ПОВЫШЕНИЕ КАЧЕСТВА РЕШЕНИЯ ЗАДАЧИ topN КОЛЛАБОРАТИВНЫМИ РЕКОМЕНДАТЕЛЬНЫМИ СИСТЕМАМИ

ACCURACY IMPROVEMENT OF TOPN TASK FOR COLLABORATIVE FILTERING RECOMMENDER SYSTEMS

D. Ponizovkin

Summary. The subject of the article is recommender systems. The main functionality of these systems is to help their users to do quick search for actual and useful information on the base of recommendations provided by a system. This task is difficult but actual in modern conditions, when a huge amount of information is available on the Internet and different devices. There are different techniques applicable in recommender systems in order to implement the target functionality. This work considers one of the most popular and successful techniques — collaborative filtration. This technique is filtration based on connections, and could be made either by users or system objects. We will consider filtration by objects. With the help of this technique the task of definition of a subset of objects, similar in characteristics to user preferences, with cardinality of N is solving. The task solution is qualitatively found if the defined by the system subset consists of N objects similar to the user. In order to implement filtration, recommender systems calculate function values named similarity measures. If the similarity measure value of items is greater than some threshold value, then there is similarity relation between these items and these items do not filtered. We propose sufficient condition of efficient solving of the topN task for collaborative systems. This condition is the transitivity property of the similarity relation. Accomplishment of the sufficient condition depends on a choice of the similarity measure and threshold value. We propose the method of the modeling of collaborative systems for which sufficient condition is accomplished.

Keywords: collaborative filtreing, recommender systems, topN task, similarity measure, transitivity of similarity relation

Понизовкин Денис Михайлович
Программист, IT-Aces,
Ярославская обл., г. Переславль-Залесский
denis.ponizovkin@gmail.com

Аннотация. В работе рассматриваются рекомендательные системы. Основная функциональность этих систем заключается в реализации помощи их пользователям производить быстрый поиск актуальной и нужной информации на основании предоставляемых системой рекомендаций. Данная задача является сложной и актуальной в современных условиях, когда доступно огромное число информации через интернет и различные устройства. Существуют различные техники, которые применяются в рекомендательных системах для реализации целевой функциональности. В работе рассмотрена одна из самых известных и успешных техник — коллаборативная фильтрация. Эта техника заключается в фильтрации на основании взаимосвязей, которая может быть произведена по пользователям или по объектам системы. В работе рассматривается фильтрация, производимая по объектам. С помощью такой техники решается задача определения подмножества объектов мощности N, близких по характеристикам к предпочтениям пользователя. Решение задачи качественно, если определенное системой подмножество состоит из N близких к пользователю объектов. Для того, чтобы провести фильтрацию, рекомендательные системы производят вычисление значений функций, называемых мерами близости. Если значение меры близости больше некоторого порогового значения, то тогда выполняется отношение близости и такие объекты не отфильтровываются системой. В статье приведено достаточное условие, при котором коллаборативные системы гарантируют достижение качественного решения,— свойство транзитивности отношения близости объектов. Выполнение свойства транзитивности зависит от того, какая функция используется в качестве меры близости, и ее пороговое значение для определения выполнения отношения близости. В статье предложен способ построения таких рекомендательных систем, которые при применении коллаборативной фильтраии по множеству объектов гарантируют выполнение достаточного условия качественного решения.

Ключевые слова: рекомендательная система, коллаборативная фильтрация, задача topN, качество решения, мера близости.

Введение

С интенсивным развитием веб-технологий, огромным и постоянно растущим числом информации, доступной через интернет с помощью множества различных устройств, популярными становятся рекомендательные системы (далее РС) [1]. РС облегчают процесс поиска пользователем интересующей его информации. Это осуществляется РС путем рекомендации пользователю подмножества объектов заданной мощности N, которые будут близки по свои характеристикам к предпочтениям пользователя. Задача определения такого подмножества называется задачей $topN$ [2,3].

Известными веб-сервисами, использующими РС, являются Netflix и Youtube — для просмотра видео, Google News и Yahoo! News — для просмотра новостей, LastFm и Spotify — для прослушивания музыкальных произведений, Amazon и Ebay — для приобретения товаров, и т.п.

Существуют различные рекомендательные техники, которые используются РС для решения задачи $topN$.

В работе рассматривается техника, именуемая коллаборативной фильтрацией [6,7], которая является одной из наиболее изученных [14], популярных [12] и успешных техник [13]. Коллаборативные РС делятся на два типа по фильтруемому множеству [4]: может фильтроваться либо множество пользователей, либо множество объектов. Будем называть последние объектно-ориентированными (далее ОРС) [9]. В основном, ОРС решают задачу $topN$, поэтому именно они и будут рассматриваться в статье [2,3].

