The Research of Recommendation Algorithms in P2P Lending

Yanmei Zhang, Xiangyu Wang, Ya Qian and Hengyue Jia

ABSTRACT

With the rapid development of P2P lending in China in recent years, how to recommend bid projects to the bidder to meet its preferences will become a research hotspot, but there are few studies in this area. And since the P2P lending platform is not a dominant scoring mechanism, it is an urgent need to solve the problem of how to generate a reasonable score matrix. First of all, from the PP Dai.com to climb over more than 320 thousand tenders of the last two years. Then, based on the deeply analysis of the data characteristics of the P2P lending areas, three types of different types of scoring matrix (Boolean, continuous and mixed) are generated. Then, three kinds of classical collaborative filtering recommendation algorithms are used in the experiment. The results show that the performance of the recommendation algorithm has important influence on the performance of the recommendation algorithm, and the performance of cooperative filtering algorithms is different in P2P lending compared with in some other areas. Finally, the experimental results are analyzed, and the relevant suggestions are given for the P2P lending areas.

KEYWORDS

P2P (Peer to Peer) lending refers to loans that directly realized by individuals on the Internet platform. With the rapid development of P2P lending in China, how to recommend bidding projects that meet bidders’ preferences to bidders will be a new hot spot. If platforms can recommend suitable bidding products to users according to preference, it will be helpful to improve users’ satisfaction, and plays an important role in enhancing the user viscosity and increasing the economic benefits of P2P lending platform. However, through literature search and P2P lending platform survey, it can be found that there are just a small number of researches in this field. The idea of this paper is to use existing bidding data of P2P lending platform to
identify bidders’ risk preference, and adopt collaborative filtering algorithm to
determine the nearest neighbor relationship with the same tender preference, finally
recommend bidding project to target bidder according to the nearest neighbor.

Through observing bidding records of P2P lending platform, it can be found that
P2P lending platform data is relatively sparse. Consequently, how to generate an
effective scoring matrix based on tender records become the focus of this paper.
Research idea of this paper is as below: the first step is to acquire historical data from
PPDAI, then transform the tender records into three different forms of scoring matrix,
namely, Boolean matrix, continuous matrix and mixed matrix. Finally, select several
classical collaborative filtering algorithms, and test, compare and analyze the
recommended effects of each algorithm under different scoring matrices.

The structure of following organization is as follows: part 2 introduces selection
basis of collaborative filtering algorithm; Part 3 studies scoring matrix generation
method of P2P lending. Part 4 displays the result of simulation experiment and
analysis. Part 5 shows the conclusion and prospect.

THE SELECTION OF COLLABORATIVE FILTERING ALGORITHM

Memory-based Collaborative Filtering Algorithm

Memory-based collaborative filtering algorithm comprises UCF[ i ] [ ii ]
algorithmand ICF algorithm[iii] [iv]. UCF algorithm is based on the assumption that
users who are interested in the same project have similar interest. The basic idea is that
users who give similar scores to a series of projects have similar interest. When
recommending projects to one of the users that he didn't comment, the score that he
gives to projects can be forecasted according to how his neighbors score. Finally, the
highest K projects can be calculated and recommended to the user. Similar to UCF
algorithm, the thought of ICF algorithm is as follows: items that make the same users
generate interest are similar. The score that users give to projects can be calculated by
calculating similarity between projects. The pseudo code of both ICF and UCF are
omitted.

Model-based Collaborative Filtering Algorithm

Model-based collaborative filtering algorithm recommends projects by training a
model, such as matrix decomposition[v] [vi] and graph-based method [vii] [viii], etc.
The LDA mode[ix][x] is a kind of theme generation models in the field of machine
learning. Generative model agrees that each document word chooses a topic with a
certain probability, and chooses word from the theme to produce with a certain
probability. Topic model extracts themes concentrated in the document impliedly by
studying and then returns documents-matrix θ and themes-word matrix Ø, and
respectively shows document’s probability distribution on various topics and topic’s
probability distribution on various words. When used in personalized
recommendation, the LDA model considers user as document, project as a word. The
process that user generates theme is similar to the process that corpus generates a
series of themes. The project list which is recommended to the user is similar to words
that belong to the document in the list with the highest probability. This paper uses
Gibbs sampling method to solve the LDA model, and calculate θ, Ø matrix. Gibbs
sampling pseudo code is omitted.
GENERATION THE SCORING MATRIX

