Rating and Comments Mining Using TF-IDF and SO-PMI for Improved Priority Ratings

Jinah Kim[1] and Nammee Moon[2*]
[1] Department of Computer Engineering, Hoseo University
Asan-si, Republic of Korea
[e-mail: jina9406@gmail.com]
[2] Division of Computer Information Engineering, Hoseo University
Asan-si, Republic of Korea
[e-mail: nammee.moon@gmail.com]
*Corresponding author: Nammee Moon

Received February 11, 2019; revised April 20, 2019; accepted June 10, 2019; published November 30, 2019

Abstract

Data mining technology is frequently used in identifying the intention of users over a variety of information contexts. Since relevant terms are mainly hidden in text data, it is necessary to extract them. Quantification is required in order to interpret user preference in association with other structured data. This paper proposes rating and comments mining to identify user priority and obtain improved ratings. Structured data (location and rating) and unstructured data (comments) are collected and priority is derived by analyzing statistics and employing TF-IDF. In addition, the improved ratings are generated by applying priority categories based on materialized ratings through Sentiment-Oriented Point-wise Mutual Information (SO-PMI)-based emotion analysis. In this paper, an experiment was carried out by collecting ratings and comments on “place” and by applying them. We confirmed that the proposed mining method is 1.2 times better than the conventional methods that do not reflect priorities and that the performance is improved to almost 2 times when the number to be predicted is small.

Keywords: Data mining, Text Mining, Comments Mining, TF-IDF, SO-PMI

1. Introduction

The development of the Internet and the widespread use of mobile devices have prompted the continuous production of various contents and corresponding information. Extracting and analyzing meaningful data from among these different forms of information has become a significant field of research. Studies using a variety of data mining techniques—such as Classification, Association Rule, and Clustering—are currently being conducted [1, 2]. Although they primarily used purchase data, more diverse types of data such as location data (GPS, Wifi, and Bluetooth) and Social Network Service (SNS) data (likable information and data relating to social networks), sensor data (eye tracking and temperature data) are also available [3, 4, 5]. Moreover, they are widely adopted in diverse field such as business, genetics, transportation, and distribution. Yet they have been perhaps most widely adopted in marketing [6, 7].

The purpose of utilizing data mining in marketing is to identify the users' intentions and make predictions and recommendations based on those intentions. Banks or financial institutions can propose appropriate financial products by analyzing the transaction history or account data of customers. Stock market can be predicted by combining social data such as SNS or news [8, 9]. Stores can identify purchase patterns by analyzing the purchase history of their customers. Based on the results of their analysis, retailers can profitably change the location of product displays or otherwise manage their inventory [10, 11].

In order to improve user satisfaction, data mining methods have been extended to include personal consulting based on preference. There are two approaches to this. One is to improve the quality of data through various data combinations and data filtering. The other one is to use data of other users with similar preferences. Many studies have been conducted either to improve similarity calculations or else to find the most similar user types by using weights and applying a variety of novel variables [12, 13].

Most of the data mining data types have quantitative and explicit characteristics and are easier to analyze than qualitative and unstructured data. Yet these types of data cannot be said to present better analytic possibilities when it comes to determining user preference. If user A and user B give the same five ratings when purchasing a mobile phone, and if user A focuses on design, while user B focuses on practical use, it is impossible to identify the priorities of each user in spite of the similar ratings provided by each. Therefore, user priority needs to be figured out by analyzing qualitative data. It is also necessary to convert it into quantitative and explicit forms for easy analysis.

Text mining is frequently used for this purpose. It is an analysis technique that discovers hidden meaningful information from unstructured text data. The analytic tools of text mining are based on Natural Language Processing. Studies on this subject have been conducted for various purposes by collecting and analyzing data such as review, news, and SNS [14, 15]. TF-IDF (Term Frequency-Inverse Document Frequency) is frequently used among other text mining techniques. TF is a frequency that presents how many specific words appear in the text. As the TF value of a given term gets bigger, it is more likely that the term is important. However, IDF needs to be calculated also, because there can be words that are not important but nevertheless frequent such as definite articles, indefinite articles or prepositions. Although there are different ways to calculate IDF, one of the most frequently adopted methods consists of dividing the entire number of documents by the number of documents containing the

relevant words and logging the result. This technique is widely employed in calculating the similarity among documents and mainly used as a weight [16].

