# A Book Recommendation Algorithm Based on Collaborative Filtering

Yuanqing Zhu
Capital Normal University, Beijing 100048, China

***Abstract*—With the development of web technology and human society, people accumulate more and more data on the Internet. It contains numerous resources in this mass of data. How to explore the use of these resources has become an urgent problem. If you analyze these data sets, you can extract the hidden information or discover new knowledge. The method to obtain useful information from computer and techniques is called data mining. Searching engine is a common tool in today's information world. People use Google, Baidu, Bing and other search engines every day. These search service providers have been almost satisfied with the needs of most people. However, because of the universal nature, they can't give the accurate searching results according to users' different backgrounds and needs. Faced with this situation, it proposes the concept of personalized service, personalized recommendation system by establishing a binary relation between the user and information products, using the existing selection process or similar relationship, mining potential objects of interest to each user, and doing personalized recommendations. This article will do a research about the collaborative recommendation algorithm and propose a method about k-means clustering and k-nearest neighbor.**

***Keywords—data mining; personalized recommendation; web spider; the user interest model; collaborative filtering.***

## I. INTRODUCTION

According to China Internet Network Information Center's 38th *Report on China's Internet Development in Statistics*, it pointed out that the number of China's Internet users has reached 710 million people, and that the Internet penetration has reached 51.7% by June[1]. With the popularity of the Internet, the problems that people are facing have been high lighted. The problem of lacking information in the last century, now has gradually changed into over-loaded information, etc. The data processing method can't catch up with the increase of information. Many non-structured data are stored on servers around the world. The huge amounts of information, which contain data mining is the process of discovering insightful, interesting, and novel patterns,[2]will never be discovered if we do not use scientific analysis and data mining. Search engine has solved most problems, such as the overloaded information. But when a user can't describe their needs clearly, the results of the search engines will not be that accurate. In this context, the recommended system came into being. Recommended system tasks are designed to realize user's portraits, describe user's preferences, calculate products according to the user's interest, and then recommend these to users.

Data mining is to obtain data from a large number of novels, potentially useful models [3]. This project is to discover the information that may interest users, and to transform web unstructured data into structured data from the mass of data. We use lexical or syntactic analysis by extracting pages in the document text, and then we can use natural language-related technology to filter or extract useful information about the content. According to the linking structure of sites, we can get the knowledge of these pages and learn useful information from these pages.

The web creates new challenges for information retrieval [4].Personalized recommendation system has been partially applied in various sites, such as the shopping sites which are based on users' recent browsing history. Music service providers always send a daily song to users and these recommendations are based on user's listening records. But the existing library retrieval systems often do not consider the differences among individual users. Faced with different users, the recommendation will give the same book search result. With the increase in users' demands for personalized service, today's service providers have gradually completed services so that they can provide customized service in the future.

In the recommendation system, collaborative filtering is commonly used, based on the recommended method of articles and the like. Different questions need targeted analysis, so in this way it can be sure to use the most appropriate recommended methods. Douban. Inc. (www.douban.com) is a provider of audio books and other information on the work site. The site allows users to tag themselves, record what they read or want to read, and score the book while reading. This paper will do the research according to part of the data in Douban. The situation is just set for the user who has reading record in Douban, rather than new users. It is only for the present data mining.

Retrieval time will return two results, one search result returned by search engine, another from the recommendation system.

## II. RELATED WORK

According to the characteristics of Douban network, we designed the spider to collect certain data of user needs for tests, and then cleaned the data and eliminated redundant data which were useless for the experiment. After forming the structured data, we stored it in local place. Because of the user's modeling techniques which are commonly used, the experiments need to select the vector space model based on the end user as a model.

A large site may have a lot of users. If we use real-time computing, we can't deal with requests on time. If we choose real-time computing to recommend books, the system must calculate all data points in each sieve, which might cause problems. So we decided to let users come into some group offline and then recommend books online. We use K-means clustering method to divide user groups. Once a user searches for some books, the system will first make sure where the user belongs. Then it will recommend books according to the user's group. In this way, it makes decision on what recommendation it will give.