In personalized recommender systems, users-project evaluation matrix usually has two forms. One is continuous matrix, and the other is Boolean matrix. Generic continuous matrix is applied to the scenario of rating explicitly [xi]. For example, in the Movielensix data set, users rate the films on a five-point scale. Users need to score some films actively and the score is the basis of recommendation. In real scenes, collecting user's explicit score is difficult, but users’ preferences can be collected by analyzing users’ behavior like clicking, no clicking, purchasing, or no purchasing [xii]. This kind of score matrix acquired implicitly is Boolean type. Boolean matrix can only show whether user is interested in each project, not represent the level of interest because data is Boolean. Continuous matrix avoids occurrence of this kind of problem, and it can be better on representing interests of users.

The tendering and bidding process of P2P lending platform is similar to commodity trading process of electronic commerce website. Borrowers who are equal to sellers create loan programs in order to obtain long-term or short-term working capital loan project. Loan projects are equal to goods. The borrowers actually sell their credit. Lenders who are equal to buyers choose right projects to invest after comparing and assessing. However, P2P lending platform is different from the electronic commerce website. The commodity price in E-commerce website is generally determined. Buyers buy different commodities. It is impossible to illustrate the differences of different buyers’ interest degree through comparing various commodity costs. Bidders can determine tender amounts on P2P lending platform. When bidders spend a lot of money on one project, it indicates his preference of return and risk is consistent with the project, and interest score is higher. Few investment amounts indicate user believes the project's risk and income do not meet his preferences and score on this project is lower. Based on the above analysis, P2P lending data is consistent with the following methods, and generate Boolean matrix and continuous matrix:

1. Boolean matrix. When bidder investing a project, the value of the scoring matrix is set to 1 and the rest is set to 0.

2. Continuous matrix. When bidder investing a project, the value’s calculation formula of scoring matrix is as follows:

   \[
   \text{rating}_{ij} = 100 \times \frac{\text{amount}_{ij}}{\text{sum}_i} \tag{1}
   \]

   In the formula (1), \( \text{rating}_{ij} \) express the score user i grade project j; \( \text{amount}_{ij} \) express bid amount user i invest in project j; \( \text{sum}_i \) express the total amount of the tender bidders i.

   In the simulation test, we found that the data processing method of formula (1) can make some bidders’ score data inflated, thus affecting the recommendation effectiveness. The possible reasons of inflated data are as follows: ① A small number of investment projects led to great chance. ② bidders encounters cash flow difficulties, and they consequently do not have excess funds to encounter projects they are interested in. ③ the bidder's inadvisable investment behavior will make the score on the bidding project to be overvalued. Continuous matrix, therefore, requires further processing to make lower numerical data keep same, and high numerical multiply factor to form a hybrid matrix between Boolean matrix and continuous matrix.
(3) Mixed matrix. On the basis of continuous matrix, the high value multiply factor. When processing data, scores narrow down to 1-5 points according to the distribution of the scale. 20% of higher scores are set to 5 points, while the scores between 20% to 40% are set to four points, and so on.

Therefore, in this article, as far as characteristics, we will process the P2P lending data to three different forms of rating matrices, namely Boolean matrix, continuous matrix and mixed matrix. Now we will test the recommendation effect of several typical collaborative filtering algorithms under three different scoring matrices.

**SIMULATION EXPERIMENT**

**Data Acquisition and Preprocessing**

320,000 bid projects, more than 400,000 records from October 2013 to March 2015 are fetched from PPDai by LocoySpider as the data set used in study. First of all, on data sets, the pretreatments are conducted by nullity-error way (including zero value and empty value) and repetition-error way (the same bidder bid on the same tender many times). Then, 9,373 records are selected as the final data through random sampling, 90% of them as training set, and 10% as test set. Finally, three types of scoring matrixes are generated in accordance with section 4 methods.