Терминология и обозначения

Исходные данные РС — это $P = \{\rho(u,i):\rho(u,i): \neq\perp\}$, где:

- $u \in U \subset \mathbb{N}$ — идентификаторы пользователей РС.
- $i \in I \subset \mathbb{N}$ — идентификаторы объектов предметной области РС. Например, фильм кинематографической РС. Для простоты изложения не будем каждый раз говорить идентификатор пользователя или объекта, а

просто — пользователь или объект.

- $\rho(u,i)$ — значение расстояние между пользователем u и объектом i. Эти значения являются показателем близости объекта i к пользователю u по некоторым характеристикам, они могут значить, к примеру, степень того, насколько объект «нравится» пользователю. Будем говорить, что между пользователем u и объектом i выполняется отношение близости R, если $\rho(u,i)\leq\varepsilon_0$, где ε_0 — некоторое пороговое зачение. Будем называть таких пользователей и объектов близкими. Как правило, расстояния задаются самими пользователями за время работы с системой.
- $\rho(u,i)=\perp$, если расстояние неизвестно.

Во введенной терминологии задача $topN$ примет следующий вид: $I_{topN} = \{i:u_aRi)\wedge\rho(u_a,i) = \perp\}\wedge|I_{topN}| = N$, где U_a — активный пользователь, для которого в данный момент решается задача.

Решение задачи $topN$ РС производится за счет анализа информации о характеристиках объектов и исходных данных пользователя. Обозначим символом Y множество характеристик объектов, например, наименования кинематографических жанров. Значением характеристики объекта является значение весовой функции $v:I\times Y\rightarrow[0,1]$. Значения весов могут вычисляться экспертами, алгоритмически РС и т.д. Для излагаемого материала данный аспект не имеет значения. Структуру данных, представляющую информацию о характеристиках объекта i назовем контентом объекта и обозначим $C_Y(i)$.

Объектно-ориентированные коллаборативные РС

Решение задачи $topN$ в ОРС основано на эвристическом утверждении [8,9,10], которое гласит, что если пользователю нравится объект k, который близок по характеристикам к объекту l, то пользователю понравится объект l. Во введенной терминологии данное утверждение примет следующий вид:

$$(u_aRi) \wedge (i\underset{I}{R}j) \Rightarrow u_aRj \quad (1)$$

$\underset{I}{R}$ — отношение близости объектов.

$$i\underset{I}{R}j \Leftrightarrow (1 - \delta_i(i,j)) \leq \varepsilon_0,$$

где $\delta_i:I \times I \rightarrow [0,1]$ — мера близости объектов. Распространенная мера близости ОРС — это косинус угла между контентами, которые представляются в виде векторов [4]. Объекты, между которыми выполняется отношение близости, называются соседями.

Для решения задачи $topN$ в ОРС нужна информация только о тех объектах, для которых выполняется отношение u_aRi, поэтому будем считать, что $P = \{\rho(u,i)\leq\varepsilon_0\}$.

Решение задачи topN и его оценка

Для того, чтобы оценить применяемый метод решения, проводится тестирование, в ходе которого исходные данные разбиваются на обучающее и тестовое множества: P_0 и $P_\perp$ соответственно. $P_0 \cup P_\perp = P$. Если $\rho(u,i) \in P_0$, то будем обозначать такие объекты i_0. Если $\rho(u,i) \in P_\perp$, то будем обозначать такие объекты $i_\perp$.

Решение проводится на основании информации, принадлежащей обучающему множеству. Далее результирующее множество, полученное в ходу решения, сравнивается с тестовым для выявления качества решения.

Решением задачи $topN$ является множество соседей I_{topN}, который строится за счет фильтрации объектов, производимой по значениям меры близости δ_i. При решении задачи отфильтровываются те объекты, которые не являются соседями для объектов, о которых известно по обучающим, что они близки к активному пользователю:

$$I_{topN} = \{i:(\exists i_0|(i\underset{I}{R}i_0)\} \quad (2)$$

Качество решения задачи оценивается по значению функции оценки ε_{topN}. Существует несколько функций, которые используются в качестве ε_{topN}, к примеру:

- Точность: $P = \frac{\sum_{i\in I_{topN}}(s(i))}{N}$

- Точность по списку длины L:$P@L$ $\frac{\sum_{i \in I_{topN}}(s(i))}{L}$, $L = 1..N$
- Средняя точность:

$$AveP = \frac{1}{\sum_{n=1}^{N} s(n)} \cdot \sum_{L=1}^{N} P@L$$

- $NDCG = 1 - \frac{DCG}{IDCG}$,

где $DCG = s(i_1) + \sum_{n=2}^{N} \frac{s(i_n)}{log_2(n)}$,

где $IDGG$ — идеальное DCG, для которого $\forall\, n = 1..N\, s(i_n) = 1$.