Meanwhile, opinion mining is another example of emotion analysis that judges the polarity of text data among text mining. Since text analysis can identify user preference data and recommend expected preferences among users, research on recommendation services based on emotion analysis has been widely conducted [17, 18]. The technique of judging positivity and negativity includes Point-wise Mutual Information (PMI) analysis, which calculates the possibility that a given reference word and another word will both be present in a certain document, and has been developed as a method based on the theory of probability. With PMI, use and calculations are simpler than with other techniques, and exact results are predictable [19]. In addition, SO-PMI is another method that distinguishes the polarity with the preset positive word group and negative word group. It is widely used in many studies by providing an answer to the problem that analyzed results significantly vary according to the selection of reference words when analysis is performed with PMI [20].

Therefore, it is possible to identify the user intention from unstructured data and quantify through text mining. This paper proposes the method of Rating and Comments Mining (RCM) to identify user priority from comments and to obtain improved ratings based on them. RCM was applied in “place,” which means that the data was classified into structured data (location and rating) and unstructured data (comments) and then collected for analysis. Location data identifies the priority of preferential regions through statistical analysis of places that users have visited. Comments are used to derive the priority category and grasp the priority order for each category. The priority item varies depending on the type of place. In the case of a restaurant, the priority item may be a “taste” or an “atmosphere,” and in the case of a clothing store, it may be a “style,” a “material,” and a “price”. Priority categories are derived by constructing a dictionary of synonyms in advance, and priorities are identified through TF-IDF analysis. In addition, rating is more materialized than previously collected ratings with SO-PMI-based opinion mining by using rating and comments to improve the quality of information. And improved rating is predicted by calculating rating as well as the similarity among users on priority by region and category.

The structure of this paper is as follows. Section 2 describes RCM and Section 3 gives an explanation of the method proposed for RCM. Section 4 explains an experiment and its results. Lastly, this study and its implications are discussed in Section 5.

2. Rating and Comments Mining (RCM)

RCM is for Online to Offline (O2O) system that recommends consumer behaviors in an offline setting by using ratings and comments data created online in response to offline consumption. The goal is to obtain a user's priority by combining the structured data and the unstructured data, and to obtain a new and highly predictable rating. The most notable feature of the proposed method is that it reflects different priorities among users. It calculates the rate of importance according to priority categories among users and reflects this in predicting improved ratings by finding users with similar priorities by calculating the similarity among users.

In this paper, location, rating, and comments on “place” were collected and applied in RCM. Since the priority categories considered by users regarding place vary according to the characteristics of a given place, we have limited ourselves to the context of the “restaurant”. The process of RCM consists of constructing and implementing a Data Collector, a User Priority Extractor, a Rating Calculator, and a Predictor, as shown in **Fig. 1**.

The Data Collector collects information including place name, region, and review information by using web crawling on selected regions. Review information including reviewer, rating, and comment is collected. Filtering for eliminating redundant data is processed. Additionally, preprocessing is performed when it is needed for each process—such as morpheme analysis or Part-Of-Speech tagging for extracting priority or calculating materialized ratings.

