Paper:

Proposal of Generation of Rating Matrix Based on Rational Behaviors of Users

Kenshin Moriyoshi, Hiroki Shibata, Yasufumi Takama

Tokyo Metropolitan University 6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan
E-mail:moriyoshi-kenshin@ed.tmu.ac.jp

Abstract. This paper proposes a method to generate an artificial rating matrix based on user's rational behavior, with the aim of generating a large-scale rating matrix at low cost. Collaborative filtering is one of the major techniques for recommender systems, which is widely used because it can recommend items using only a history of ratings given to the items by users [1]. However, collaborative filtering has some problems such as the cold-start problem [2] and the sparsity problem [3], both of which are caused by the shortage of ratings in a database (rating matrix). This problem is particularly serious for services that have just started operation or do not have a large number of users. The proposed method generates a rating matrix with no missing values using users' rating probabilities, which are obtained from the distribution of their actual ratings. The final synthetic rating matrix is generated after adjusting its sparsity by introducing missing values. The validity of the proposed method is evaluated by comparing the synthetic data in terms of the similarity of the distribution of several statistics with that of the real data. Synthetic data are generated by the proposed method and generated based on regular but meaningless behavior.

Keywords: Recommendation, Synthetic Data, Long-tail

1. Introducution

In recent years, with the development of information technology, a large amount of information has been distributed on the Internet. It makes it difficult for users to find data that matches their tastes and intentions among the huge amount of data. To solve such a problem, recommendation systems are widely used as a technology for finding and presenting useful information to users.

Techniques used for recommendation can be broadly classified into content-based filtering and collaborative filtering. Collaborative filtering is widely used because it can recommend items using only a history of ratings given to the items by users [1]. However, it is known to have some problems such as the cold-start problem [2] and the sparsity problem [3], both of which lead to the degradation of recommendation accuracy. Cold-start problem is caused by the shortage of the ratings of new users and items. Sparsity problem corresponds to the situation where the number of ratings in a database (rating matrix is considerably few compared to the number of items and users).

As a solution to these problems, this paper focuses on the generation of synthetic rating matrix. Generating synthetic data as alternatives for real data has been studied in the machine learning domain [4, 5]. As synthetic data is effective as an extension or substitute for real data, it is also expected to be useful for recommendation. Although related techniques such as the complement of missing values in a rating matrix has been studied, to the best of our knowledge, generating a full synthetic rating matrix has not been fully studied.

This paper proposes a method to generate a synthetic rating matrix by assuming user's rational behavior, with the aim of generating a large-scale rating matrix at low cost. Rational behavior is defined as a behavior that can be explained rationally, such as " a user who likes popular items will give high ratings to popular items."

The proposed method classifies items into pop items and niche items. Users are also classified into pop-item-loving users, niche-item-loving users, and others. By assuming different rating behaviors for different user types, a rating matrix without missing values is generated. The final synthetic rating matrix is generated after adjusting its sparsity by introducing missing values.

In the experiment, synthetic rating matrix is generated based on the statistics of Movielens[1] dataset. The quality of the synthetic rating matrix is evaluated by calculating the similarity of the distribution of several statistics such as item's average rating and precision between real and the synthetic data. Furthermore, the validity of the employed rational behavior is evaluated by comparing the synthetic data generated by the proposed method to that generated from regular but meaningless behavior.

2. Related Work

2.1. Popular / Niche Items

It is known as the rule of thumb that 80% of the outcomes or results comes from 20% of causes. It is also called Pareto law. In offline retail stores, it is believed that the top 20% of products sold (pop items) accounts

1. https://grouplens.org/datasets/movielens/

for 80% of total sales. Therefore, hit items have been prioritized for handling due to store space and management costs.

On the other hand, online retailers do not need store space and have lower management costs, which makes it easy to handle niche items in addition to popular items. As a result, sales of niche items could exceed those of hit items in addition to popular items. Anderson proposed the long tail theory [6], which states that elements and factors that occur infrequently account for a non-negligible proportion of the total. It has been adopted as the strategy by Amazon.com[2] and Netflix[3] [7].

Recommendation methods that take into account popular items and niche items have been studied. For example, the elimination of popularity bias, which causes a tendency to recommend popular items, is one of the major issues for recommender systems. Yoshida et al. proposed a method to remove the popularity bias by weighting the predicted ratings of item-based collaborative filtering using the user profile created from the interest level of the target user and the item's inverse popularity [8]. Experimental results showed that the frequency of recommending niche items increased [9].