## III. DATA COLLECTION

The amount of information on the web is growing rapidly[4], so web crawler is able to automatically download the page program. The spider collects information from multiple sites, and focuses on further analysis and mining. The basic principle of the in-order spider is: a time of such reptiles get a web page, not considering the full use of resources. According to the set topic keywords, the program maintains a seed URL queue. Then the spider uses seed queue URL to access the destination page, saves useful information or extract URL and adds the seeds end of the queue. When URL queue is empty or until certain conditions are met, the crawler will stop.

Douban provides Application Programming Interface (API) to facilitate the development of research development for developers. Therefore, this article uses the spider program to collect the set of user testing data. We selected thirty users who had a large quantity of reading books to be seed users randomly, then chose other users who had followed those seed users. At the same time, we needed to get the number of books those users read. If the number was smaller than 200, we thought this user didn't have a good reading history.

### A. Book describe

Because personalized service system uses a computer as a platform, the description of user's model must be a computable representation. In other words, this user model doesn't use a natural language to describe user's preference. The model must be targeted to algorithm, or it should have some standardized data structure.

In this case, the user preference model is described by the number of types of books which users have read before. Each book contains a number of tags that describe the content of the book and book type. Because Douban books allow users to create labels, if each label is used as a basis for describing the book, probably because different users have different preferences and some users will create original labels. This is likely to cause an offset on the final recommendation result. Therefore, we chose 380 popular artificial labels and those were divided into 58 categories. Finally, the book content is described by using the space vector model. The data for each dimension is the number of tags which are contained in a class.

### B. User Profile

User data were obtained by spider program, including all the books which users had read, books' tags, identification number, and users' ratings. The specific format is shown in Table I

TABLE I. EXAMPLE OF USER DATA

| *Book* | *Tags* | *Book ID* | *Rating* |
|---|---|---|---|
| The Life and Times of Su Tungpo | biography /LinYutang/Su Tungpo/ Biographies / history /SuShi/ literature / figure | 1000614 | 3 |
| The Minto Pyramid Principle | thought / Pyramid principle / logistic / writing / management / McKinsey / contemplate / reference book | 1020644 | 3 |
| Mountains of Spices | YuJie/ Mountains of Spices /Love/ novel / Letters / Chinese literature / Literature /China | 1030470 | 1 |
| A decoy on the Indus | Senoo Kappa / India /Travel/paint/ travel notes /Japan/ culture /A decoy on the Indu | 1032645 | 4 |
| The Pride and The Prejudice | The Pride and The Prejudice / Austin / novel/ foreign literature / England /Love/ classic / | 1045926 | 3 |

However, since the user has a strong subjective opinion, their behavior and preferences may change over time. So the user's interest is always variable and diverse. At the same time, it will cause some problem if a new user is added to the system. Other factors might interfere with the recommendation process. It will decrease the user experience if the system gets incorrect results,

Models can be used to test and compare various scenarios of behavioral responses and management decisions [5]. As can be seen from the above table, if all the labels describe the user's preferences, you can describe the effect accurately, but due to less popular tags in subsequent computing, it may cause sparse data problem. We use 58 categories of books to describe the model of users. Each dimension of vector means the number of book types which users had read. As formula says:

$$M = \left\langle C_1, C_2, ....., C_{58} \right\rangle \quad (1)$$

## IV. USER CLUSTERS

K-means clustering algorithm is one of the simplest pattern recognition classification techniques.[6]

Clustering is usually divided into hierarchical clustering and partitioning clustering. In hierarchical clustering, the final classification does not specify the number of partitions. Through repeated calculation to improve convergence, the optimal solution under the current situation is found. The split is also divided into hierarchical clustering and aggregation algorithm. Splitting algorithm means that at the beginning of each sample of the entire sample set, it will be divided into several subsets. Each of them is then divided into smaller sets. Cycle the above operation, until a coarse to fine partition sequence is generated. Agglomerative clustering method is to take each object as an initial cluster, and then turn these classes combined into larger clusters. In the cluster partition, each partition is obtained as a data partition, instead the structure is generated as clustering tree. In general, partitioning clustering method will calculate a local functions or globally defined functions to generate a cluster.

### A. Demand Analysis

If it's found in the data set that some users have similar history, or their scoring trend is similarity. We can say that these users belong to the same class, they have similar preferences. In this article, so as to make the similar users into a class, we use K-means clustering method.

