An Improved Similarity Calculation Algorithm Used in News Recommender System

SHI Xinchen
Southern University of Science and Technology, Shenzhen, China
518055
shixc@sustc.edu.cn

LUO Zongwei
Southern University of Science and Technology, department of computer science, Shenzhen, China
luozw@sustc.edu.cn

Abstract:

Recommender system applies machine learning and focuses on solving the information explosion problem. The mainstream recommendation algorithms are: content based algorithm, collaborative filtering, and hybrid recommender system. All of them are facing the same problem, it is called similarity computation. Recommender systems usually use similarity to measure the items and users. In this paper, we first introduced some basic knowledge and the development trend of the recommender systems. Then we illustrate an improved similarity algorithm in news recommender system to make sure the algorithm pay more attention to users' interest. The algorithm will modify their weight through users' action. In the experiment, we use the AJS AWC-BC and AK-Means algorithms in real news recommender system and compared with some other algorithm. The result showed that the improved algorithm have better performance in vertical news recommend system.

Keywords:

Recommender system; Similarity Calculation; Jaccard Algorithm; K-Means Algorithm

1 Introduction

News recommendation is a branch of recommender system. Different with traditional recommender system, news is a kind of information which updated very fast and the validity is very short. Fresh news is appeared and disappeared all the time. Traditional Item Based Collaborative Filtering(ItemCF) unable to support such high frequency updates. So, in news recommendation area, User Based Collaborative Filtering(UserCF) is widely used. UserCF is still facing two tricky problems: cold-start and user similarity calculation.

Cold-start problem is to design a personalized recommender system without a large amount of user data and make the user satisfied with the recommendation results. The method to solve this problem is: 1. Using user registration information to solve cold-start problems, that is, using age, gender and other data to recommend some popular items (this method is coarser in size). For example, if it's women, they recommend products that women love using user-tags or importing some external data.

User similarity calculation is the core issues of recommender system. For each feature *F*, the degree of preference for each item of the user with this characteristic is calculated, marked as *P(F, i)*. The more demographic characteristics of users, the more accurate the user's interest can be predicted. *P(F, i)* can be simply defined as the popularity of article *i* in a user with F characteristics. This approach will lead to popular items in the various characteristics of users have a higher weight, not quite consistent with the requirements of personalized recommendation, it is difficult to recommend the user to personalize their characteristics.

In this paper, first we proposed an adjusted Jaccard Similarity Algorithm (AJS) and illustrate its application in news recommendation. Second, we proposed an adjusted K-means algorithm (AK-means). We also provided an experimental study on evaluate users' similarity via users' different behavior on reading news. We defined some action to distinguish whether the user is really interested in certain items and adjusted the weight of the user-item pair when user is interested in the items.

SIMUTOOLS '17, September 11–13, 2017, Hong Kong, China

ACM ISBN 978-1-4503-6388-4/17/09…$15.00
https://doi.org/10.1145/3173519.3173532

2 Related Works

In [1] authors proposed a novel personalized news recommendation framework via implicit social experts. [2-3] studied the long-term and short-term reading preferences of users' interest. [4] uses hypergraph to model various high-order relations among different objects in news data, and formulate news recommendation as a ranking problem on fine-grained hypergraphs. Recommender system needs users' historical actions and interests to predict users' future behavior. Cold-start problem is usually occurred in a new user. When a new user comes, recommendation cannot recommend through his historical data. [5] proposed a browsing-based models to predict next page the user will visit to solve cold-start problem. [6] presented the implementation of our solution to efficiently recommend specific news articles from a large corpus of newly-published press releases in a way that closely matches a reader's reading preferences. [7] introduced a class of news recommendation systems based on context trees to solve anonymous visitors. [8] proposed some novel geographical topic feature models to solve cold-start problem.

3 The Adjusted Jaccard Similarity Algorithm

The Adjusted Jaccard Similarity (AJS) algorithm, combined with the characteristics of news recommender system, would generate better results than traditional content-based recommendation algorithms. AJS algorithm improved the similarity calculation method on the basis of Jaccard similarity. The specific of the algorithm will be introduced as follows in this section.

3.1 The Similarity Calculation Method

Item-based recommendation algorithms depend on the items which user viewed before, and usually recommend via the items similarity to the viewed items. The utility of the project i to user u defined as $\varphi(u,i)$, according to the score of user u to project i. For the news recommendation, Item-based recommendation algorithms usually use the news title, keyword and other information to filter the mass news information.

