ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 147

International Conference on Network, Communication, Computer Engineering (NCCE 2018)

Study on Sparsity of Recommender System in University Library

Xu Zhao[1, a)], Guang Liu[2, b)]

[1] *Key Laboratory of Modern Teaching Technology, Ministry of Education, P. R. China, Library, Shaanxi Normal University, Xi'an 710119, China.*
[2] *Network Information Center, Shaanxi Normal University, Xi'an 710119, China.*

[a)] zhaoxu@snnu.edu.cn, [b)] liuguang@snnu.edu.cn

Abstract. Due to the lack of readers' rating data, university library recommender system is facing the problem of sparsity of data. This study proposes a general logarithmic transformation model which can convert the reader's implicit feedback data to the rating, thus alleviating the sparsity to a certain extent. Because logarithmic transformation can use different base, several logarithmic transformation methods are analyzed and compared from different angles. The model-based collaborative filtering is used to compare the recall and accuracy of these methods to make full use of the technical advantage based on matrix factorization to further alleviate the data sparsity problem. The experimental results show that the proposed general logarithmic transformation model can play the role of modifying the data skew, compressing the variable scale, reducing the value of the calculation and so on, and the results of the model are interpretable. Moreover, when the suitable k value is selected, the recommended results of different logarithmic transformation methods can approach the optimal solution in a finite experiment.

Key words: Book Recommender Systems, Model-based Collaborative Filtering, Sparsity, Logarithmic Transformation.

INTRODUCTION

The recommendation algorithm is the core technology of the recommender system [1], and collaborative filtering (CF) [2] is the most widely used recommendation algorithm which relies on the users' rating of the items [3]. However, the lack of readers' rating on books in the library management system will lead to data sparsity seriously when building a university library recommender system. But a lot of readers' implicit feedback data are easily obtained through the library management system, and these data imply readers' preferences for books. So, if we can make full use of the implicit feedback data and convert them into explicit and available rating data, the problem of data sparsity can be alleviated to a certain extent.

Through the analysis of readers' borrowing behavior, we found that the reader's borrowing time to books, one of the readers' implicit feedback, can indicate the readers' preference to books to some extent. If the reader is not satisfied with the borrowed books, the book is generally returned on the same day or the second day, and if the borrowed book is very useful to the reader, then the reader will borrow the book for a long time or even repeatedly. Therefore, readers' borrowing time can be used as the source of readers' rating to the books.

By analyzing the data in the library management system, we can see that some borrowing time are only short in one day, while others are hundreds of days. If the borrowing time is directly taken as a rating, the value is too large, which is not conductive to the calculation. Therefore, a general logarithmic transformation method is proposed in this paper, which makes logarithmic transformation to the reader's total borrowing time on a book, and then the converted value is used as the reader's rating for the book in the recommender system. With the rating matrix, the sparsity of recommender system is alleviated effectively. This part is described in detail in section second.

Apart from the lack of rating data, another main reason for sparsity is the high dimension of the scoring matrix [3]. Therefore, the model-based CF is chosen as the specific recommendation algorithm in the experiment. That reason is

that the model-based CF is based on matrix factorization technology [4], and this technology can decompose the high-dimensional matrix, and further alleviate the sparsity. This part is described in the fourth section.

The rest paper is organized as follows. Section 2 describes the general logarithmic transformation model and explains its function. The sources and processing methods of the data used in the experiment are explained in Section 3. In Section 4, several different methods of logarithmic transformation are compared from different angles. And the model-based CF algorithm is used to compare the recall and accuracy of several logarithmic transformation methods when comparing the recommended effect. It also includes a brief discussion of the results obtained and finally we conclude the paper in the last section.

A GENERAL LOGARITHMIC TRANSFORMATION MODEL

A General Logarithmic Transformation Model

In the book management system, we have the time information of reader 's borrowed books and returned books. The time between borrowing books and returning books is called the length of the reader's borrowing time. If a reader borrowed a book many times, such a borrowing behavior actually implies a kind of implicit feedback, that the reader is interested in this book. Therefore, it is obviously feasible to make use of such implicit feedback to characterize the reader's preference for books and to generate a reader's rating on books under the absence of a direct available rating.

Through the analysis of the borrowing data in the library management system, the borrowing time is from one day to hundreds of days. If the borrowing time is directly used as a rating, some values are large and inconvenient to calculate. We use the following methods. First, for a book, the number of times a reader borrows and the length of each loan is counted. Then the borrowing time is added to the total, and after logarithmic transformation of the total borrowing time, the transformed value is used as the reader's rating for the book. According to this idea, the general logarithmic transformation model is as follows.