2.2. Synthetic data

Supervised learning requires training data with correct labels, and publicly available datasets have been prepared for various domains such as MNIST[4] and Movielens. However, if a sufficient amount of data is not available, it is necessary to manually assign labels to data for creating datasets, which is time-consuming and costly. As a solution, an approach to artificially generate synthetic dataset of which the statistics is close to real data has been studied.

As mentioned above, synthetic data is effective for extending, augmenting, and substituting real data by adding artificially generated data to target datasets. In addition, it is also effective for protecting the privacy of real data. From these viewpoints, various methods have been proposed for generating synthetic data.

Typical methods of synthetic data generation are classified into two types: statistical methods and machine learning-based methods. As for statistical methods, Alexander et al. have proposed a method for generating synthetic training data by automatically labeling the unlabeled data. To determine labels to be assigned to data, a labeling function that is defined based on heuristics. They showed that the classifiers trained with the generated data could achieve high accuracy for classifying diseases [4]. As for machine learning-based methods, Maayan et al. have proposed a method for generating synthetic data of medical images by using GAN (Generative adversarial networks). In addition to an original dataset, the generated data was used to train CNN (Convolutional Neural Network) for medical image classification. They showed that the accuracy of image classification was improved by augmenting the original small dataset with the generated synthetic data [5].

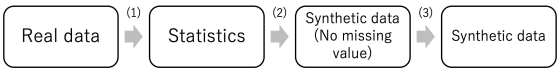


Fig. 1. Outline of Proposed Method

3. Method

3.1. Outline of Proposed Method

Figure 1 shows the outline of the proposed system. The system generates synthetic data including missing values with the following procedures.

(1) Extracting statistics to be used for generating synthetic data from real data

(2) Defining and applying Rating Generation Function to generate rating matrix with no missing values using statistics as arguments

(3) Defining and applying Missing Value Generation Function to insert missing values into the rating matrix by (2) so that its density can be close to that of real data

Section 3.2, 3.3, and 3.4 discuss (1), (2), and (3), respectively, and section 3.5 describes the metrics used to evaluate the similarity of the generated synthetic rating matrix to the real data.

3.2. Extracting Statistics

The proposed method classifies items into "pop items (V_{pop})" and "niche items (V_{niche})" using the statistics that are extracted from real data (rating matrix). This paper defines pop items as follows, both of which are examined in the experiments.

pop1. A set of items with much ratings than others, which accounts for 80% of the total number of ratings in a dataset.

pop2. A set of items of which average ratings is higher than 3.5

Let X be a set of users. User $x_i \in X$ is regarded as either pop-item-like users, niche-item-like users, or others based on the ratio of pop items rated by x_i ($user_pop_rate_i$). Eq. (3.2.1), (3.2.2), and (3.2.3) respectively define the condition of pop-item-like users (X^{pop}), niche-item-like users (X^{niche}), and otherwise (X^{other}), in which t_1 and t_2 are predefined thresholds ($0 \leq t_1 \leq t_2 \leq 1$).

$$0 \leq user_pop_rate_i \leq t_1 \Leftrightarrow x_i \in X^{niche} \quad (3.2.1)$$

$$t_1 < user_pop_rate_i < t_2 \Leftrightarrow x_i \in X^{other} \quad (3.2.2)$$

$$t_2 \leq user_pop_rate_i \leq 1 \Leftrightarrow x_i \in X^{pop} \quad (3.2.3)$$

2. https://www.amazon.co.jp
3. https Internet services such as ://www.netflix.com/jp
4. http://yann.lecun.com/exdb/mnist/

3.3. Rating Generation Function Based on Rational Behavior

Rating Generation Function is used to generate a rating matrix without missing values. It assumes user's rational behavior such as "If a user has a preference for an item type, the function returns a high rating for items matching the preference, and a low rating for items not matching the preference."

