УДК 004.891.3 **doi: 10.26906/SUNZ.2018.3.073**

В. А. Лещинский, И. А. Лещинская

Харьковский национальный университет радиоэлектроники, Харьков, Украина

УСОВЕРШЕНСТВОВАНИЕ МЕТОДА КОЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ С НЕЯВНОЙ ОБРАТНОЙ СВЯЗЬЮ НА ОСНОВЕ РАНЖИРОВАНИЯ ОТРИЦАТЕЛЬНЫХ РЕЗУЛЬТАТОВ В МАТРИЦЕ ИСХОДНЫХ ДАННЫХ

Предметом изучения в статье являются процессы выявления предпочтений пользователей в рекомендательных системах. **Целью** является разработка усовершенствованного метода колаборативной фильтрации на основе ранжирования пропущенных и отрицательных результатов в матрице исходных данных. **Задачи:** Формализовать свойства исходных данных, включая пропущенные данные для задачи коллаборативной фильтрации; разработать подход к ранжированию исходных данных для коллаборативной фильтрации с неявной обратной связью, включая пропущенные и отрицательные результаты; усовершенствовать метод колаборативной фильтрации путем предварительного ранжирования пропущенных и отрицательных результатов в исходных данных. Используемыми **методами** являются**:** коллаборативная фильтрация, машинное обучение. Получены следующие **результаты.** Формализованы свойства исходных данных для коллаборативной фильтрации. Такие данные упорядочиваются для каждого пользователя как последовательность предпочтений интересующих пользователя объектов. На основе свойств исходных данных показано, что при коллаборативной фильтрации с неявной обратно связью необходимо упорядочивать не только данные о поведении пользователя, но и пропущенные и неточные данные. Предложен подход к упорядочиванию таких данных на основе их попарного сравнения. Усовершенствован метод коллаборативной фильтрации на основе уточнения весовых коэффициентов для обучающей выборки с учетом предварительного ранжирования входных данных. **Выводы.** Научная новизна полученных результатов состоит в следующем: усовершенствован метод коллаборативной фильтрации с неявной обратной связью путем присвоения дополнительных весов элементам в матрице исходных данных на основе ранжирования пропущенных и отрицательных результатов. Метод позволяет повысить точность рекомендаций по критерию AUC с учетом неполноты исходных данных.

Ключевые слова: коллаборативная фильтрация; машинное обучение; ранжирование; факторизация матриц; кривая ошибок; обучающая и тестовая выборки.

Введение

Коллаборативная фильтрация используется в рекомендательных системах. Рекомендательные системы традиционно используются для прогнозирования выбора интересных пользователю объектов на основе имеющейся информации, как о текущем пользователе, так и о других клиентах. В качестве объектов, представляющих интерес для пользователя, рассматриваются: новости, веб-сайты, фильмы, товары и т.п. [1-3].

В зависимости от типа входных данных рекомендательные системы применяют два вида обратной связи: явную (explicit feedback) и неявную (implicit feedback) [4].

Explicit feedback основана на явных отзывах и оценках пользователей, отражающих их интерес к книге, фильму или иному объекту.

Такую обратную связь использует, в частности, компания Netflix при формировании рейтинга для фильмов и телевизионных программ [5]. Пользователи имеют возможность выставлять оценки в данном сервисе.

Однако в большинстве случаев получение явных оценок пользователей связано с трудностями, что и приводит к необходимости использования неявной обратной связи от пользователя. Implicit feedback отражает предпочтения пользователя на основе наблюдений за его поведением [6].

В качестве входных данных при неявной обратной связи обычно используют: историю покупок; просмотренных сайтов; шаблоны поиска; шаблоны движения мыши, и т.п.

Ключевым подходом, который используется при построении рекомендательных систем с неявной обратной связью, является коллаборативная фильтрация (collaborative filtering) [2-7].

Данный подход основывается на принципах отбора по схожести пользователей либо по схожести объектов, с которыми взаимодействуют пользователи [8].

При коллаборативной фильтрации с неявной обратной связью выявляют скрытые факторы, которые влияют на выбор пользователя.

При реализации коллаборативной фильтрации возникают проблемы, связанные с разреженностью данных, а также наличием релевантных данных.

Разреженность данных связана с тем, что в большинство коммерческих рекомендательных систем основано на большом количестве данных о товарах и незначительном количестве оценок пользователей.

В результаты матрица «объект – пользователь» получается очень большой с малым количеством данных о покупках/просмотрах. Это не позволяет повысить точность рекомендаций. Указанная проблема особенно актуальна для новых, только что созданных систем [2].