For each reader u and each book b, N represents the total number of times that reader u borrowed book b, and n refers to one of them. b_n denotes the date of reader u's nth borrowing of book b. r_n refers to the date of reader u's nth return of book b. And $t_n = r_n - b_n$ are the length of time for the user u the nth borrowed the book b.

So, the total days for user u borrowing the book b is $\sum_{n=1}^{N} t_n$. And the score for the user u borrowing the book b is:

$$\mathrm{P(u)} = \log_a \sum_{n=1}^{N} t_n \quad , a > 0, a \neq 1, t_n > 0 \tag{1}$$

The equation is the rating model. Because logarithmic transformation can use different base numbers, we try to use different logarithmic transformation in the experiment, and the base of logarithm is 2, e, 3, 4, 5, 6, 7, 8, 9 and 10 respectively. The specific content of the experiment is in the fourth section.

The Role of Logarithmic Transformation

Logarithmic Transformation is a common way of data transformation and has many applications [5]. When it is used to transform implicit feedback into rating to solve sparsity, there are following advantages.

First, logarithmic transformation can modify the skewness of data. Before logarithmic transformation, the distribution of the borrowing time is shown in the Fig.1. The horizontal axis represents the length of borrowing time, and the longitudinal axis refers to the number of borrowing times. Obviously, the tail on the right side of the distribution is longer, indicating that there are some very large values, and most of the sample data are concentrated on the left side.

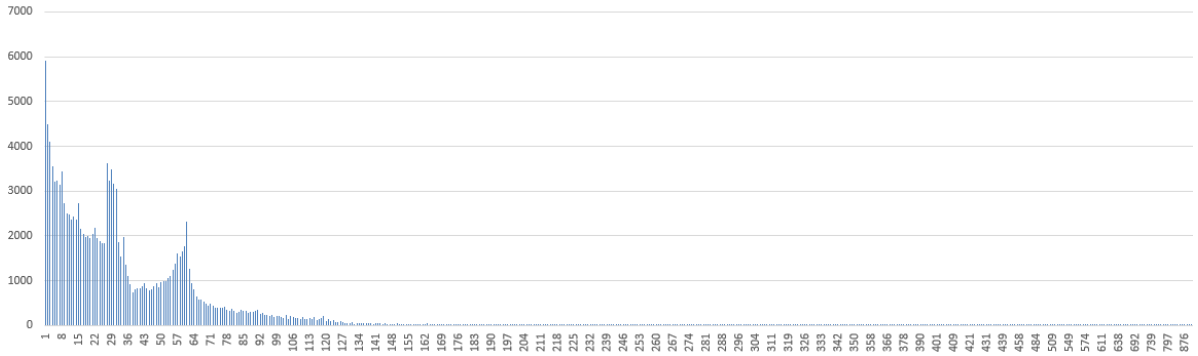

FIG. 1. The distribution of the original data

The logarithmic transformation can reduce the value greater than the median in a certain proportion. After logarithmic transformation, the distribution of data has changed. Taking lnN as an example, the distribution after logarithmic transformation is shown in the Fig.2.

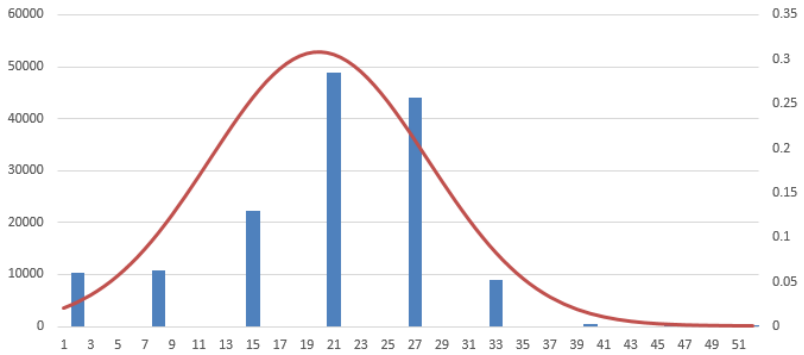

FIG. 2. Data distribution after logarithmic transformation

The horizontal axis represents the fractional interval of rating, which is divided into 52groups. andthe longitudin al axis represents frequency of the rating at each interval. As we can see from the Fig.2, the curves drawn according to histogram are very similar to normal distribution curves. The skewness and kurtosis coefficients of the transformation are calculated. The results show that the data after the logarithm of the original data (the amount of data is greater than 140 thousand) approximately obeys normal distribution, indicating that logarithmic transformation corrects the skewness of data to some extent.

