An Explanatory Matrix Factorization with User Comments Data[*]

Donghyun Kim
Department of Industrial and Systems Engineering
Korea Advanced Institute of Science and Technology
South Korea
dhk618@kaist.ac.kr

Hayong Shin
Department of Industrial and Systems Engineering
Korea Advanced Institute of Science and Technology
South Korea
hyshin@kaist.ac.kr

ABSTRACT

Matrix factorization is one of the crucial algorithms of the Recommendation system. It implies that the relationship between user and contents can be explained by hidden latent variables. However, it is not intuitive to understand the meaning of these hidden latent variables. Therefore, this study suggests a way to learn the meaning from supplementary data such as comments and use in matrix factorization. The data used in this study is user comment data from Naver which is the largest web platform and also the largest Webtoons (Web comics) platform in South Korea. We show that the suggest method which uses the supervised latent variable also fits well with users with the distinct tendency compare to conventional matrix factorization.

KEYWORDS

Matrix Factorization, Explanatory Analysis, Latent Dirichlet Allocation

1 INTRODUCTION

Recommendation system (RS) is referred to collecting information to analyze user's taste. Numerous methods have been proposed for the RS, and one of overwhelming method is matrix factorization (MF) which is predicting a missing value of a score matrix composed of evaluation for contents given by the user [3]. MF improved the quality of RS significantly, but there are some issues such as a cold-start problem, insufficient explanatory power, etc. MF decomposes into low rank matrices with K latent features and make the original score matrix treatable, but it was difficult to analyze the meaning of each latent features.

There are a lot of works that uses user reviews to assist RS. Also, it is shown that the appropriate latent factor model using topic selection with LDA is better than the existing model [2, 4]. However, previous studies assumes that there exist score matrix and use reviews to make better while not only this study does not have score matrix but also this focus on the explanatory power of MF not the RMSE itself. The methodology itself is not new as part of research using user reviews to make better RS, but the two popular methodologies, MF and Latent Dirichlet Allocation (LDA), have been mixed appropriately and give exploratory power. It also differ as using user comment data about Webtoons which was not used previously.

2 EXPLANTORY MATRIX FACTORIZATION

In this study, we introduce a method to utilize domain knowledge by combining LDA and MF. The LDA has explanatory power on topics, and the MF is explaining the relationship between the user and the contents with hidden latent variable. However, the meaning of each latent variable is difficult to grasp. Therefore, we first derive explanatory power from the supplementary data $(S_i, i = 1 \ldots I)$, such as user comments or the report using LDA. The LDA assumes that several topics are mixed in each document, and the analysis results can analyze the themes of documents. There are many LDA algorithms to infer topic [5], so we do not describe conventional LDA algorithms in detail. Let $t_k, k = 1, \ldots, K$, be a topic from LDA, then each user i can be represented as vector $u_i = (u_{i1}, \ldots, u_{iK})^T$, where $u_{ij} = P(t_j | S_i)$. This value indicates how much a specific user talks about a particular topic, which is an indirect indicator that shows what the user likes. Therefore, we can make a user-topic relationship matrix $U = (u_1, u_2, \ldots, u_I)$ for whole user and use it directly in MF. The core of MF is to divide an user-contents rating matrix M into two low-rank matrices $M = U^T L$ which needs to be estimated. However, if U can be obtained sufficiently from supplementary data, MF turned into a simple matrix inverse problem. Thus, L can be obtained simply through the Moore-Penrose pseudoinverse with $L = (UU^T)^{-1}UM$.

3 EXPERIMENT

3.1 Data Collection and Refinement

We use user comment data from Naver, which is Korea's largest web and Webtoon platform. However, since Naver does not provide any formalized data, the data was collected and refined through web crawling by Python. The collected data contains 4 features (Title-Episodes-userID-Comment). The raw data has over 100K users, 151 Webtoons, 21927 episodes and over 110 million comments. Since this raw data needs more than 10TB of capacity, due to hardware limitations, we limited to small size data. Also, unlike commonly used reviews, there are a lot of useless data because comments can be written without any restrictions such as

RecSys 2017 Poster Proceedings, August 27-31, 2017, Como, Italy

multiple comments is allowed in same item. Therefore, in this study, we chose 3,000 users who kept the grammar as much as possible and wrote over a reasonable length (more than 70 characters in average) for certain period consistently (at least 12 weeks). Compared to all data, 3000 users are quite small numbers, but since the data used in this paper is very different from the user review usually used in other papers, it was important to refine useful data before analysis. This data contains 149 Webtoons, 1.1 million comments.