Проблема релевантности оценок часто возникает в случае холодного старта, поскольку новые объекты или пользователи затрудняют создание релевантных рекомендаций [3].

Для решения указанных проблем целесообразно использовать не только разреженные «положительные» данные – информацию о количестве покупок, просмотров, но и данных об «отрицательных» предпочтениях.

Однако существующие версии коллаборативной фильтрации ориентированы на положительные данные. Это не дает возможность получить генерализованную модель и может приводить к возникновению проблемы переобучения.

Изложенное свидетельствует об актуальности проблематики данной статьи.

Целью статьи является усовершенствование метода коллаборативной фильтрации с неявной обратной связью на основе дополнительного ранжирования «отрицательных» исходных данных с тем, чтобы снизить количество ошибок распознавания.

Ранжирование отрицательных результатов в матрице исходных данных при построении рекомендаций с неявной обратной связью

В качестве исходных данных при построении рекомендаций используются:

– список пользователей

$$U = \{u_i\};$$

– список объектов, интересующих пользователей

$$E = \{e_j\};$$

– матрица рейтингов/покупок

$$R \subseteq U \times E;$$

пользователи взаимодействуют с объектами (товарами), формируя матрицу рейтингов в рекомендационной системе:

$$R = \{r_{ij}\},$$

где r_{ij} – рейтинг j – предмета e_j у пользователя u_i.

Очевидно, что каждый из пользователей ставит рейтинг (в случае явной обратной связи), покупает (в случае неявной связи) только незначительное подмножество товаров из общего списка. Поэтому матрица рейтингов обычно сильно разрежена, большинство ее элементов равны «0». Иными словами, только небольшая часть из возможных взаимодействий «пользователь-товар» будет присутствовать в наборе данных.

Результатами построения рекомендаций являются либо упорядоченный список объектов E_U для каждого пользователя либо упорядоченный список пользователей U_E, соответствующий каждому объекту.

Для обоих списков должно выполняться условие: объекты и пользователи присутствуют в матрице R:

$$E_U = \{e_i \mid \exists r_{ij} \in R\}; \quad (1)$$

$$U_E = \{u_j \mid \exists r_{ij} \in R\}. \quad (2)$$

Формируемый рекомендационной системой порядок пользователей либо объектов обладает свойствами всеобщности (3), транзитивности (4), антисимметричности (5):

$$\begin{gathered}\forall (e_i \neq e_k) \in E_U \exists \underset{u_j}{(>)} : e_i \underset{u_j}{>} e_k \vee e_i \underset{u_j}{<} e_k, \\ \forall (u_j \neq u_l) \in U_E \exists \underset{e_i}{(>)} : u_j \underset{u_j}{>} u_l \vee u_j \underset{u_j}{<} u_l,\end{gathered} \quad (3)$$

$$\begin{gathered}\forall (e_i \neq e_k \neq e_n) \in E_U : \\ e_i \underset{u_j}{>} e_k \wedge e_k \underset{u_j}{>} e_n \Rightarrow e_i \underset{u_j}{>} e_n, \\ \forall (u_j \neq u_l \neq u_m) \in U_E : \\ u_j \underset{u_j}{>} u_l \wedge u_j \underset{u_j}{>} u_l \wedge u_j \underset{u_j}{>} u_m,\end{gathered} \quad (4)$$

$$\begin{gathered}\forall (e_i = e_k) \in E_U : e_i \underset{u_j}{>} e_k \wedge e_i \underset{u_j}{<} e_k, \\ \forall (u_j = u_l) \in U_E : u_j \underset{u_j}{>} u_l \wedge u_j \underset{u_j}{<} u_l,\end{gathered} \quad (5)$$

Важное отличие систем с неявной обратной связью состоит в том, что вместо положительных и отрицательных рейтингов они используют только позитивные данные (например, количество покупок).

Остальные данные отражают как объекты, не представляющие интереса для пользователя, так и пропущенную информацию о покупках, просмотрах и т.п.

В процессе обработки матрицы R в рекомендационных системах обычно выполняется бинаризация данных. При этом позитивные данные представляются единицей, а остальная информация – нулем.

В дальнейшем в процессе машинного обучения представленные единицей данные рассматриваются как принадлежащие к классу, а все остальные, включая пропущенные - как лежащие за пределами класса.

Построенная в результате обучения модель должна предсказывать единицы в матрице исходных данных R.

Следовательно, пропущенная информация исключается из дальнейшего рассмотрения.

Однако полученная модель должна упорядочивать все объекты для каждого пользователя согласно описанных выше свойств (1) – (5), что требует ранжирования отрицательных результатов в матрице исходных данных.