### B. Method Description

K-means is undoubtedly the most widely used partition clustering algorithm. [7] First, we randomly selected k users as a seed to form an initial k class. The only point of each class is the focus of the current class. Sequentially calculating the distance between the focus and each user in each class, each data point can be classified according to the distance. If the data points are far from the focus of each class, we ignore this clustering point. If there is an independent data point that does not belong to any class, we say the algorithm is not completed. At this point, we recalculate each class's focus and the distance between it and the rest of the data points. After each round, we see if there still are unclassified data points. If there are, we repeat the steps above. If not, we say that the data set clustered.

To calculate two users' vector, the system uses the Euclidean distance to decide if the users are similar:

$$\sqrt{\sum_{i=1}^{58}\left(V_{mi}-V_{ni}\right)^2} \qquad (2)$$

where 'm' and 'n' represent two users, $C_{mn}$ represents the Euclidean distance between two users. We assume that k = 6, select a small number of users as the test data, and obtain the results shown in the following table II:

TABLE II. ONE RESULT (K=6)

| *Result* | *Group1* | *Group2* | *Group4* | *Group4* | *Group5* | *Group6* |
|---|---|---|---|---|---|---|
| First results | Group 1 : 13 | Group 2 : 4 | Group 3 : 49 | Group 4 : 20 | Group 5 : 25 | Group 6 : 2 |
| Second results | Group 1 : 5 | Group 2 : 6 | Group 3 : 35 | Group 4 : 7 | Group 5 : 2 | Group 6 : 64 |
| Third results | Group 1 : 14 | Group 2 : 5 | Group 3 : 50 | Group 4 : 21 | Group 5 : 26 | Group 6 : 3 |
| Fourth results | Group 1 : 14 | Group 2 : 5 | Group 3 : 50 | Group 4 : 21 | Group 5 : 26 | Group 6 : 3 |

In each row is the result of each cluster, as can be seen from the table, after the third cluster, data begin to converge.

## V. RECOMMENDING BOOKS ONLINE

Suppose a user is searching books. As in the previous step users are already classified, now the system can get the correct class where a user belongs. Because the result of this algorithm has a character that the data in same class is similar, but different between different classes, we can consider that there are other users with similar hobbies in the class where the user belongs. If we recommend books according to this classification, the possibility of users' acceptance is at its largest.

KNN shows the maximum accuracy as compared to the Naive Bayes and Term-Graph[8]. Thus, at the time the user is searching books, we not only receive the result of searching engine, but also the recommendation system.

We select three most similar classes based on previous work and use these classes to finish the next task. First of all, we need to compute if other users in the remaining classes have similar hobbies with this user. In this step, the similarity of two users' vector is described by space distance. The distance is determined by Pearson correlation coefficient. Pearson correlation coefficient is a measure of correlation between two variable indexes, if there is not a standardized data, it tends to give better results. In this method, we can avoid recommending offset caused by huge difference between two users' different score. Before we start this work, the system needs to filter out the users who haven't read the book before.

Because three users will affect the final recommend result, the influences to the final results of these three users are not the same. In order to determine the extent influence to the result of each user, we use the following formula to calculate each user accounted for weight ($E_i$):

$$E_i = \frac{P_i}{\sum_{i=1}^{n} P_i} \qquad (3)$$

where $P_i$ represents to the user's Pearson correlation coefficient.

The system selects some books which were read and scored by these three users. Then we can calculate the forecasting score to the books. The forecasting score ($Q_i$) is calculated as follows:

$$Q_i = \sum_{i=1}^{n} U_{i,j} * E_i \qquad (4)$$

where $U_{i,j}$ is the score which the user 'i' rated the book 'j'. If the forecasting score of the book is high, it means that the user would pay attention to this book and develop an interest. So the system will return some books with highly forecasting score and these books would be the final recommend result.

## VI. EXPERIMENTS

The experiment data is from Douban. We select a user named 'er-mao' randomly to be the test data. Then we get the class where the user belongs and calculate which books he would like to spend time in reading. With this program, we realize that she would like to read Chinese literature and poetry books. So in the next step, we find other users who like literature and poetry books.

Because this user belongs to a class that's clustered in previous steps, we can get the class number and find out other users in this class. In this class, we select three people who

have same hobbies. According to the data of these users, we draw a diagram of their books' score. (Fig.1).