$s(i) = 0$, если $\exists\, i_\perp : i \underset{I}{R} i_\perp$, $s(i) = 1$ иначе. Будем говорить, что если $\varepsilon_{topN} \le \varepsilon_0$, то решение качественно.

Достаточное условие получения качественного решения

Теорема. *Достаточным условием получения качественного решения при применении ОРС (1) является выполнение свойства транзитивности отношения сходства $\underset{I}{R}$ на подмножестве*

$$I_0 \cup I_\perp \cup I_{topN}, I_0 = \{i_0\},\ I_\perp = \{i_\perp\}.$$

Покажем, что если эвристическое утверждение верно и выполняется достаточное условие, то тогда $\varepsilon_{topN} \le \varepsilon_0$.

По решению задачи верно, что $\forall\, i \in I_{topN}\, \exists\, i_0 : i \underset{I}{R} i_0$.

По эвристическому утверждению (1) верно, что $\exists\, i_\perp : i_0 \underset{I}{R} i_\perp$. Пусть транзитивность выполняется. Тогда

$$(i \underset{I}{\mathcal{R}} i_0) \wedge (i_\perp \underset{I}{\mathcal{R}} i_0) \Rightarrow (i \underset{I}{\mathcal{R}} i_\perp).$$

Так как $\forall\, i$ верно, что $i \underset{I}{R} i_\perp$, то $\forall\, i$: $s(i) = 0 \Rightarrow \varepsilon_{topN} = 0$, то есть $\varepsilon_{topN} \le \varepsilon_0$ для любых ε_0.Тем самым мы показали, что выполнение транзитивности отношения сходства является достаточным условием получения качественного решения

Выполнение достаточного условия получения качественного решения зависит от того, какие функции применяются в качестве меры близости, и параметра ε_0.

К примеру, традиционной функцией, которая используется в качестве меры близости в ОРС, является косинус угла между векторами [4]. При, например, $\varepsilon_0 = 0{,}49$ и следующих контах $C_Y(i) = (1,1,0), C_Y(j) = (0,1,1), C_Y(k) = (0,0,1)$, получим, что $i \underset{I}{R} j$, $i \underset{I}{R} k$, но $i \underset{I}{R} k$ не выполняется.

Нечеткая объектно-ориентированная РС

Будем представлять контенты объектов системы в виде нечетких подмножеств множеств характеристик [16]:

$$C_Y(i) = \{y | v(i,y))\},\ v(i,y) \in [0,1],\ y \in Y$$

Контент пуст: $C_Y(i) = \emptyset$, если $v(i,j) \equiv 0$.

Определим операцию объединения и пересечения контентов:

$$c_Y(i) \cup c_Y(j) = c_Y(k) :\ \forall y \in Y : v(k,y) = \max(v(i,y), v(j,y))$$

$$c_Y(i) \cap c_Y(j) = c_Y(k) :\ \forall y \in Y : v(k,y) = \min(v(i,y), v(j,y))$$

Между объектами системы зададим функцию расстояния в виде обобщенного расстояния Хэмминга:

$$\rho_i(i,j) = \begin{cases} \text{Неопределено, если } c_Y(i) \cap c_Y(i') = \emptyset \\ \frac{1}{|Y|} \cdot \sum_{k=1}^{|Y|} |v(i,y_k) - v(j,y_k)|, \text{иначе} \end{cases}$$

Определим отношение близости объектов:

$$i \underset{I}{\mathcal{R}} j \Leftrightarrow \rho_i(i,j) \le \varepsilon_0$$

Расстояние ρ_i обладает метрическими свойствами.

Метод улучшения объектно-ориентированных коллаборативных моделей

Для того, чтобы выполнялось достаточное условие получения качественного решения задачи $topN$, в множество соседей будем помещать только те объекты, для которых $\rho_i^1(i,c) = 0$, где $c = \{i_0 : u_a R i_0\}$, $\rho_i^1(i,c) = min(\rho_i(i,j) : j \in c)$.