Для выполнения ранжирования целесообразно выполнить попарное сравнение объектов e_j для каждого пользователя u_i, которые отражены в матрице R.

Предлагаемый подход к решению этой задачи включает в себя такие шаги:

1. Определение относительного порядка между объектами для каждого пользователя. На данном шаге выполняется попарное сравнение результатов (покупок, просмотров) для каждого пользователя.

Для пользователя формируется квадратная матрица, элементы которой указывают, какой из объектов является более интересным для него. Знак + в этой матрице означает, что элемент в строке более интересен, чем элемент в столбце.

Пример реализации данного шага для пользователя u_1 приведен на рис. 1.

	e_1	e_2	e_3	e_4
u_1	0	9	6	0

⇩

	e_1	e_2	e_3	e_4
e_1	=	9	6	0
e_2	+	=	+	+
e_3	+	-	=	+
e_4	?	-	-	=

Рис. 1. Определение относительного порядка между парами объектов для каждого пользователя

Так, знак + в ячейке (2,1) означает, что для пользователя u_1 элемент e_2 предпочтительнее элемента e_1. Действительно, пользователь выбрал элемент e_2 9 раз, а элемент e_1 - 0 раз.

Знак вопроса в ячейке квадратной матрицы означает, что предпочтения установить не удалось. Например, пользователь не выбирал оба элемента e_1 и e_4 поэтому в ячейке (4,1) стоит вопросительный знак.

2. Определение весов для объектов. На данном шаге на основе отношений между объектами определяются веса, отражающие возможный интерес пользователя к этим объектам.

Основная идея шага заключается в следующем. Нам необходимо в результате обучения построить достаточно обобщенную модель, которая отражала бы общие тенденции и предпочтения пользователей и позволяла бы избежать отдельных флуктуаций, отражающих специфику конкретного набора данных.

Традиционно такая задача решается путем регуляризации. Коэффициент регуляризации подбирается таким образом, чтобы сгладить флуктуации отдельных переменных относительно требуемой закономерности.

В нашем случае причинами пропуска объектов пользователем (например, e_1 и e_4) могут быть:

– отсутствие интереса;

– отсутствие доступа;

– ошибка в записях.

При традиционном подходе все эти причины объединяются и рассматриваются как отсутствие интереса пользователя к объекту.

Это может привести флуктуациям и соответствующему искажению общих закономерностей в полученной модели. Для исключения такой зависимости от конкретных данных предлагается добавить небольшие значения пропущенным элементам в соответствии с установленной на шаге 1 упорядоченностью.

Для представленного на рис.1 упрощенного примера получаем такую упорядоченность элементов: $e_2(9)$, $e_3(6)$, e_1 и $e_4(?)$.

Линейная экстраполяция полученной последовательности элементов e_i очевидно приводит к таким их значениям: 9; 6; 3; 3. Для исключения флуктуаций используем коэффициент регуляризации, например $\lambda = 0{,}1$. Тогда упорядоченная последовательность значений элементов принимает такой вид: 9; 6; 0,3; 0,3. Последние 2 значения в последовательности элементов отражают потенциальный интерес пользователя к элементам e_1 и e_4 как результат отсутствия доступа или ошибки в исходных данных.

Очевидно, что представленная в примере линейная интерполяция является частным случаем общей закономерности, связывающей веса элементов $e_1 - e_4$.

В более общем случае требуется построение регрессии методами машинного обучения для того, чтобы рассчитать значения пропущенных элементов.

Значение коэффициента λ подбирается экспериментально.

Для этого в рамках коллаборативной фильтрации необходимо провести нескольких циклов машинного обучения с разными значениям λ.

Сравнение полученных моделей выполняется по метрике AUC, как будет показано в описании метода.

Коллаборативная фильтрация с неявной обратной связью с уточнением пропущенных данных

Усовершенствованный метод коллаборативной фильтрации с неявной обратной связью включает в себя следующие этапы:

– подготовка отфильтрованной матрицы исходных данных R^*;

– формирование обучающей и тестовой выборок из матрицы R^*;

– выявление неявных связей между клиентами и товарами на основе матричной факторизации;

– оценка полученных рекомендаций.

На этапе подготовки матрицы R^* выполняются такие шаги: фильтрация данных; проверка пригодности исходных данных; построение матрицы R «пользователи-объекты» из исходной базы данных; построение матрицы приоритетов.

На шаге фильтрации данных из исходных записей удаляются ошибочные строки, т.е. такие, в которых отсутствует информация о пользователе либо об объекте его интереса (товаре).