Let R be the range of ratings and R_{high}, R_{low}, and $R_{random}(=R)$ be the sets of high, low, and random ratings, respectively. And let a_{ir} be the event that user $x_i \in X$ gives a rating $r \in R_t (t \in \{hight, low, random\})$. The Rating Generation Function returns a rating given by x_i based on the sampling with replacement events a_{ir} from Ω_t.

$$\Omega_t = \{a_{ir} | x_i \in X, r \in R_t\}. \tag{3.3.1}$$

That is, the probability of rating a_{ir} returned by the Rating Generation Function is denoted as $P(a_{ir}|\Omega_t)$.

$$P(a_{ir}|\Omega_t) = \frac{user_count_{ir}}{\sum_{k \in R_t} user_count_{ik}}, \tag{3.3.2}$$

where $user_count_{ir}$ is the frequency that x_i gave a rating r to items in real data. The rating for an item v_j by a user x_i is determined based on the following rules, the conditions of which depend on the types of users and items.

- $x_i \in X^{pop} \wedge v_j \in V^{pop} \Rightarrow P(a_{ir}) = P(a_{ir}|\Omega_{high})$
- $x_i \notin X^{pop} \wedge v_j \in V^{pop} \Rightarrow P(a_{ir}) = P(a_{ir}|\Omega_{low})$
- $x_i \in X^{niche} \wedge v_j \in V^{niche} \Rightarrow P(a_{ir}) = P(a_{ir}|\Omega_{high})$
- $x_i \notin X^{niche} \wedge v_j \in V^{niche} \Rightarrow P(a_{ir}) = P(a_{ir}|\Omega_{low})$
- $x_i \in X^{other} \Rightarrow P(a_{ir}) = P(a_{ir}|\Omega_{random})$

R, R_{high} and R_{low} will be explained in section 4.3.

3.4. Missing Value Generation Function for Density Adjustment

To adjust the density of the rating matrix generated with the method described in section 3.3, Missing Value Generation Function is applied to generate missing values. Its input is the generated rating matrix in Sec. 3.3 and the output is a rating matrix with missing values (M_g), of which density is close to that of real data (M_r). This paper proposes the following three functions, which are compared each other in the experiment. Here, a cell in a rating matrix that corresponds to a user x_i and an item v_j is denoted as c_{ij} and n_i is the number of $x_i's$ unrated items in M_r.

miss1. M_g has exactly the same missing values as M_r: c_{ij} is a missing value in $M_g \Leftrightarrow c_{ij}$ is a missing value in M_r

miss2. For each user x_i, set a missing value to c_{ij} in M_g with the probability $\frac{1}{n_i}$

miss3. For each user x_i, randomly select n_i cells from i-th row in M_r and set missing values to those cells

Table 1. Summary of a dataset used in the experiment

#User	#Item	#Rating	Average rating	sparsity
941	1,682	99,942	3.529	0.0632

3.5. Metrics

Kullback-Leibler divergence (KL divergence) and Jensen-Shannon divergence (JS divergence) are used as metrics to calculate the similarity of the distribution of several statistics such as item's average rating and precision between real and synthetic data.

KL divergence is a measure of the difference between two probability distributions. For two discrete probability distributions of a random variable x, Q, Q', KL divergence $D_{KL}(Q|||Q')$ of Q to Q' is defined as Eq. (3.5.1).

$$D_{KL}(Q||Q') = \sum_x Q(x) \log \frac{Q(x)}{Q'(x)}. \tag{3.5.1}$$

KL divergence is asymmetric: $D_{KL}(Q||Q') \neq D_{KL}(Q'||Q)$ $(Q \neq Q')$ and diverges to infinity if there exists an x where $Q'(x) = 0$. Eq. (3.5.2) defines JS divergence $D_{JS}(Q||Q')$ of Q to Q', which is a symmetric version of the KL divergence.

$$D_{JS}(Q||Q') = \frac{1}{2}(D_{KL}(Q||M) + D_{KL}(Q'||M)), \tag{3.5.2}$$

where $M = \frac{1}{2}(Q + Q')$.

4. Experiment

4.1. Data Set

Synthetic rating matrix is generated based on the statistics of Movielens 100k dataset. The dataset contains 943 registered users and 100,000 ratings on a 5-point scale for 1,682 movies. In the experiment, the dataset excluding users who only gave negative (2 or less) or positive (4 or more) ratings is used, of which summary is shown in Table 1.