Так как $\rho_i^1(i,c) = 0$, то $\exists\, i_0 \in c : \rho_i^1(i,i_0) = 0$. По эвристическому утверждению известно, что $\exists i_\perp : i_0 \underset{I}{R} i_\perp$, то есть $\rho_i(i_0, i_\perp) \le \varepsilon_0$. Так как введенная функция расстояния ρ_i обладает метрическими свойствами, то по свойству неравенства треугольника верно, что: $\rho_i(i, i_\perp) \le \rho_i(i_0, i_\perp) + \rho_i(i, i_0) = \varepsilon_0$. Так как $\rho(i, i_0) = 0$, то $\rho_i(i, i_\perp) \le \varepsilon_0$, то есть $i \underset{I}{R} i_\perp$, и $\varepsilon_{topN} \le \varepsilon_0$ на таком решении.

Практические результаты

Было разработано программное обеспечение, реализующее нечеткую ОРС и ОРС и решение задачи $topN$

Таблица 1. Подзадача top-N

Тип РС	P	AveP	NDCG
ОРС	0.0105	0.011	0.0419
Нечеткая ОРС	0.0004	0.0004	0.0318

в этих системах. Тестирование проводились на множестве данных MovieLens. Характеристики этого множества:

- $|N^k| = 6000$
- $|N^m| = 10000$ — множество фильмов
- $|P| = 1000\,000$
- $|Y| = 18$ — множество кинематографических жанров

Далее в таблице приведены значения оценок при решении задачи $topN$ в нечеткой ОРС и ОРС.

Заключение

В статье проанализированы коллаборативные объектно-ориентированные рекомендательные системы и решение задачи $topN$ в этих системах. Показано достаточное условие, при выполнении которого гарантируется достижение качественного решения данными системами. Однако выполнение достаточного условия зависит от таких параметров этих систем, как функция, используемая в качестве меры близости и пороговое значение меры близости, по которому определяется выполнение отношения близости. Предложена модификация рассматриваемых систем, для которой достаточное условие выполняется.

ЛИТЕРАТУРА

1. P. Resnick and H. R. Varian, «Recommender systems», Communications of the ACM, vol. 40, no. 3, pp. 5658, 1997.
2. Karypis G. Evaluation of item-based top-N recommendation algorithms // in Proceedings of the International Conference on Information and Knowledge Management. 2001. с. 247–254.
3. Deshpande M., Karypis G. Item-based top-N recommendation algorithms // ACM Transactions on Information Systems. 2004. С. 143–177.
4. Su X., Khoshgoftaar T. A survey of collaborative filtering based social recommender systems // Computer Communications. 2014. Т. 41. С. 1–10.
5. Herloker J. Evaluating Collaborative Filtering Recommender Systems // ACM TRANSACTIONS ON INFORMATION SYSTEMS. 2004. Т. 22. С. 5–53.
6. Explaining Collaborative Filtering Recommendations, Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl
7. G. Adomavicius and A. Tuzhilin, «Toward the next genera- tion of recommender systems: a survey of the state-of-the- art and possible extensions,» IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734749, 2005.
8. Umyarov A., Tuzhilin A. Improving Collaborative Filtering Recommendations Using External Data // ICDM '08 Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. 2008. С. 618–627.
9. Wang Jun. Unifying user-based and item-based collaborative filtering approaches by similarity fusion // SIGIR '06 Proceedings of the 29th annual international ACM SIGIR. 2006.
10. Berkovsky S., Kuflik T., Ricci F. Cross-Domain Mediation in Collaborative Filtering // Proceedings of the 11th international conference on User Modeling. 2007. С. 355–359.
11. Item-based collaborative filtering recommendation algorithms / B. Sarwar, G. Karypis, J. Konstan [и др.] // Proceedings of the 10th international conference on World Wide Web. 2001. С. 285–295.
12. Yao W., Xudong L., Min X., Ester M., Qing Y. CCCF: Improving Collaborative Filtering via Scalable User-Item Co-Clustering// WSDM '16 Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, pp 73–82
13. R. Hu and P. Pu. Using personality information in collaborative filtering for new users. Recommender Systems and the Social Web, page 17, 2010.
14. D. Asanov Algorithms and Methods in Recommender Systems// Berlin Institute of Technology // https://www.snet.tu-berlin.de/fileadmin/fg220/courses/SS11/snet-project/recommender-systems_asanov.pdf
15. Miha Grchar and Dunja Mladenich and Marko Grobelnik Data sparsity issues in the collaborative filtering framework.Proceedings of the WebKDD'05 Proceedings of the 7th international conference on Knowledge Discovery on the Web: advances in Web Mining and Web Usage Analysis, pp58–76 Pages 58–76
16. А. Амелькин., Д. П. Понизовкин. Математическая модель задачи top-N для контентных рекомендательных систем // Известия МГТУ МАМИ, 2, с. 26–31

Журнал «Современная наука: актуальные проблемы теории и практики»