**Evaluation Index**

In peer-to-peer lending bidding record data set, if a bidder bide a bidding project, we think the bidder is interested in the bidding project, and system generates corresponding score. The projects that bidder is not interested in cannot be measured. That is to say, the data set can only confirm projects that bidders are interested in, and not confirm the project that bidders are not interested in. So the accuracy (Precision) index cannot be computed, therefore the recall rate, diversity and novelty[xiii] are used to evaluate the effect of recommendation algorithms.

1. Recall rate (Recall)
   
   Recall rate indicates proportion the recommended results number "hit" to the projects number "sample" the user is interested in. Formula is as follow:
   
   \[
   \text{recall} = \frac{\text{hit}}{\text{Sample}}
   \]

   The higher recall rate means the better performance of recommendation system.

2. Diversity (Hamming, HM)

   Diversity indicates the differences of each user's recommended list and whether the system recommends each user according to users’ personality characteristics. Diversity generally use the Average Hamming distance (Average Hamming short) to measure. Hamming distance formula of user i and user j recommended list is as follow:

   \[
   \text{Hamming}_{ij} = 1 - \frac{Q_{ij}}{L}
   \]

   Among them, \(Q_{ij}\) express the same project that appear in recommendation lists of user i and user j; \(L\) express the length of recommendation list. The range of Hamming distance is 0 to 1 indicates two users are recommended the same list. 1 indicates two users’ recommendation lists are completely different. The diversity of the
recommendation system can be measured by the average hamming distance between any two users. The greater average hamming distance means the higher diversity.

(3) The Novelty (Novelty)
Novelty indicates the ability that recommendation system recommends unfashionable project to users. If recommendation system only recommends higher overall popular projects, the recommended result is not innovative though accuracy is high. Novelty can use average recommended project popularity to indicate. The formula is as follow:

\[ \text{Novelty} = \sum \frac{\text{pop} \cdot \text{itemNum}}{\text{itemNum}} \]

Lower average popularity means greater novelty. Higher average popularity means smaller novelty.

Results and Analysis

PARAMETER SELECTION

Reference to related research on parameter setting method of the LDA algorithm hyperparameter \( \alpha \) and \( \beta \), this article selects conventional values, namely \( \alpha = \frac{50}{\text{topics}} \), \( \beta = 0.01 \). In order to fairly compare the performance of the algorithm, related parameters of each algorithm is adjusted to make algorithm achieve optimal performance. Thus, through repeated tests, the neighbor number makes UCF algorithm and ICF algorithm have the best performance as well as the topics make the LDA algorithm have the best performance can be concluded. Specific as shown in table 1:

| Algorithm | Boolean matrix | Continuous matrix | Mixed matrix |
|-----------|----------------|------------------|--------------|
| LDA       | 4              | 8                | 6            |
| UCF       | 110            | 120              | 110          |
| ICF       | 280            | 300              | 280          |

THE ALGORITHM PERFORMANCE COMPARISON UNDER DIFFERENT RATING MATRIXES

In order to fully compare performance of UCF, ICF and LDA algorithm, the Recall rate (Recall), diversity (HM) and Novelty of each algorithm are tested respectively under Boolean, continuous and mixed three types of score matrixes. As shown in figure 1, figure 2 and figure 3.

![Figure 1. the performance of each algorithm under Boolean matrix.](image)
As is shown in figure 1 to figure 3: First, in terms of the effect of three kinds of recommendation algorithms, the effect of ICF algorithm under three kinds of matrixes are always better than UCF algorithm. When LDA algorithm is directly used in Boolean matrix and continuous matrix, the result is bad. When rating matrix is Boolean matrix, matrix binary sexual determines each word in the user documentation comes once when LDA algorithm progresses. The ability of theme generation is weak, and LDA algorithm effect is bad. When scoring matrix is continuous, the binary matrix problem is overcame, but LDA algorithm effect is still poor because high value data that interfere with theme generation process is not considered. Based on continuous matrix, the high value data are multiplied by factors to weaken influence, and then form mixed matrix. The topic emergency is relatively accurate. LDA algorithm is effective.

