401	15	
SPAA	0.1333	16	(1, 3/16)

Рис. 4: Пример выдачи рекомендаций

На Рис. 4 представлена рекомендация пользователю, у которого

шесть конференций: STACS, ESA, ISTCS, FOCS, STOC, SPAA. С помощью кросс-валидации данные конференции были разбиты на две части, половина пошла в тренировочную, половина — в тестовую выборку.Зеленым обозначены тренировочные, а желтым — тестовые данные. Рис. 3 показывает precision-recall кривую для данной рекомендации. Первые три конференции из тренировочной выборки, и они не участвуют в расчете кривой. Рассмотрим четыре точки на кривой. Первая — $(0, 0)$, она говорит о том, что в начале рекомендуется не релевантная конференция. Вторая — $(\frac{1}{3}, \frac{1}{3})$ — первая релевантная конференция находится на третьем месте, соответственно precision и recall равны $\frac{1}{3}$. На четвертом месте снова идет релевантная конференция, precision равен $\frac{1}{2}$, recall — $\frac{2}{3}$. Последняя рекомендуется на шестнадцатом месте, соответственно precision равен $\frac{3}{16}$, recall равен 1.

Количественной мерой, показывающей качество алгоритма является площадь под графиком (AU-PR). У идеального алгоритма она равна единице. В таком случае precision равен одному при любом recall и первыми рекомендуются релевантные конференции.

Теперь будет рассмотрен пример усредненного графика precision-recall кривой для 1500 пользователей.

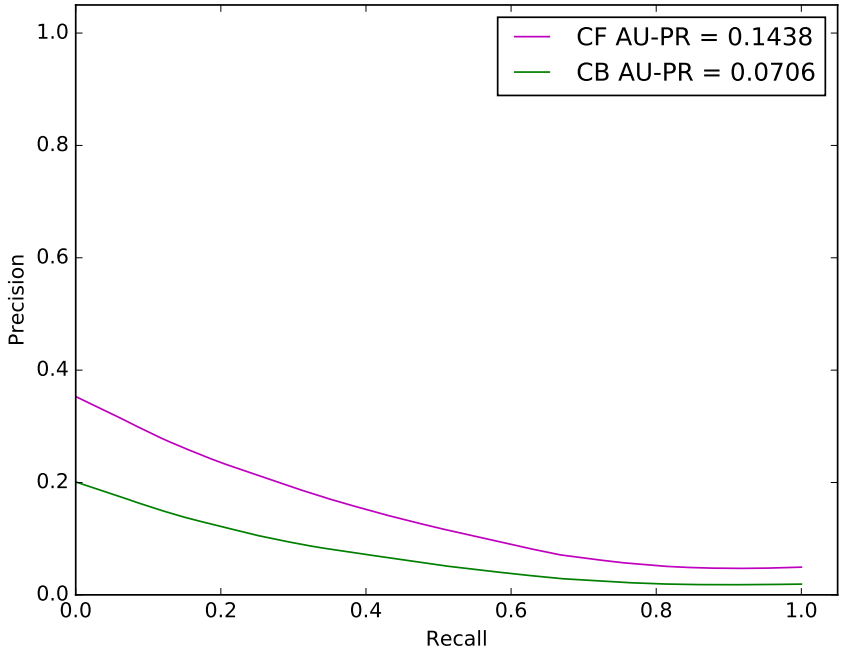

Рис. 5: Усредненная precision-recall кривая

На Рис. 5 фиолетовым обозначен коллаборативный, а зеленым — контентно-ориентированный алгоритм. Как видно на графике, площадь под коллаборативным алгоритмом примерно в 2 раза больше. Значит в среднем он работает лучше.

7.5. Гибридный алгоритм

Как уже было описано, гибридный алгоритм был реализован с использованием взвешенной суммы. Тестирование состоит в подборе веса α в формуле $r_h = \alpha * r_{cf} + (1 - \alpha) * r_{cb}$. Для тестирования пользователи были разбиты на две группы — на тех, у кого от двух до шести конференций и на тех, у кого больше шести.