Second, logarithmic transformation can compress the scale of a variable. Long borrowing time is compressed by the logarithmic transformation, and it is true in fact, because the longer the borrowing time, the less obvious the difference between readers' interests. While short borrowing time is expanded through the logarithmic transformation, and it is actually true, because the shorter the borrowing time, the more distinct the difference between readers' interests. For example, there is little different in preferences between 90 days and 120 days of borrowing, while there is quite different between interests of 10 days and 1 day of borrowing.

Third, logarithmic transformation will not change the relative relationship of data, but it will reduce the numerical range and facilitate calculation. This is due to the monotonicity of logarithmic function, when a>1, it is a monotone increasing function on the domain of definition. Therefore, the relative relation between the logarithmic data is invariable, but the logarithmic value is beneficial to the calculation. Because the total value of borrowing time is large, the calculation of these values will make the result range too large, while logarithmic transformation can shrink the value of borrowing time, so that the generated scoring matrix is easy to calculate.

Fourth, after logarithmic transformation, the value is reduced, but it does not cause information loss. Taking natural logarithm as an example, from Fig.2, we know that the transformed data distribution is approximate normal

distribution. Since a positive random variable X is log-normally distributed if the logarithm of X is normally distributed. And its mathematical expectations and variance are as follows [6].

$$\mathrm{E}(X) = e^{\mu+\frac{1}{2}\sigma^2} \tag{2}$$

$$\mathrm{Var(X)} = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) \tag{3}$$

It can be seen that the statistical characteristics of the transformed data can in turn deduce the statistical characteristics of the original data, and there is no loss of data information.

Fifth, after logarithmic transformation, the data of 1 day of borrowing time is changed to 0 after calculation, and we can see that this part of the data is not very few by analyzing the data. Usually, if a book is borrowed and returned on the same day, it shows that the reader does not have preference for the book. Therefore, the data can naturally be used as a negative sample.

DATASETS

We analyzed and processed a four years period data from the sample school's library system, including data export, storage, cleaning, calculation, analysis and so on. And we choose 1000 readers' borrowing records as the datasets whose records are more than 100 times. The final datasets have 146109 records, which contain 1000 readers and 81817 books. For protecting the reader's privacy, we renumber the reader's id and book's id.

According to the base number, the different logarithmic transformation formula is used to calculate the reader's score separately, and the number is integer. The specific score is in 4.2 sections.

Then we adopt 10-fold cross-validation model to split the datasets into training set and testing set.

ANALYSIS AND COMPARISON OF DIFFERENT LOGARITHMIC TRANSFORMATION METHODS

Since the logarithmic transformation can use different base numbers, the experiment is based on the general logarithmic transformation model, and several different logarithmic transformations are selected and compared from several different angles. The first section shows the transformation curves corresponding to logarithmic transformation of different base numbers. The second section is the interpretation of the value after the logarithmic transformation. The third section compares the actual recommended effect in recall and accuracy based on different transformation.

The Corresponding Transformation Curve

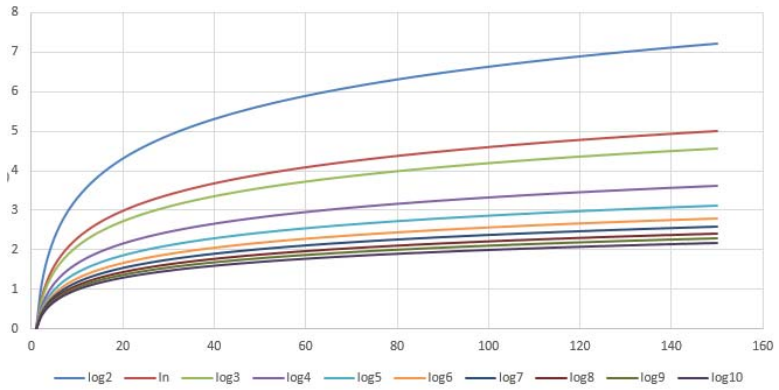

FIG. 3. Logarithmic transformation curve

The logarithmic transformation can extend the small part of the borrowing time, and compress the large part of the borrowing time, so that the reader's borrowing preference can be more clearly defined. The effect of logarithmic

transformation can be intuitively understood from the logarithmic diagram of Fig.3. The horizontal axis of Fig.3 indicates the borrowing time, and the longitudinal axis represents the score after logarithmic transformation. The logarithmic transformation curves correspond to the base of 2, e, 3 to 10 respectively. As we can see from Fig.3, the larger the base number, the greater the extension of the smaller part of the borrowing time, and the stronger the compression of the large part of the borrowing time.