4.2. Distribution of The Statistic Being Compared

The Statistics of which distribution are compared in the experiment are shown in Table 2. Predicted ratings are calculated with user-based collaborative filtering [1]. Precision [10] and recall [10] for a user x_i are calculated as Eq. (4.2.1), (4.2.2), respectively. In the equations, V_i^+ and L_i are the set of items to which x_i gives $r \in \{4,5\}$ in the real data and the generated synthetic data, respectively.

$$Precision_i = \frac{|V_i^+ \cap L_i|}{|L_i|}, \tag{4.2.1}$$

$$Recall_i = \frac{|V_i^+ \cap L_i|}{|V_i^+|}. \tag{4.2.2}$$

4.3. Experiment Summary

In the experiment, Rating Generation Function in section 3.3 is defined using by $R = \{1,2,3,4,5\}$ and we tested two different settings of R_{high} and R_{low} such as follows.

gen1. $R_{low} = \{1,2,3\}$, $R_{high} = \{3,4,5\}$

gen2. $R_{low} = \{1,2,3\}$, $R_{high} = \{4,5\}$

To evaluate the validity of the employed rational behavior, Rating Generation Function based on regular but meaningless behavior is defined as follows.

regular. Given x_i and v_j, return $((r_i + j - 2) \bmod 5) + 1$. r_i is selected from R with a uniform probability.

For example, when r_i is 3, ratings for v_1, v_2, v_3, v_4, v_5 are 3,4,5,1,2, respectively.

As described in Sec. 3, the proposed method generates a synthetic rating matrix by combining the definition of pop item, Rating Generation Function, and Missing Value Generation Function. In the experiment, the combination used for generating a rating matrix is referred to as x-y-z.

x. Definition of pop item: r (regular), 1 (pop1), 2 (pop2)

y. Rating Generation Function: r (regular), 1 (gen1), 2 (gen2)

z. Missing Value Function: 1 (miss1), 2 (miss2), 3 (miss3)

As the definition of pop item doesn't affect the regular (Rating Generation Function), we use the notation 'r-r-z' when using the regular to unify the notation. In the experiments, a synthetic rating matrix is generated with the following procedure as follows.

Step.1 Obtain the statistics of the real data required for generating a synthetic rating matrix.

Step.2 Repeat steps 3 to 8 with all combinations of x, y, z.

Step.3 Classify items into pop and niche items in accordance with x.

Step.4 Repeat steps 5 to 8 while changing t_1 and t_2 used in step 5 with 0.05 increments within the following ranges.

Step.5 Classify users into X^{pop}, X^{niche}, X^{other}.

- x = pop1 $\Rightarrow 0.3 \leq t_1 \leq t_2 \leq 1$
- x = pop2 $\Rightarrow 0 \leq t_1 \leq t_2 \leq 1$

Step.6 Generate a rating matrix with no missing values with Rating Generation Function y.

Step.7 Generate a synthetic rating matrix with Missing Value Function z.

Step.8 Calculate the similarity between the real data and the synthetic data in terms of statistics shown in Table 2 by KL divergence and JS divergence.

Table 2. Statistics used in the experiment

Statistics	Notation	Min	Max	Class width
Rating	*value*	1	5	1
Average rating of item	*item_mean*	1	5	0.5
Average rating of user	*user_mean*	1	5	0.5
User-based pearson correlation	*pearson*	-1	1	0.1
Precision	*precision*	0	1	0.1
Recall	*recall*	0	1	0.1

5. Result

5.1. Comparison among proposed methods

For each statistic, methods and thresholds (t_1, t_2) with which the synthetic rating matrix has the closest distribution to the real data were determined. Table 3 and Table 4 show the results when tie similarity of distribution is calculated based on KL divergence and JS divergence, respectively. The tables also show the similarity of distribution between real data and the synthetic rating matrix generated with r-r-z.

It can be seen that the generated synthetic rating matrix became similar to the real data when pop2, gen1, and miss1 were used. It is also found that for the distribution of *recall*, the synthetic rating matrix using gene2 was the most similar to the real data.

As for the thresholds t_1 and t_2, the distributions of *precision* and *recall* are more similar to the real data when much niche-item-like users (X^{niche}) exist compared to other statistics. However, in the cases of other statistics, it can be seen that the synthetic rating matrix get to be similar to the real data when the number of niche-item-like users (X^{niche}) is too small.

Table 3. Similarity of the distribution between real and synthetic data (KL). - denotes it cannot be calculated.