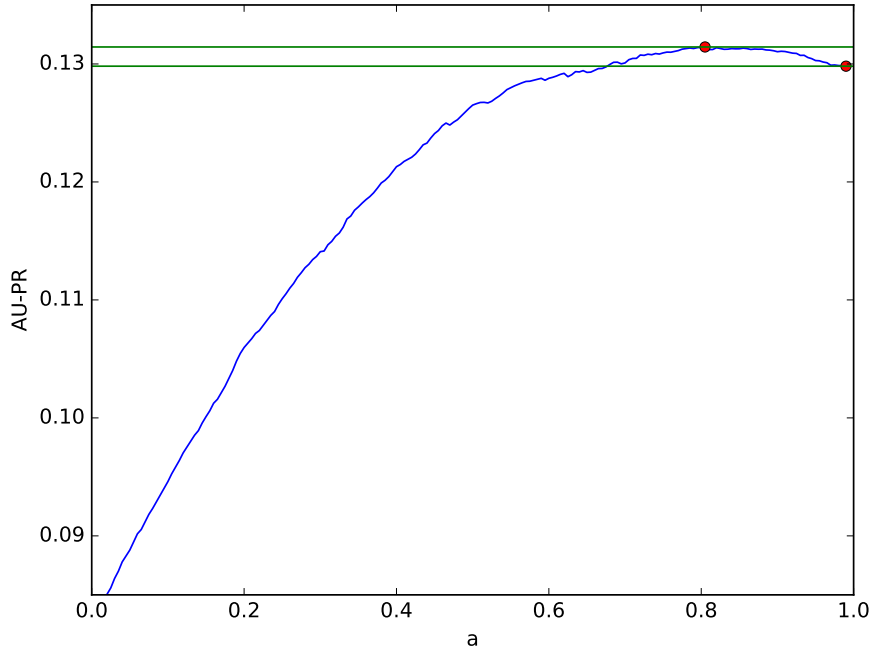

Рис. 6: Значение AU-PR относительно весового коэффициента для пользователей, у которых от двух до шести конференций. Верхняя зеленая линия обозначает максимальное значение AU-PR. Нижняя — значение AU-PR при коллаборативном подходе.

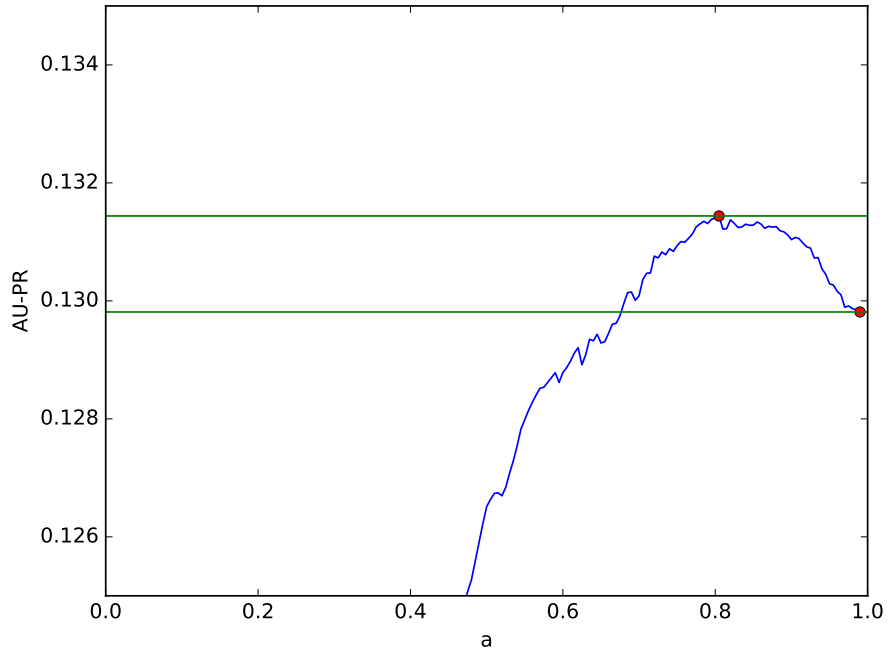

Рис. 7: Значение AU-PR относительно весового коэффициента для пользователей, у которых от двух до шести конференций в увеличенном масштабе.

Как видно на Рис. 6 и Рис.7 гибридная модель дала прирост относительно коллаборативной для данной группы пользователей. Максимальное значение AU-PR для гибридной модели равно 0.1314 (при α равном 0.805), а для коллаборативной 0.1298. Таким образом прирост составил 1.25% по данной метрике.

Далее таким же образом рассматривается отношение AU-PR и веса α для группы пользователей, у которых больше шести конференций.