Taking lgN as an example, 40 of the horizontal axis corresponds to 1.6 of the longitudinal axis, that is, the data that is borrowed for 1 to 40 days is extended to the interval of 0 to 1.6 after logarithmic transformation. While the remaining 40 to 150 days' borrowing time, data were projected to the interval of 1.6 to 2.2 only after logarithmic transformation. This achieves the expansion and emphasis on the small part of the borrowing time, while compressing and weakening the function of the large part of the borrowing time.

The Interpretation of the Value After the Logarithmic Transformation

TABLE 1. The value after the logarithmic transformation and the corresponding level

Base	Range of Score	Corresponding Level
2	0-9	10 level (1-10)
e	0-8	9 level (1-9)
3	0-7	8 level (1-8)
4	0-5	6 level (1-6)
5,6,7	0-4	5 level (1-5)
8,9,10	0-3	4 level (1-4)

Based on logarithmic transformation of different base, the original data is transformed into different ranges, as shown in Table 1. The third column corresponds to the rank of the Likert scale [7]. The reason for this correspondence is that the Likert table is a psychological response scale, which reflects the attitude of the respondents. The scores after logarithmic transformation reflect readers' preference on books, therefore, we can interpret our data through correspondence with Likert scale.

The Likert scale is usually divided into 5 levels, and sometimes 4 is also used. The format of a typical 5 level Likert scale is as follows: 1. Strongly disagree. 2. Disagree. 3. Neither agree nor disagree. 4. Agree. 5. Strongly agree. [8] Table 1 shows that there are different degrees of scores in the transformation. Taking the common 5 scores as an example, refer to the description of the Likert scale, it can be interpreted as the reader's 5 levels of preference for a book. Since the score of logarithmic transformation is not a direct statement of the reader's subjective attitude, it is a preference expressed by the reader's actual actions. So, the explanation of the value after the logarithmic transformation is not exactly the same as the Likert scale, which is explained in turn: the reader's demand for the book is very strong, the reader is in a strong demand for the book, the reader is in demand for the book, the reader is in general demand for the book, and the reader does not need the book at all.

Analysis and Comparison of Recall and Accuracy of Recommended Results

Through the first two sections, we have an intuitive understanding and a clear explanation of the results of the logarithmic transformation. In this section, we will compare the different logarithmic transformation methods on the results of the recommendation.

Recall and Accuracy

The prediction accuracy of TopN recommendation system is usually measured by recall and accuracy [9]. $\mathrm{TR(u)}$ denotes a list of books recommended by the reader u based on the behavior of reader u on the training set, and $\mathrm{TE(u)}$ denotes the actual borrowing list of reader u on the test set. The definition of recall and accuracy of the recommended results are shown in formula 4 and formula 5 respectively.

$$\mathrm{recall} = \frac{\sum_{u \in U} |TR(u) \cap TE(u)|}{\sum_{u \in U} |TE(u)|} \tag{4}$$

$$\text{precision} = \frac{\sum_{u \in U} |TR(u) \cap TE(u)|}{\sum_{u \in U} |TR(u)|} \tag{5}$$

LFM

Latent Factor Models (LFM) [10] is a model-based CF method. By taking advantage of the matrix factorization technique, it can tackle the sparsity and predict the unobserved preferences.

The LFM equation is shown as follows:

$$\hat{R}_{ui} = \sum_{k=1}^{K} P_{uk} Q_{ik} \tag{6}$$

For each reader u, P_{uk} denotes the implicit relationship between the reader's preference and the implicit feature k. And for each book i, Q_{ik} is the implicit relationship between the book i and the implicit feature k.

The dot product of P_{uk} and Q_{ik} is approximated to the reader u's score of book i, which is denoted by R_{ui}, and all the values of R_{ui} make up the readers-books rating matrix.