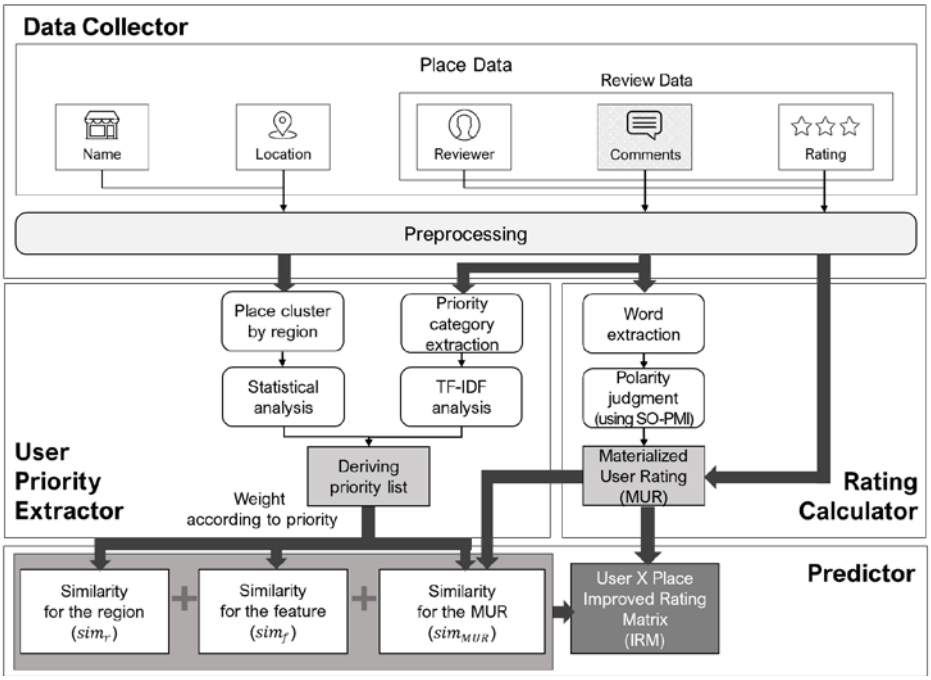

Fig. 1. Process of the Rating and Comments Mining (RCM)

The User Priority Extractor is a process of identifying priority among users by utilizing preprocessed location and comments data. The statistical analysis of the number of visits to the region by the user and the TF-IDF analysis based on the dictionary of synonym of the priority are carried out. Eventually the importance of the priority category of users with reference to distinct places can be acquired and they are listed.

The Rating Calculator is a process of obtaining the materialized ratings by using rating and comments. The materialized rating is obtained by integrating the result of judging the polarity based on SO-PMI, and the existing rating.

Lastly, the Predictor creates the improved rating matrix based on the priority list which is derived from the User Priority Extractor and the materialized user rating, which has been previously derived from the Rating Calculator. This reflects user-specific priorities, and the prediction is carried out.

3. Component of the RCM

3.1 User Priority Extractor

The User Priority Extractor is divided into two types, one for structured data (location) and one for unstructured data (comments). The User Priority Extractor analyzes the importance of user priority and combines the resulting data to extract the priority list. The process for this is as follows.

First, we undertake the analysis process for the location. This is done by statistical analysis that extracts the region through the location data (GPS) of the place that the user has visited before and obtains the frequency of visit by region. This analysis intends to identify the inclination of users to select specific places, and to determine whether they visit a specific region or several regions. In this way it is possible to specify the priority of regions for the user's preference of place.

Second, we undertake the analysis process for the comments. The analysis consists of three steps: selecting priority categories, constructing a dictionary of synonyms, and analyzing by means of TF-IDF. TF analysis reviewing “restaurant” was conducted as shown in **Table 1** to select priority categories. In addition, we examined the existing analysis cases of “restaurant” [21, 22, 23, 24, 25]. As shown in **Table 2**, the main focus was on Taste, Price, Atmosphere, and Service. Other factors (Etc) included Menu, Design, Category, Communication with Employees, and Cleanliness. As a result, we concluded that the priority categories were Taste, Price, Atmosphere, and Service.

Table 1. TF analysis result for restaurant reviews

Rank	Word	Frequency
1	Taste	1,191
2	Price	528
3	Food	441
4	Delicious	379
5	Atmosphere	216
6	Restaurant	191
7	Best	187
8	Meat	162
9	Menu	161
10	Service	142

Table 2. Analysis result of evaluation items of existing research on restaurant reviews

Existing Research	Taste	Price	Atmosphere	Service	Etc
Pantelidis, I. S. (2010) [21]	O	O	O	O	Menu, Design
urafsky, D., Chahuneau, V., Routledge, B. R., & Smith, N. A. (2014) [22]	O	O			Category

Chaves, M. S., Laurel, A., Sacramento, N., & Pedron, C. D. (2014) [23]	O	O		O	Communication, Menu, Cleanliness
Yan, X., Wang, J., & Chau, M. (2015) [24]	O	O	O	O	
Mubarok, M. S., Adiwijaya, & Aldhi, M. D. (2017) [25]	O	O	O	O	

Next, we set about building a Dictionary of Synonyms (DS). The DS is constructed based on the extracted priority categories as shown in **Table 3**. Based on the structure of the standard Korean language dictionary, synonyms for the priority category are searched for and classified according to the priority category. At this time, there may be no appropriate words referred to in the review data. The words with a high frequency of reference are sorted from the review data and, if the words are included in the priority category, the DS is constructed through the process of clustering.