На шаге проверки пригодности исходных данных проводится проверка разреженности матрицы. На практике в качестве порога разреженности обычно рассматривают значение 99,5%. Если разреженность не превышает этот порог, то исходные данные считаются пригодными для формирования рекомендаций.

На шаге построения матрицы «пользователи-объекты» реализуются инженерные решения, позволяющие преобразовать имеющиеся записи о покупках (просмотрах) в стандартный формат исходных данных.

На шаге построения матрицы весов приоритеты c_{ij} для «положительных» данных формируются путем нормирования количества покупок для элементов r_{ij}.

Последовательность формирования приоритетов для «отрицательных» данных была приведена выше.

На этапе формирования обучающей и тестовой выборок выполняется обработка исходной матрицы R^*.

Обе выборки создаются из этой матрицы. На данном этапе выполняются такие шаги: формируется обучающая выборка путем маскирования части данных; формируется тестовая выборка путем бинаризации исходной матрицы.

В качестве обучающей выборки используется подмножество элементов матрицы R^*. Основная задача данного шага заключается в том, чтобы отразить характер взаимодействия максимального количества пользователей с большинством объектов. Поэтому около 30% элементов данной матрицы удаляется случайным образом.

Тестовая выборка формируется путем замены количества покупок/просмотров значением 1.

Задача факторизации методами машинного обучения [9, 10] заключается в минимизации взвешенного квадрата отклонений для всех элементов исходной матрицы R:

$$\min \sum_{i,j} c_{ij} (r_{ij} - x_i^T y_j)^2 ,$$

где x_i^T – транспонированная i – строка матрицы «пользователь - латентная переменная»;

y_j – строка j матрицы «латентная переменная – объект».

На этапе факторизации матриц предлагается использовать метод ALS (Alternating Least Squares Method).

Особенность данного метода состоит в поочередном нахождении минимума:

то для элемента «пользователь – переменная» при фиксации элемента «переменная – объект»,

то для элемента «переменная – объект» при фиксированных элементах «пользователь – переменная».

На этапе оценки полученных рекомендаций используется кривая ошибок ROC (Receiver Operating Characteristic) и соответствующая метрика AUC . Данная кривая отражает соотношение между TPR и FPR для всех элементов выборки.

Чем ближе кривая к вертикальной оси, тем меньше ошибка классификации. При этом увеличивается площадь под кривой. Поэтому площадь под кривой ошибок AUC (Area Under the Curve) используется для оценки результатов прогнозирования. AUC можно рассматривать как вероятность того, что случайно выбранные корректные рекомендации окажутся выше в списке случайно выбранных плохих подсказок.

Выполненная экспериментальная проверка усовершенствованного метода показала, что для выборки объемом 250 тыс. записей, отражающей покупки в сети интернет-магазинов, AUC после ранжирования отрицательных результатов увеличилась с 0,862 до 0,897.

Выводы

В статье усовершенствован метод коллаборативной фильтрации с неявной обратной связью путем присвоения весов элементам в матрице исходных данных с учетом ранжирования отрицательных результатов.

Ранжирование основано на попарном сравнении предпочтений каждого пользователя по отношению к объектам в матрице исходных данных.

По результатам попарного сравнения формируется относительный порядок элементов для каждого пользователя, что позволяет присвоить веса элементам матрицы исходных данных.

В практическом плане метод позволяет повысить точность рекомендаций с учетом неполных исходных данных.

REFERENCES

1. Adomavicius G. and Tuzhilin A. (2005), “Towards the Next Generation of Recommender Systems” *A Survey of the State-of-the-Art and Possible Extensions, IEEE Transactions on Knowledge and Data Engineering*, No. 17, pp. 634–749.
2. Ekstrand, M.D., Riedl J.T., and Konstan J.A. (2011), “Collaborative filtering recommender systems”, *Foundations and Trends in Human-Computer Interaction* No. 4(2) , pp. 81–173.
3. Herlocker J.L., Konstan J.A., Terveen L.G. and Riedl J.T. (2004), “Evaluating collaborative filtering recommender systems” *ACM Transactions on Information Systems (TOIS) 22, 1*, pp. 5–53.
4. Guo G., Zhang J. and Yorke-Smith N. (2015), “TrustSVD: Collaborative Filtering with Both the Explicit and Implicit Influence of User Trust and of Item Ratings” *AAAI*, pp. 123–129.