3.2 Experiment Settings and Result

The topic is modeled through the LDA with selected 1.1 million comments. In this experiment, the number of topic K was set to 10, 20, and 30, and the hidden latent variables of MF were also set to be the same in each case. Topic selection is very important task, but it is too vague to use the whole as it is. Therefore, some topics are collected through each Webtoon, and the some topics are obtained by whole data. Some of the noticeable topics are listed in Table 1. Topic 3 is mainly composed of words about stories of comic, and Topic 7 is made of the drawing style of comic. Even though not all topics can be identified as the above topics, but there are more topics that can be interpreted, such as the attitude of the artiest, etc.

Table 1: Noticeable topics from comment data ($K = 30$)

Topic 3	Topic 7	Topic 21
Sick of	Beautiful	☺
Crazy	Sick of	Funny
Main character	Drawing	Best comments
Story	Color	Clear

Each user vectors u_i are constructed by the topics we obtained. We use cosine similarity which is most commonly used. Using this similarity, we construct a matrix U to be used in MF and simply obtain other matrix L. We compare RMSE with original MF. In this paper we use most basic MF algorithm [6] as conventional MF algorithm. The score matrix M used in this experiment is composed of 0 and 1 which indicates whether the user sees a certain comic, not the score rating. We assume that the user only sees the comics they commented on. In order to measure the RMSE, about 15% of each user data was randomly deleted. Therefore, we learned with 85% of the data and observe the difference between the erased 15% actual data and the predicted data. As can be seen from the results Figure 1, it cannot be concluded that the overall data performance is better than conventional MF. However, when compared only for those with distinct tendencies, whose $max(u_i) - mean(u_i) > \alpha$ (in this paper $\alpha = 0.4$ is used) which means user i has at least one noticeable topic that can be categorized more clearly than other users, it can be seen that the suggested method using the LDA is slightly better than the conventional MF method. In other words, we can see that the unsupervised latent variable which is conventional MF fits better with users who judge the contents with a complex view, and the suggested EMF which uses the supervised latent variable fits well with users with a simple view. This result cannot be regarded as meaningful for RMSE itself, but it can be implied that it has a similar RMSE even though it is obtained by simple matrix inversion using a relatively interpretable latent variable rather than the existing method. The reason why RMSE is lower than other studies is because it is not to predict the score, but to determine whether user sees a specific Webtoon, so we calculate RMSE with a rounded value which is 0 or 1.

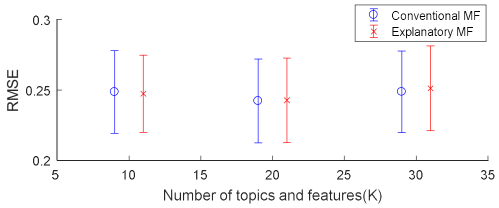

Figure 1: RMSE plot with the full user data

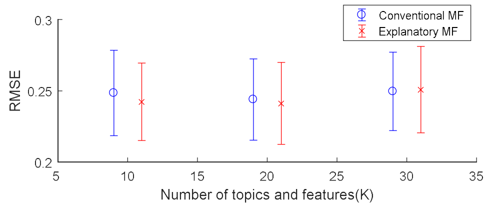

Figure 2: RMSE plot with the distinct tendency user data

4 CONCLUSIONS

In this paper, the two popular methodologies, MF and LDA, have been mixed appropriately and shows some extra synergy. We conducted experiments with comment data of Webtoons and shows the suggest method works quite well as much as conventional MF, and some cases it works better. This study did not fully use the comment data that is currently available. Webtoon is a content that is published one episode a week, so we think it will be very influential in time. Also, this study is domain specific research and the proposed algorithm is used only in this domain, so the extensive research with certified data set is needed to generalize the algorithm.

ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT & Future Planning (2017R1A2B4006290).

REFERENCES

[1] Bobadilla, Jesús, et al. 2013. Recommender systems survey. Knowledge-based systems 46 (pp.109-132).
[2] Seroussi, Yanir, Fabian Bohnert, and Ingrid Zukerman. 2011. Personalised rating prediction for new users using latent factor models. Proceedings of the 22nd ACM conference on Hypertext and hypermedia (pp.47-56).
[3] Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. 2013. Recommender systems survey. Knowledge-based systems, 46, 109-132.
[4] Chen, Li, Guanliang Chen, and Feng Wang. 2015. Recommender systems based on user reviews: the state of the art. User Modeling and User-Adapted Interaction 25(2) (pp. 99-154).
[5] Alghamdi, Rubayyi, and Khalid Alfalqi. 2015. A survey of topic modeling in text mining. I. J. ACSA 6.1 (pp. 147-153).
[6] Koren, Yehuda, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. Computer 42(8).