To learn P_{uk} and Q_{ik}, the system minimizes the regularized squared error function on the set of historical ratings:

$$\min_{p_*, q_*} \sum_{(u,i) \in \kappa} (r_{ui} - q_i^T p_u)^2 + \lambda(\|q_i\|^2 + \|p_u\|^2) \tag{7}$$

And the constant λ controls the extent of regularization and is commonly determined by cross-validation. More details about LFM can be inferred to Koren's paper [4].

Recommendation Results and Analysis

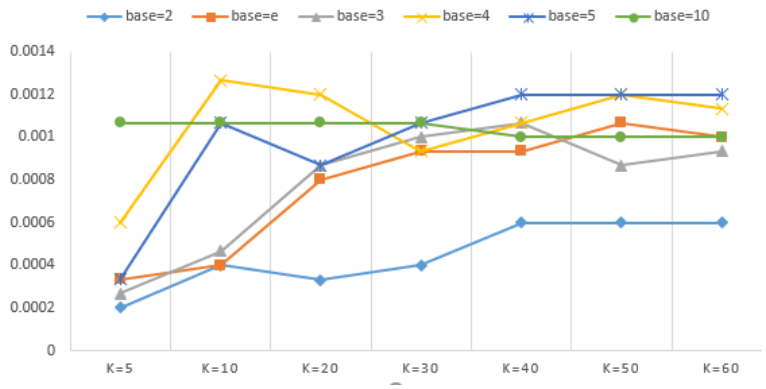

FIG. 4. Recall and accuracy based on different logarithmic transformation

The horizontal axis in Fig.4 represents the implicit feature k, and the longitudinal axis refers to the recall and accuracy. The line charts show changes of recall and accuracy based on the scores generated by different logarithmic transformation.

As we can see from Fig.4, in this experiment, the maximum value of recall and accuracy is the value calculated by the log4N when k=10.

With the change of k value, the recall and accuracy calculated based on lgN is relatively stable, and the results are even better at k<30. In fact, in k=5, it has achieved a good result. The fluctuation of recall and accuracy based on the scores generated by several other logarithmic transformation is more obvious, indicating that they are more sensitive to the change of k value.

In the process of k value change, in addition to the base=2, at least one of the recall and accuracy calculated by each of the other logarithmic transformations is close to the maximum of the experiment. It means that based on the scores generated by these logarithmic transformations, selecting a suitable k value can achieve better recommendation results. In addition, it should be noted that the greater the k value, the longer the computation time will be.

SUMMARY

To solve the sparsity of book recommendation based on CF, we propose a general data processing method based on logarithmic transformation, which converts feedback data into rating to solve the problem of lack of explicit rating. The different logarithmic transformation effects are compared and analyzed through experiments, and the process of analysis gives a way to choose the more appropriate method. When the results of this paper are used in the similar recommendation environment, it is necessary to analyze the actual situation and choose the suitable method based on the general logarithmic transformation model.

ACKNOWLEDGMENTS

Supported by Interdisciplinary Incubation Project of Learning Science of Shaanxi Normal University (SYSX201504).

REFERENCES

1. Resnick P, Varian H R. Recommender systems [J]. Commun. ACM, Vol. 40 (1997) No. 3, p. 56-58.
2. Schafer J B, Dan F, Herlocker J, et al. Collaborative Filtering Recommender Systems [M]. Springer Berlin Heidelberg, 2007, p.291-324.
3. Zhang Juanjuan. Collaborative Filtering Recommendation Algorithm on Data Sparsity Problem form Statistical Perspective [D]. Chongqing Technology and Business University, 2016.
4. Koren Y, Bell R, Volinsky C. Matrix Factorization Techniques for Recommender Systems [J]. Computer, Vol. 42(2009) No. 8, p.30-37.
5. Zumel N, Mount J. Practical Data Science with R [M]. Manning Publications Company, 2014, p.64-79.
6. Johnson Norman L, Kotz Samuel, Balakrishnan N. Continuous Univariate Distributions, Volume 1, 2nd Edition [M]. New York: John Wiley & Sons, 1994, p.207-258.
7. Likert R. A Technique for the Measurement of Attitudes [J]. Archives of Psychology, Vol. 22(1932) No. 140, p.1–55.
8. Fink, A.G. How to Conduct Surveys: A Step-by-Step Guide [M]. Chongqing: Chongqing University Press,2016, p.57.
9. Xiang Liang. Recommender Systems [M]. Beijing: Posts & Telecom Press, 2012, p26.
10. Koren Y. Factorization meets the neighborhood: A multifaceted collaborative filtering model [M]. 2008, p.426-434.