Based on the content recommendation method, the most matching item is recommended according to the user's viewed project. $Content(j)$ represents the description file for the item J. In this paper, we use the TF*IDF method to extract feature sets from the content of news J. $ContentBasedProfile(i)$ represents the users' news reading preference. In this paper, we use typical article which user have read before. We use Jaccard matrix to measure similarity.

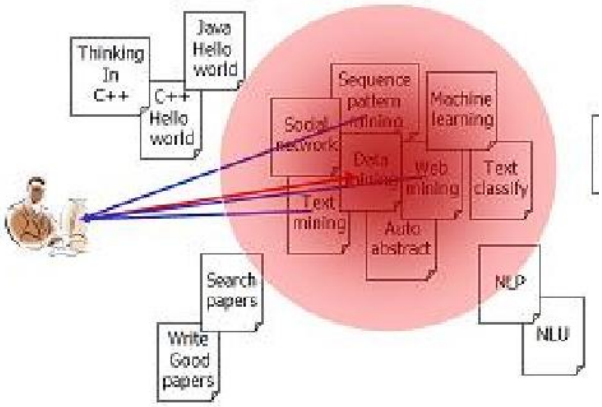


Fig. 1. Content based recommendation

3.2 The Description of AJS

In consideration of portal news content update fast, small content difference. Traditional Jaccard similarity calculate method hardly get the difference from the content and hardly get the interest of the user. We developed an improved Jaccard similarity calculation method called Adjusted Jaccard Similarity(AJS) algorithm. The algorithm can be divided into two parts.

(1) Get n articles from the user have read which can represents user's reading preference we call it RH_i. Use formula 1 and formula 2 to calculate the users' reading preference.

$$RH_i = \{N_{in} | Max_N(E_i)\} \tag{1}$$

$$RH_i \in i_i E_i = \frac{\alpha t_{readed} + \beta O_i}{(t_{current} - t_{published})} \tag{2}$$

In formula 2, t_{readed} means the time spent by the user in reading the news and it is cumulative computation. O_i represents the

number of the user operating behavior to the news i. α and β is the parameter.

(2) We use formula 3 and formula 4 to calculate user i to news j 's action function $\varphi(i,j) = Score\big(ContentBasedProfile(i), Content(j)\big) = Similarity(RH_i, N_{new})$.

$$similarity(N_m, N_n) = \frac{|N_m \cap N_n|}{|N_m \cup N_n|}$$

(3)

$$Similarity(RH_i, N_{new}) = \frac{1}{N}\sum_{n=1}^{N} similarity(N_{i1}, N_{new})$$

(4)

In formula 3 & 4, $similarity(N_m, N_n)$ represents Jaccard similarity of news m and news n. N_m is the text features in the news. AJS algorithm finds out the n_1 news which is the most similar to the news that the user has read through the calculation of the Jaccard similarity between the N news and the recommended news. Traditional similarity methods usually considered the word frequency and timeliness, but in the news recommendation area, because of the limitation of the news field and the authority of information, the words frequency is insufficient to describe users' preference in detail. The AJS algorithm developed is focused on the scoring method of the news. It means that the AJS algorithm screens out the news which represents users' long-term preference better. For the news which was read for a long time, or read repeatedly, this preference presented users' long-tern demand. This kind of news can reflect the user's preference better and has higher discrimination. The AJS algorithm emphasizes the analysis of this kind of news.

3.3 The Description of AWC-BC

Apriori Regular Categorization Based on Content (ARC-BC) is a kind of optimization algorithm. Apriori Weighted Categorization Based on Content (AWC-BC) algorithm is developed based on the traditional association rules ARC-BC. AWC-BC algorithm is no longer limited in the text classification, but can generate the personalized recommendation list. AWC-BC also can be obtained as an expert knowledge of certain area.