Secondly, in terms of three types of score matrixes, each algorithm’s performance is best under mixed matrix. In particular, when recommended amount is less than 50, the recommended effect under three types of score matrixes are similar. When recommended amount is between 50 to 100, the recommended effect under mixed matrix and continuous matrix are better. When recommended amount is more than 100, the recommended effect under mixed matrix is best.

CONCLUSION

In this article, through analyzing lending data characteristics in P2P field, three different forms of scoring matrixes are formed. Based on three types of score matrixes, several typical collaborative filtering algorithms process simulation experiments. Through experimental study, it can be found that LDA algorithm performance well in some areas but not better than UCF algorithm and ICF algorithm when applied to P2P lending. Score matrix under different forms (Boolean, continuous and hybrid) has important influence on algorithm performance. In practical application, we are hoping to choose appropriate score matrix and recommendation algorithm according to platform data characteristics. The score matrix and recommendation algorithm can be determined by deducting P2P lending platform data in one period.
In the next step of research work, for high value data processing problems under mixed matrix, how to determine a reasonable threshold and penalty factor to make the score reflect user's interests more accurately is also worthy of improvement. In addition, how to effectively handle the sparse data sets to make the results more accurate needs further study.

ACKNOWLEDGEMENTS
This paper is supported by National Natural Science Foundation of China (No. 61309029, 61273293).

REFERENCES
1. J. Herlocker, J. Konstan, A. Borchers, and J. Riedl, An Algorithmic Framework for Performing Collaborative Filtering, in Proc. 22nd Int'l ACM SIGIR Conf. Res. Dev. Inf. Retrieval, pp. 230-237 (1999).
2. R. Jin, J. Chai, and L. Si, An Automatic Weighting Scheme for Collaborative Filtering, in Proc. 27th Int'l ACM SIGIR Conf., pp. 337-344 (2004).
3. [1] M. Deshpande and G. Karypis, Item-Based Top-n Recommendation Algorithms, ACM Trans. Inf. Syst., vol. 22, no. 1, pp. 143-177 (2004).
4. [11] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, Item-Based Collaborative Filtering Recommendation Algorithms, in Proc. 10th Int'l WWW Conf., pp. 285-295 (2001).
5. [1] Salakhutdinov R., Mnih A. Probabilistic matrix factorization. Advances in neural information processing systems, 2008, 20: 1257-1264.
6. [1] Liio X., Ouyang Zhang X. Improving Matrix Factorization-Based Recommender via Ensemble Methods. International Journal of Information Technology and Decision Making, 2011,10(3): 639-561.
7. [1] Fouss F., Pirotte A, Renders J.M.? et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. Knowledge and Data Engineering, IEEE Transactions on, 2007, 19(3): 355-369.
8. [1] Song X., Tseng B.’ Lin C.’ et al. Personalized recommendation driven by information flow. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006. 509-516.
9. [1] Qi Liu, Enhong Chen, Hui Xiong, Chris H. Q. Ding, and Jian Chen, Enhancing Collaborative Filtering by User Interest Expansion via Personalized Ranking IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, 42 (1), 2012.
10. [1] Blei D., Ng A., Jordan M. Latent dirichlet allocation. The Journal of Machine Learning Research, 2003, 3: 993-1022.
11. [1] Koren Y., Bell R., Volinsky C. Matrix factorization techniques for recomimender systems. Computer,2009, 42(8): 30-37.
12. [1] Pan R., Zhou Y., Cao B., et al. One-class collaborative filtering. Proceedings of Data Minings 2008. ICDM'08. Eighth IEEE International Conference on. Ieee, 2008. 502-511.
13. [1] Yuxiao Zhu, Linyuan Lv. Evaluation Metrics for Recommender Systems. Journal of University of Electronic Science and Technology of China, 2012, 41(2): 163-175.