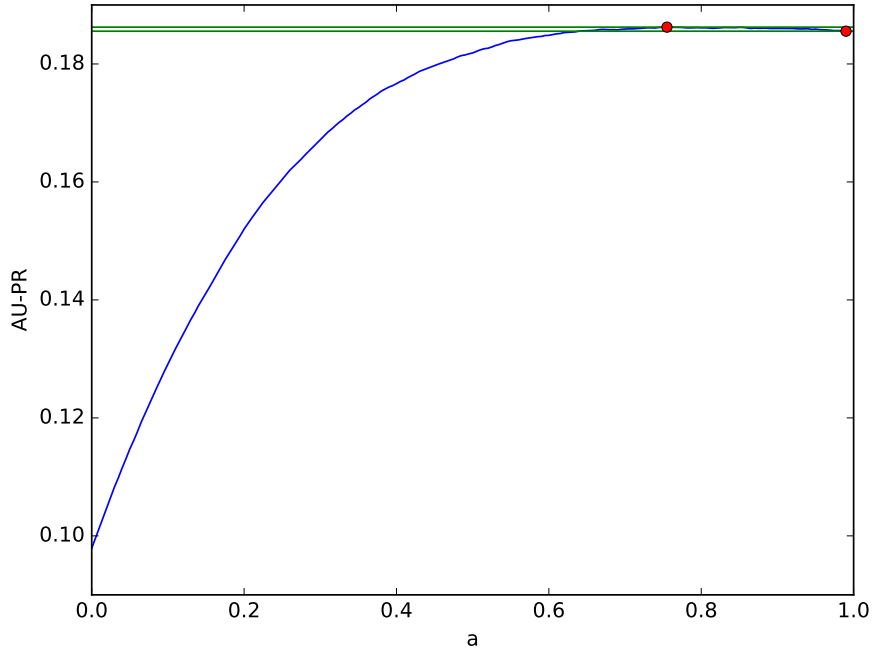

Рис. 8: Значение AU-PR относительно весового коэффициента для пользователей, у которых больше шести конференций. Верхняя зеленая линия обозначает максимальное значение AU-PR. Нижняя — значение AU-PR при коллаборативном подходе.

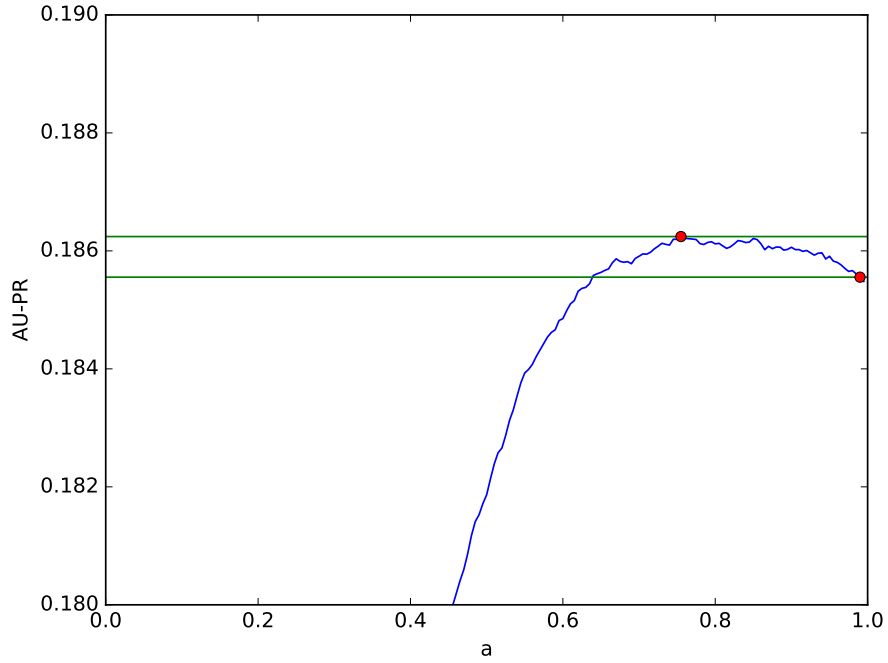

Рис. 9: Значение AU-PR относительно весового коэффициента для пользователей, у которых больше шести конференций в увеличенном масштабе.

Как видно на Рис.8 и Рис.9 для данной группы пользователей гибридная модель также дала прирост, но он оказался меньшим. Возможно это потому, что для пользователей у которых много конференций коллаборативная модель работает лучше, чем для пользователей, у которых количество конференций невелико. Максимальное значение AU-PR для гибридной модели равно 0.1862 (при α равном 0.755). Для коллаборативной модели значение AU-PR равно 0,1855. Таким образом гибридная модель дала прирост 0.37%.