Association analysis is the search for correlations between different items that occur in the same event. Formal description of association rules mining problem assumed that $I = \{i_1, i_2, \ldots, i_m\}$ is the item set. T is a random record in the database D and satisfy $T \in I$, T has a flag called TID. Implication X=>Y, where $X \in I, Y \in I, X \cap Y = \emptyset$ is up to, under the condition of the same B, reflected the similarity of item A and item B. The intensity of the correlation is measured by its Support and Confidence. The formula of $Support_{X=>Y}$ and $Confidence_{X=>Y}$ is defined below:

$$Support_{X=>Y} = \frac{\sigma(X \cup Y)}{N}$$

(5)

$$Confidence_{X=>Y} = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

(6)

Where $\sigma(X) = |\{T_i | X \in T_i,\ T_i \in D\}|$, N is the total number of things. If $Support_{X=>Y} \geq Support_{min}$ and $Confidence_{X=>Y} \geq confidence_{min}$, called the association rule is strong association rule.

The basic idea of ARC-BC algorithm is to discover the frequent patterns of each subset instead of finding the frequent patterns of the whole training set. Then, the classification rules are composed of the frequent patterns of each subclass as the antecedent and the class alias as the latter. While ARC-BC is a kind of local association classification method for the text classification, it is divided into two stages: the construction classifier stage and the classification prediction stage. Assume the text set D has n categories. Each category is represented as $C_1, C_2, \ldots, C_n$. The number of J document of the category i is presented as $d_{ij} = \{w_{i1}, w_{i2}, \ldots, w_{im}\}$, m is the number of features. When the k feature in the feature space appears in the document, $w_{ik} = 1$, otherwise $w_{ik} = 0$. For the AWC-BC algorithm, the main steps were shown below:

(1) According to different system requirements, the experts will be divided into n categories of news.
(2) Using other classified news, Apriori algorithm is used to construct classifier.
(3) According to the constructed classifier, the news is classified according to the formula 7 (for some automatic classification systems, the procedure is omitted and the default confidence is 1).
(4) Class preference prediction for the users.
(5) After step (3) and (4), get users preference of a certain news item.

(6) Get the top N news in recommendation list through the formula (9).

$$Support_{max-j} = \begin{cases} 1 & if\ the\ news\ j\ is\ assigned\ on\ category\ C_i \\ Max\left\{Support_{j \to C_j}\right\} & otherwise \end{cases} \quad (7)$$

$$Support_{max-i} = Max\{Support_{i \to C_i}\} \quad (8)$$

$$P_{ij} = Support_{max-i} \times Support_{max-j} \quad (9)$$

P_{ij} means based on AWC-BC algorithm, the effect of user *i* to news *j*. $Support_{max-j}$ means the *Support* of the news j belongs to the C_i class. $Support_{max-i}$ means the *Support* of the user i belongs to the C_i class.

Traditional ARC-BC algorithm mainly applied to text classification. In order to realize the news recommendation function for vertical portals, in this paper, we improved ARC-BC algorithm. Through the association mining of P_i in the user model, we can discover the potential reading interest of the user. In order to make up for the singleness of recommendation results in content recommendation method, the AWC-BC algorithm is developed for improving ARC-BC considering the user's own model, and the user's interest to dig new interest points.

3.4 Adjusted K-Means Algorithm

The Adjusted K-Means Algorithm is improved on the basis of clustering algorithm. The candidate class is introduced into the clustering result to made the recommendation result more diversified and alleviate the click sparse problem of the vertical portal users.

We use formula (10) to estimate the degree of popularity of the user *i* to news. Formula (11) reflected the probability of the user *i* belongs to the class *J*. Through the formula 11 we know that $\mathrm{Cluster}_{\mathrm{i,new}} \in [0,1]$, $|cluster_j|$ represent the user number of class *J*. $|cluster_{j=>new}|$ represent the user in class *J* who have read the news.

$$Cluster_{i,new} = P_{ij} \frac{|cluster_{j=>new}|}{|cluster_j|} + P_{ij'} \frac{|cluster_{j'=>new}|}{|cluster_{j'}|} \quad (10)$$

$$P_{ij} = \frac{l(i,u_j)}{l(i,u_j)+l\left(i,u_{j'}\right)}, P_{ij'} = \frac{l\left(i,u_{j'}\right)}{l(i,u_j)+l\left(i,u_{j'}\right)} \quad (11)$$

The recommender system is improved on the basis of the traditional K-means algorithm, and the user is divided into two categories to meet the high requirements of the vertical portal users on the recommendation results. The preference for news new by the users in the main group and the candidate group is calculated, and the degree of preference for news new by the user I is calculated. Through candidate groups, the accuracy of recommendation will be improved, and the probability of invalid results caused by data sparseness is also reduced.