Distribution	Method	t_1	t_2	KL	r-r-1	r-r-2	r-r-3
value	2-1-1	0.05	0.90	5.0×10^{-6}	0.141	0.143	0.142
item_mean	2-1-1	0.05	0.55	0.056	-	-	-
user_mean	2-1-1	0.00	0.70	0.001	0.317	-	-
pearson	2-1-1	0.00	0.50	0.007	-	1.22	1.20
precision	2-1-1	0.30	0.90	0.043	-	-	-
recall	2-2-3	0.15	0.70	0.156	-	-	-

Table 4. Similarity of the distribution between real and synthetic data (JS)

Distribution	Method	t_1	t_2	JS	r-r-1	r-r-2	r-r-3
value	2-1-1	0.05	0.90	1.0×10^{-6}	0.037	0.038	0.038
item_mean	2-1-1	0.05	0.55	0.013	0.069	0.221	0.218
user_mean	2-1-1	0.00	0.70	2.89×10^{-4}	0.283	0.277	0.272
pearson	2-1-1	0.00	0.50	0.002	0.200	0.204	0.205
precision	2-1-1	0.30	0.90	0.011	0.196	0.302	0.302
recall	2-2-2	0.05	0.80	0.039	0.266	0.278	0.260

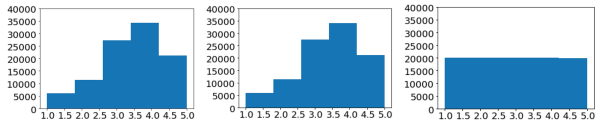
5.2. Comparison between proposed and comparative methods

Figures 2 to 7 compare the distributions of each statistic between the real data and synthetic rating matrix. Each figure contains the distribution of the real data (a), that

of the synthetic rating matrix closest to the real data in terms of JS-divergence by the proposed method (b), and that closest to the real data by r-r-z (c).

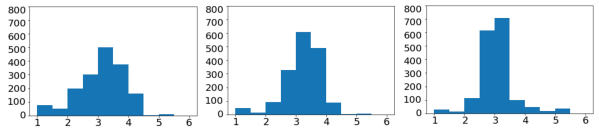
From (a) and (b) of Figures 2 to 6, It can be seen that the distribution of the synthetic rating matrix generated by the proposed method is similar to the real data. However, Figure 7 shows many users tend to have extremely low (0) or high (1.0) recall in the case of the proposed method. As the reason for that, Table 4 shows that the distribution of only *recall* is the most similar to that of the real data when miss2 is used, unlike other statistics. While miss1 and miss3 always generate the same number of missing values as the real data for each user, miss2 does not guarantee it: it is supposed that some users have fewer positive ratings than in the real data. As a result, the recall of those who have too few positive ratings tends to be 0 or 1.0.

Tables 3 and 4 show that the similarity of the synthetic data generated by regular is lower than that of the proposed method. From (a) and (c) in Figures 2 to 7, it can be seen that the distributions are quite different. These results show that the rational behavior assumed in the proposed method reflects the real user behavior to some extent.



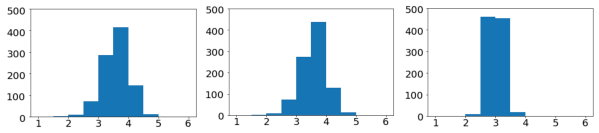
(a) Real data (b) Proposed (c) Compared

Fig. 2. Distribution of *value*. horizontal axis: value, vertical axis: number of ratings. Compared: regular but meaningless behavior (r-r-z).



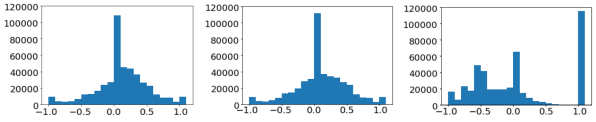
(a) Real data (b) Proposed (c) Compared

Fig. 3. Distribution of *item_mean* horizontal axis: item_mean, vertical axis: number of items. Compared: regular but meaningless behavior (r-r-z).



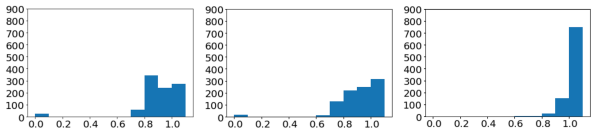
(a) Real Data (b) Proposed (c) Compared

Fig. 4. Distribution of *user_mean*. horizontal axis: user_mean, vertical axis: number of users. Compared: regular but meaningless behavior (r-r-z).