Table 3. Example of Dictionary of Synonym (DS)

Priority category	Synonym
Taste	meal, food, delicious, plain, clean, not oily, light …
Price	value, amount, affordable, reasonable, expensive, cheap …
Atmosphere	mood, luxurious, interior, style, design …
Service	amount, employee, kindness, goodness, waiting, parking lot …

Next, the TF-IDF analysis proceeds. $TF-IDF_{s,c}$ is the same as equation (1) if synonym s in the DS appears for each priority category in all comments c of the user. The combined TF-IDF value for all priority categories is the Final User Priority feature and it is listed to derive a priority list. In equation (1), $tf_{s,c}$ is the frequency of synonym s in comments c. df_s is the number of comments containing synonym s. N is the total number of comments.

$$TF-IDF_{s,c} = tf_{s,c} \times log\left(\frac{N}{df_s}\right) \quad (1)$$

3.2 Rating Calculator

User's review data (rating, comments) are used to calculate materialized ratings on places. Since the data is collected through web crawling, a complicated preprocessing is undertaken to filter out insignificant data such as special letters, URL, html tag, etc. Then we perform sentence separation, morpheme analysis, word extraction, polarity judgment, and comments scoring, in that order, as shown in **Fig. 2**.

First, the sentences are separated into distinct units to conduct SO-PMI-based emotion analysis on the collected comments. Sentences were divided within the text based on the presence of a period. Then, by performing morpheme analysis, only nouns, adjectives, and verbs are extracted from each sentence.

Then we perform the polarity judgment, which is the process of exploring whether the extracted words are positive or negative. The previously constructed Emotion Dictionary (ED) is used to make the polarity judgments. This dictionary is a collection of positive and negative words that are operative within the priority categories. In other words, if examples of "taste" are being considered, positive word collections can include "delicious" or "good," while negative word collections can include "salty" or "not good". When constructing the ED, it is vital to judge the polarity properly and include widely employed words. In the present paper we will be analyzing reviews written in Korean, and we will be making use of the KOSAC (KOrean Sentiment Analysis Corpus), which we have adopted as our standard ED. KOSAC is frequently updated with many positive and negative words. Furthermore, words were converted to their original lexical entries so that coined words like abbreviations used on the Internet could be included. Then the polarity was distinguished by calculating the SO-PMI as shown in equation (3). Equation (3) is based on equation (2). Here, $P(\mathrm{w}_1)$ is the probability of a word w_1 appearing in comments, and $P(\mathrm{w}_1, \mathrm{w}_2)$ means the probability that words w_1 and w_2 appear simultaneously in comments.