Заключение

- Была реализован гибридный рекомендательный алгоритм, состоящий из коллаборативного и контентно-ориентированного подхода.
- Было проведено тестирование с использование метрик precision и recall. По результатам тестирования гибридная модель превзошла коллаборативную. При этом лучший прирост она показала для пользователей с небольшим числом конференций.
- Итоговая формула для гибридной системы : $r_h = 0.8 * r_{cf} + 0.2 * r_{cb}$
- Код работы доступен по адресу : [8]

Список литературы

[1] Combining content-based and collaborative filters in an online newspaper / Mark Claypool, Anuja Gokhale, Tim Miranda et al. // Proceedings of ACM SIGIR workshop on recommender systems / Citeseer. — Vol. 60. — 1999.

[2] DBLP. — 2016. — May. — URL: http://dblp.uni-trier.de/.

[3] Ho Yvonne, Fong Simon, Yan Zhuang. A Hybrid GA-based Collaborative Filtering Model for Online Recommenders // ICE-B 2007 - Proceedings of the International Conference on e-Business, Barcelona, Spain, July 28-31, 2007, ICE-B is part of ICETE - The International Joint Conference on e-Business and Telecommunications. — 2007. — P. 200–203.

[4] Lang Ken. NewsWeeder: Learning to Filter Netnews // Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, USA, July 9-12, 1995. — 1995. — P. 331–339.

[5] Linden Greg, Smith Brent, York Jeremy. Amazon.com Recommendations: Item-to-Item Collaborative Filtering // IEEE Internet Computing. — 2003. — Vol. 7, no. 1. — P. 76–80. — URL: http://dx.doi.org/10.1109/MIC.2003.1167344.

[6] Lops Pasquale, de Gemmis Marco, Semeraro Giovanni. Content-based Recommender Systems: State of the Art and Trends // Recommender Systems Handbook. — 2011. — P. 73–105. — URL: http://dx.doi.org/10.1007/978-0-387-85820-3_3.

[7] Mooney Raymond J., Roy Loriene. Content-based book recommending using learning for text categorization // ACM DL. — 2000. — P. 195–204. — URL: http://doi.acm.org/10.1145/336597.336662.

[8] Pesults. — 2016. — May. — URL: https://github.com/Ershov-Alexander/rec_sys_for_sci_conf.

[9] Pytables. — 2016. — May. — URL: http://www.pytables.org/.

[10] Roh Tae Hyup, Oh Kyong Joo, Han Ingoo. The collaborative filtering recommendation based on SOM cluster-indexing CBR // Expert Syst. Appl. — 2003. — Vol. 25, no. 3. — P. 413–423. — URL: http://dx.doi.org/10.1016/S0957-4174(03)00067-8.

[11] Science Concierge: A fast content-based recommendation system for scientific publications / Titipat Achakulvisut, Daniel E Acuna, Tulakan Ruangrong, Konrad Kording // arXiv preprint arXiv:1604.01070. — 2016.

[12] Scikit-learn. — 2016. — May. — URL: http://scikit-learn.org/stable/.

[13] Scipy. — 2016. — May. — URL: https://www.scipy.org/.

[14] Scipy sparse matrix. — 2016. — May. — URL: http://docs.scipy.org/doc/scipy/reference/sparse.html.

[15] Shani Guy, Gunawardana Asela. Evaluating Recommendation Systems // Recommender Systems Handbook. — 2011. — P. 257–297. — URL: http://dx.doi.org/10.1007/978-0-387-85820-3_8.

[16] Smyth Barry, Cotter Paul. A personalized television listings service // Communications of the ACM. — 2000. — Vol. 43, no. 8. — P. 107–111.

[17] Ungar Lyle H, Foster Dean P. Clustering methods for collaborative filtering // AAAI workshop on recommendation systems. — Vol. 1. — 1998. — P. 114–129.

[18] Vozalis Manolis G., Margaritis Konstantinos G. Using SVD and demographic data for the enhancement of generalized Collaborative Filtering // Inf. Sci. — 2007. — Vol. 177, no. 15. — P. 3017–3037. — URL: http://dx.doi.org/10.1016/j.ins.2007.02.036.

[19] numpy. — 2016. — May. — URL: http://www.numpy.org/.