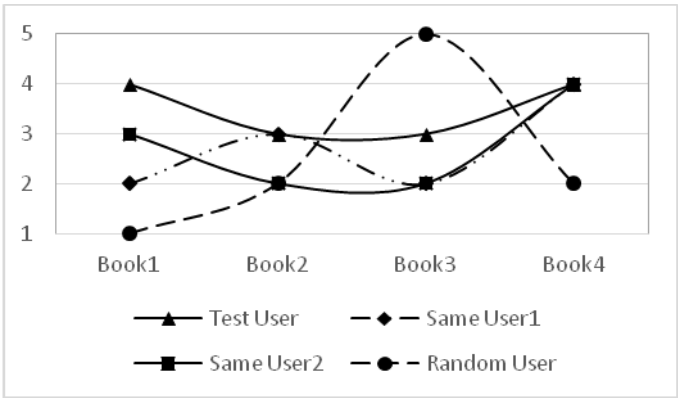


Fig. 1. Schematic diagram of different users' score.

As can be seen from the figure, the 'random user' is a user that we select randomly. This data is to explain that if someone doesn't have similar hobbies with the test-user, this data should be excluded. Otherwise it might mislead following work.

After three users are selected, the system begins to calculate the recommended books. We choose some books that these three users have read. Then we calculate prediction score on these books of 'test user'. According to the test-user's model, the prediction scores of three book are displayed in table III.

TABLE III. TEST USER'S ESTIMATE SCORE

| ***Book name*** | ***Book tags*** | ***User1's score*** | ***User2's score*** | ***User3's score*** | ***Test user's estimate score*** |
|---|---|---|---|---|---|
| Six lonely talk | Aesthetics/Essay/Culture | 3 | 3 | 4 | 3 |
| Snow man | Reasoning/Criminal psychology | 4 | 3 | 3 | 4 |
| All You Need Is Kill | Light novels/science fiction/Japan | 2 | 2 | 3 | 3 |

After the results are counted and sorted out, as can be seen in the table, those three books are the recommend books in the end.

## VII. CONCLUSIONS

With the rapid development of information technology, more and more kinds of information content gradually have fused together and formed a massive data warehouse. Every moment more data are being generated, most of which are useless. However, there are much useful data in it, including preferences and so on. To help general user to get more personalized service, recommend system come into being. The system has been used in E-commerce field, and it's also good to help enterprises to improve competitiveness. More and more researches are being done in the field of book recommendation.

In this paper, I discussed some clustering algorithm and some recommend methods. According to the characteristics of data structure, I provided a hybrid recommendation method. For users with historical data, this method can improve the accuracy of the recommendation.

## ACKNOWLEDGMENT

This work was supported in part by National Science Foundation of China under Grants No. 61303105 and 61402304, the Humanity & Social Science general project of Ministry of Education under Grants No.14YJAZH046, the Beijing Natural Science Foundation under Grants No. 4154065, the Beijing Educational Committee Science and Technology Development Planned under Grants No. KM201610028015 and KM201410028017, and Beijing Key Disciplines of Computer Application Technology.

## REFERENCES

[1] The38th， Report on China's Internet development in statistics [J]. htp://www. cnnic. net. cn, 2016.

[2] Zaki M J, Wagner Meira J. Data Mining and Analysis[J]. 2014.

[3] Latha M, Surya R. Brain Tumour Detection Using Neural Network Classifier and k-Means Clustering Algorithm for Classification and Segmentation[J]. Brain, 2016, 1(01).

[4] Brin S, Page L. Reprint of: The anatomy of a large-scale hypertextual web search engine[J]. Computer networks, 2012, 56(18): 3825-3833.

[5] Willner P. Reliability of the chronic mild stress model of depression: A user survey[J]. Neurobiology of Stress, 2016.

[6] Kantardzic, Mehmed, et al. Data Mining: Ideal, Model, Function& Algorithm. Tsinghua University Press, Beijing, 2003.

[7] Celebi M E, Kingravi H A, Vela P A. A comparative study of efficient initialization methods for the k-means clustering algorithm[J]. Expert Systems with Applications, 2013, 40(1): 200-210.

[8] Bijalwan V, Kumar V, Kumari P, et al. KNN based machine learning approach for text and document mining[J]. International Journal of Database Theory and Application, 2014, 7(1): 61-70.