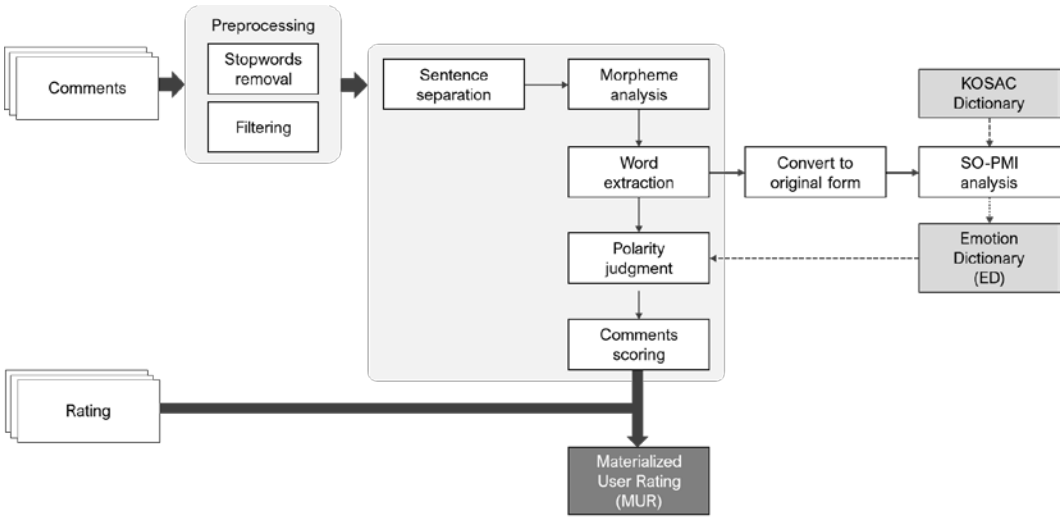

Fig. 2. Process of the Rating Calculator

As shown in equation (3), *w* is a word to identify the polarity by being extracted in the comments. *PW* is a collection of positive words and *NW* is a collection of negative words. If the result of equation (3) is a negative number, the word is judged to be negative, and if result of equation (3) is a positive number, the word is judged to be positive. The word, thus analyzed, is added to the ED according to the result. **Table 4** shows the constructed ED.

$$\mathrm{PMI}(\mathrm{w}_1, \mathrm{w}_2) = log \frac{P(\mathrm{w}_1, \mathrm{w}_2)}{P(\mathrm{w}_1)P(\mathrm{w}_2)} \tag{2}$$

$$\mathrm{SO-PMI}(w) = \sum_{pw \in PW} PMI(w, pw) - \sum_{nw \in NW} PMI(w, nw) \tag{3}$$

Table 4. Example of Emotion Dictionary (ED)

Positive			**Negative**		
recommend	good	okay	not recommend	not good	bad
best	delicious	neat	terrible	disgusting	pressured
nice	cheap	kind	awful	expensive	unkindness
free	courteous	fast	costly	not fine	lag
luxurious	high quality	quiet	poor	low quality	noisy

The final step in conducting emotion analysis is to judge the total polarity of the comments. The polarity judgment will be a value between 1 and 9, and the default value is 5. If the word in comments is positive, the value of the word is the total number of words in the comments divided by 8, because the polarity judgment has a value between 1 and 9. On the other hand, if the word is negative, it is multiplied by -1 in the previous step.

When the emotion analysis is completed for one comment, all values are summed up. Next, the process of combining it and the existing rating is performed. The existing rating is set as the default rating. Based on the decimal point, the former is existing rating and the latter is result value of calculation on comments. In this way, Materialized User Rating (MUR) on place is obtained as matrix form.

3.3 Predictor

The Predictor will make predictions about places that have not been visited. The user-place preference matrix is obtained by first considering the similarities calculated from the results of the User Priority Extractor and then combining them with the MUR calculated by the Rating Calculator. The process of calculating the similarity among users proposed in this paper additionally reflects with greater accuracy the features and location of places visited by users, in contrast to the traditional prediction method which calculates users similar to ratings. When users visit places, categories prioritized by users are generally different, and the region that can be visited varies according to user. Thus, it is necessary to reflect the feature and location of the place. In this paper, the similarity between users is obtained for the feature, location, and MUR of the place. The similarity for user u is calculated by using Pearson's correlation coefficients as shown in equation (4). The similarity value ranges from -1 to 1. As it gets closer to 1, it indicates that it is consistent with the similarity between users.

$$sim_u(a,\ b) = \frac{\sum_{u\in U}(R_{a,p}-\bar{R}_a)(R_{b,p}-\bar{R}_b)}{\sqrt{\sum_{u\in U}(R_{a,p}-\bar{R}_a)^2}\sqrt{\sum_{u\in U}(R_{b,p}-\bar{R}_b)^2}} \quad (4)$$

We calculate the similarity between users for the feature (sim_f) in order to distinguish those users whose priority is similar to the place feature. The similarity between users for the region (sim_r) is calculated in order to find those users who visit similar regions and obtain similarity within the set of visited regions. We calculate the similarity between users for the MUR (sim_{MUR}) in order to distinguish those users who have been given similar rating. After calculating the three similarity types among users, we derive the final similarity (sim_{final}) between users by combining all three of the above.