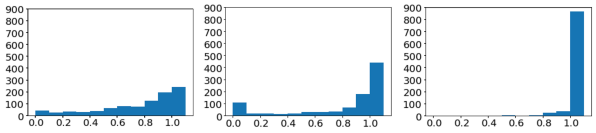
(a) Real data (b) Proposed (c) Compared

Fig. 5. Distribution of *pearson*. horizontal axis: pearson correlation coefficient, vertical axis: number of user pairs. Compared: regular but meaningless behavior (r-r-z).



(a) Real data (b) Proposed (c) Compared

Fig. 6. Distribution of *precision*. horizontal axis: precision, vertical axis: number of users. Compared: regular but meaningless behavior (r-r-z).



(a) Real data (b) Proposed (c) Compared

Fig. 7. Distribution of *recall*. horizontal axis: recall, vertical axis: number of users. Compared: regular but meaningless behavior (r-r-z).

6. Conclusion

This paper proposed a method for generating synthetic rating matrix by assuming the rational behavior of users.

In the proposed method, definition of pop items, Rating Generation Functions, and Missing Value Generation Functions are designed based on the rational behavior that "If a user has a preference for an item type (pop/niche), they give a high rating for items that match their preference and a low rating for others."

Using Movielens 100k dataset, an experiment was conducted to verify whether the proposed method can generate synthetic rating matrix similar to the real data. Experimental results show that the combination of pop2, gen1, and miss1 was effective for generating synthetic rating matrix similar to the real data. For the statistics other than precision and recall, it was confirmed that the synthetic rating matrix get to be similar to the real data when the number of niche-item-like users is too small.

To verify the validity of the assumed users' rational behaviors, the synthetic rating matrix was also generated based on regular but meaningless behavior. It was confirmed that the generated synthetic rating matrix was not similar to the real data.

In the future, clustering analysis will be applied to com-

pare user groups between the real data and synthetic rating matrix for more detailed analysis such as similarity of user characteristics. As for the potential application of the proposed method, it is expected to develop a method for detecting a cherry-picking attack by considering users who maliciously give ratings. It is also interesting to develop a recommendation system that predicts ratings for a real user based on the synthetic rating matrix.

7. Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers 19K22896, 21H03553, 22H03698.

References:

[1] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews" Proc. of the 1994 ACM Conference on Computer Supported Cooperative Work, pp.175-186, 1994.

[2] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods, and Metrics for Cold-start Recommendations" Proc. of the 25th ACM SIGIR Conference on Research, and Development in Information Retrieval, pp.253-260, 2002.

[3] S. Lee, J. Yang, and S. Y. Park, "Discovery of Hidden Similarity on Collaborative Filtering to Overcome Sparsity Problem" Proc. of the 7th International Conference on Discovery Science, pp. 396-402, 2004.

[4] A. Ratner, C. D. Sa, S. Wu, D. Selsam, and C. Ré, "Data Programming: Creating Large Training Sets, Quickly" Advances in Neural Information Processing Systems, Vol.30, pp.3574-3582, 2016.

[5] M. Frid-Adar, I. Diamanta, E. Klang, M. Amitaib, J. Goldberger, and H.Greenspana, "GAN-Based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification" Neurocomputing, Vol.321, pp.321-331, 2018.

[6] C. Anderson, "The Long Tail: Why the Future of Business is Selling Less of More" Hachette, Inc, 2006.

[7] O. Celma, and P.Cano, "From Hits or Niche?: Or How Popular Artists Can Biad Music Recommendation and Discovery" Proc. of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition, No.5, pp.1-5, 2008.

[8] M. Zhang, and N. Hurley, "Niche Product Retrieval in Top-N Recommendation" IEEE/WICA/ACM International Conference on on Web Intelligence, Vol.1, pp.74-81, 2010.

[9] S. Yoshida, T. Takagi, "Proposal of Recommender System Removed Popularity Bias by Using Information Gain," Transactions of the Japanese Society for Artificial Intelligence, Vol.30, No.5, pp.647-657, 2015 (in Japanese)

[10] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating Collaborative Filtering Recommender Systems" Proc. of ACM Transactions on Information and System Security, pp.5-53, 2004.