The AK-means algorithm extracts the user's potential reading preferences from other user's reading history. This method can help users discover new news topics and make up for the shortcomings of content recommendation methods.

4 Experiments and evaluation

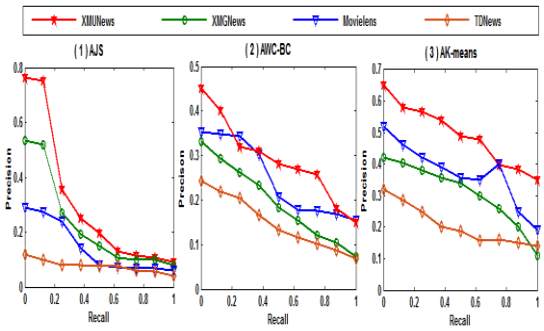
Our experiment data were four kinds of data: Xiamen University News (XMUNews), Xiamen Government News (XMGNews), Movilens and TDNews. XMUNews and XMGNews are typical vertical portals, Movilens is typical recommender system dataset, TDNews is from the NetEase news. The information of the dataset is shown in the table 1.

Table 1. 4 kinds of dataset

Names	Data Sources	User Num	Item Num	Click Num	Fields
XMU News	Xiamen University	9502	6372	932640	University News
XMG News	Xiamen Government	874	3278	78839	Local News
Movie Lens	grouple	94	16	100	Movie recom

	ns.org	3	82	000	mend
TDNews	NetEase	45	500	863	Social News

Separately applies three recommender algorithms, AJS algorithm, AWC-BC algorithm and AK-means algorithm obtains Recall-Precision line chart. From the line graph can be summed up (1) XMUNews data sets, using AJS AWC-BC AK-Means in 4 different dataset have mentioned before, TDNews data has been the lack of recommended results (2) three kinds of recommendation algorithms, AJS algorithm performs better than the other two kinds of recommendation algorithms.



5 Conclusion

In this paper, we present AWC-BC algorithm, an improvements to Jaccard Similarity, as AJS. In the offline experiment, four kinds of dataset were used to compare recommendation results. The advantages of AJS algorithm and AK-means algorithm for vertical portals are verified respectively. Considering the balance of recommended accuracy and timeliness, the hybrid recommendation method proposed in this paper can obtain higher recommendation accuracy, but there are some problems such as too long computation time and less time effectiveness. Next, from the aspects of distributed computing, the AWC-BC algorithm can improve the speed of calculation and ensure the real-time recommendation.

Acknowledgements

This work was partially supported by GDNSF fund (2015A030313782), SUSTech Starup fund (Y01236215), SUSTech fund (05/Y01051814, 05/Y01051827, 05/Y01051830, and 05/Y01051839).

Reference

[1] Lin C, Xie R, Guan X, et al. Personalized news recommendation via implicit social experts[J]. Information Sciences, 2014, 254: 1-18.

[2] Li L, Zheng L, Yang F, et al. Modeling and broadening temporal user interest in personalized news recommendation[J]. Expert Systems with Applications, 2014, 41(7): 3168-3177.

[3] Fortuna B, Moore P, Grobelnik M. Interpreting News Recommendation Models[C]//Proceedings of the 24th International Conference on World Wide Web. ACM, 2015: 891-892.

[4] Li L, Li T. News recommendation via hypergraph learning: encapsulation of user behavior and news content[C]//Proceedings of the sixth ACM international conference on Web search and data mining. ACM, 2013: 305-314.

[5] Trevisiol M, Aiello L M, Schifanella R, et al. Cold-start news recommendation with domain-dependent browse graph[C]//Proceedings of the 8th ACM Conference on Recommender systems. ACM, 2014: 81-88.

[6] Tavakolifard M, Gulla J A, Almeroth K C, et al. Tailored news in the palm of your hand: a multi-perspective transparent approach to news recommendation[C]//Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013: 305-308.

[7] Garcin F, Dimitrakakis C, Faltings B. Personalized news recommendation with context trees[C]//Proceedings of the 7th ACM conference on Recommender systems. ACM, 2013: 105-112.

[8]Chen C, Lukasiewicz T, Meng X, et al. Location-Aware News Recommendation Using Deep Localized Semantic Analysis[C]//International Conference on Database Systems for Advanced Applications. Springer, Cham, 2017: 507-524.