At this time, sim_f and sim_r are combined first; we then combine this result with sim_{MUR}. This first combining process reflects whether the user prioritizes the feature or the region. Because The more various regions of place the user visits, the more likely that the user will visit the desired place, regardless of region. In other words, it can be said that choice of feature is a more enduring and salient aspect of user behavior than choice of region. On the contrary, the main movement patterns of users are limited to specific regions. When they want to visit certain places, they are likely to select the one closest to the regions in which they regularly operate. Therefore, the weight W is calculated as a high frequency of visits to the total number of regions. The sim_{final} reflecting this is shown in equation (5). Finally, by multiplying MUR derived from the Rating Calculator by sim_{final}, a User-Place Improved Rating Matrix (IRM) is derived, and based on this, it is possible to predict the user about the few top places that have not visited.

$$sim_{final} = \frac{\{(Wsim_f + (1-W)sim_r) + sim_{MUR}\}}{2} \tag{5}$$

4. Experiment

In this study, the proposed system was applied to the place "Restaurant" in the region of Seoul, South Korea, in order to test the system's performance. Seoul-based restaurants and review information data provided by Google Map were collected by using web crawling. Google Map was selected because many people around the world use it and data associated with places are adequately accumulated and documented. **Fig. 3** shows the form of place information and review information in Google Map.

After preprocessing the data, an experiment was carried out with 4,751 reviews on 297 restaurants that 1,039 users had visited. This result was obtained by filtering restaurant reviews in such a way as to identify those people who had visited at least three among the 297 restaurants during a certain period of time. Results of the collected data are presented in **Fig. 4**.

Fig. 3. Form of place information and review information in Google Maps

Place name	ID	Rating	The number of reviews	Type	Region
꾸•••	ChlJhX	4.5	10	Tunisian	Jongro
거••••	CHIJrQK	3.1	11	General Restaurants	Yongsan
가••••	CHIJA5X	3.4	16	General Restaurants	Geumcheon
설•	ChlJyYyx	3.8	34	General Restaurants	Gangdong
더••••	ChlJ-xC	4.2	25	General Restaurants	Gangnam
다••••	ChlJ7R	4.3	31	French restaurant	Jongro

(a) Place data
(Place name, ID, Rating, The number of review, Type, Region)

Place name	Reviewer	Rating	Comments
미•••	Sang*** ****	4	가성비 적절
새•••• •••	Young *** ***	2	맛은 여느 새
ㅍ••••	박**	5	맛있다 명동에 또 생겨 사람...
석•••	M_ *** ****	4	맛있었습
이••• •••	M** *****	4	언제나 일정한 음식 퀄리티...
꿈*****	한**	5	직원들이친절하고음식맛...

(b) Review data
(Place name, Reviewer, Rating, Comments)

Fig. 4. Sample of the final collected data

Fig. 5 shows a graph representing the priority list derived by using the User Priority Extractor. The dotted line on the yellow background is the weight for a given location. The larger the range, the more likely the user prefers a particular region. Other priority categories represent user priority importance extracted from comments. We can see that taste is the most important category for most users.

Based on this result, the similarity among users can be derived by using the Predictor. The results have been represented as a Heatmap and are shown in **Fig. 6**. Within **Fig. 6**, (a) shows sim_r, (b) shows sim_f, (c) shows sim_{MUR}, and (d) shows sim_{final}. Here, the x-axis and y-axis are users. The higher the similarity is, the more the color becomes red; and the lower the similarity, the more the color becomes blue. The similarity to oneself is 1, so there is a red diagonal line for each similarity. Eventually, the User-Place IRM is derived by using similarity sim_{final} and MUR.