5. Bennet J. and Lanning S. (2007) "The Netflix Prize", *Proceedings of KDD cup and workshop*, available at : http://www.netflixprize.com. (last accessed March 22, 2018).
6. Hu Y., Koren Y. and Volinsky C. (2008), "Collaborative filtering for implicit feedback datasets", *Data Mining*, *ICDM'08. Eighth IEEE International Conference on. IEEE*, pp. 263–272.
7. Pan R., Zhou Y., Cao B., Liu N. N., Lukose R. M., Scholz M., and Yang Q. (2008), 'One-class collaborative filtering", *IEEE International Conference on Data Mining*, pp. 502-511.
8. Linden G., Smith B. and York J. (2003), "Amazon.com recommendations: Item-to-item collaborative filtering", *Internet Computing, IEEE 7, 1,* pp. 76–80.
9. Yi Fang and Luo Si. (2011), "Matrix co-factorization for recommendation with rich side information and implicit feedback" *Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems. ACM*, pp. 65–69.
10. Yehuda Koren, Robert Bell, and Chris Volinsky. (2009), "Matrix factorization techniques for recommender systems", *Computer No.8*, pp. 30–37.

Рецензент: д-р техн. наук, проф. О. О. Можаєв,
Національний технічний університет "Харківський політехнічний інститут", Харків
Received (Надійшла) 14.03.2018
Accepted for publication (Прийнята до друку) 15.05.2018

Вдосконалення методу колаборативної фільтрації з неявним зворотнім зв'язком на основі ранжування негативних результатів в матриці вхідних даних

В. О. Лещинський, І. О. Лещинська

Предметом вивчення в статті є процеси виявлення вподобань користувачів в рекомендаційних системах. **Метою** є розробка вдосконаленого методу колаборативної фільтрації на основі ранжирування пропущених і негативних результатів в матриці вихідних даних. **Завдання**: Формалізувати властивості вихідних даних, включно з пропущеними даними для задачі колаборативної фільтрації; розробити підхід до ранжирування вихідних даних для колаборативної фільтрації з неявним зворотним зв'язком, включно з пропущеними і негативним результатами; удосконалити метод колаборативної фільтрації шляхом попереднього ранжування пропущених і негативних результатів у вихідних даних. Використовуваними **методами** є: колаборативна фільтрація, машинне навчання. Отримані наступні **результати**. Формалізовані властивості вихідних даних. Такі дані упорядковуються для кожного користувача як послідовність переваг цікавлять користувача об'єктів. На основі властивостей вихідних даних показано, що при колаборативній фільтрації з неявним зворотнім зв'язком необхідно упорядковувати не тільки дані про поведінку користувача, але і пропущені та неточні дані. Запропоновано підхід до впорядкування таких даних на основі їх попарного порівняння. Удосконалено метод колаборативної фільтрації на основі уточнення вагових коефіцієнтів для навчальної вибірки з урахуванням попереднього ранжирування вхідних даних. **Висновки**. Наукова новизна отриманих результатів полягає в наступному: удосконалено метод колаборативної фільтрації з неявним зворотнім зв'язком шляхом присвоєння додаткових ваг елементам в матриці вхідних даних на основі ранжирування пропущених і негативних результатів. Метод дозволяє підвищити точність рекомендацій за критерієм AUC з урахуванням неповноти вихідних даних.

Ключові слова: колаборативна фільтрація; машинне навчання; ранжування; факторизація матриць; крива помилок; навчальна і тестова вибірки.

Collaborative Filtering for Implicit Feedback Datasets using Ranking Negative Data

V. Leshchynskyi, I. Leshchynska

The **subject matter** of the article is the processes of revealing the preferences of users in recommender systems. The **goal** is to develop an improved method of collaborative filtering based on the ranking of missed and negative results in the matrix of input data. **Tasks**: Formalize the properties of the source data, including missing data for the collaborative filtering task; develop an approach to the ranking of input data for collaborative filtering with implicit feedback, including missed and negative results; improve the method of collaborative filtering by pre-ranking the missed and negative results in the original data. The **methods** used are: collaborative filtering, machine learning. The following **results** were obtained. The properties of the initial data are formalized. Such data is arranged for each user as a sequence of preferences of objects of interest to the user. Based on the properties of the source data, it is shown that with collaborative filtering with implicit feedback, it is necessary to organize not only data about the user's behavior, but also missed and inaccurate data. An approach to the ordering of such data is proposed on the basis of their pairwise comparison. The method of collaborative filtering is improved on the basis of refinement of weight coefficients for the training sample, taking into account the preliminary ranking of the input data. **Conclusions**. The scientific novelty of the results obtained is as follows: the method of collaborative filtering with implicit feedback has been improved by assigning additional weights to the elements in the matrix of the initial data based on the ranking of the missed and negative results. The method allows to increase the accuracy of the AUC criteria taking into account the incompleteness of the initial data.

Keywords: collaborative filtering; machine learning; ranging; factorization of matrices; error curve; training and test samples.