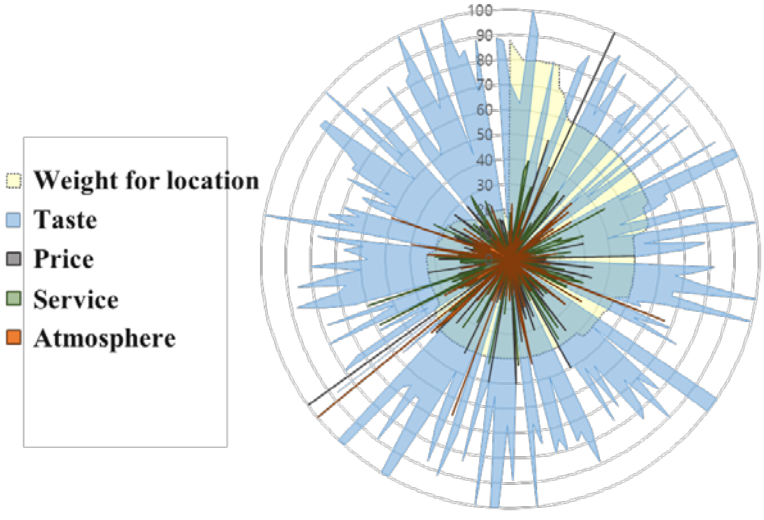

Fig. 5. Result of priority list by User Prioirity Extractor

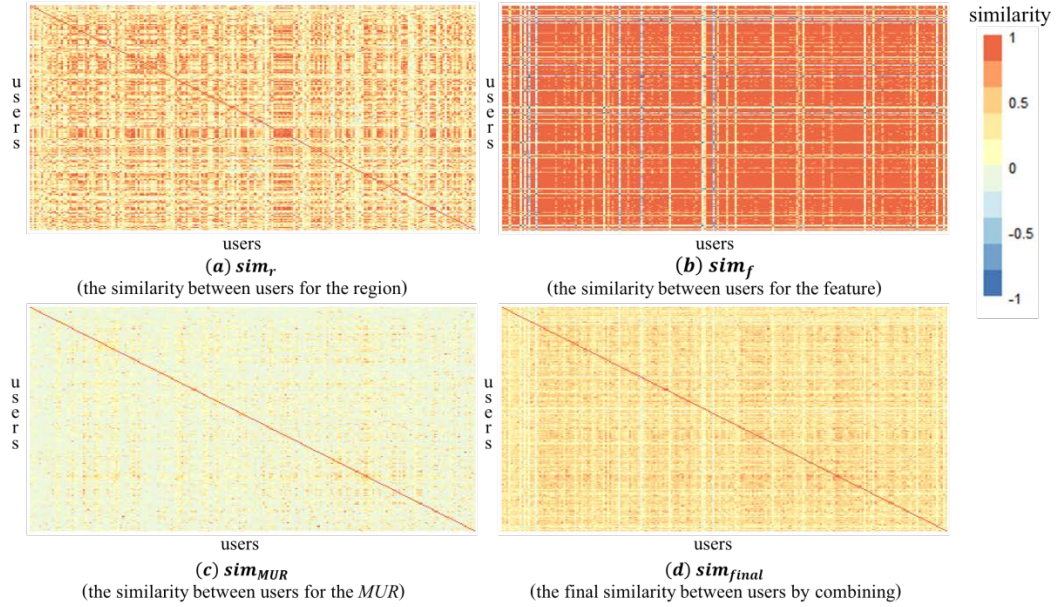

Fig. 6. Heatmap for each similarity between user

In this paper, 70% of data were set to training data and 30% as test data, conducting analysis and evaluating performance. Performance evaluation calculated precision and recall based on **Table 5**, thereby calculating F-measure values. Precision is the percentage of the restaurant that the user actually visited among the predicted restaurants, as shown in equation (6). Recall is the percentage of the restaurant that the predicted restaurant among the restaurants which is user actually visited, as shown in equation (7). F-measure is calculated as equation (8) with Precision and Recall.

Table 5. Descriptions of symbols

Actual' predicted	Predicted	Not predicted
Visit	a	b
Not visit	c	d

$$Precision = \frac{a}{a+c} \tag{6}$$

$$Recall = \frac{a}{a+b} \tag{7}$$

$$F-measure = \frac{2 \times Precision \times Recall}{Precision+Recall} \tag{8}$$

Experimental performance was verified by comparing the performance against the conventional method and the RCM. **Fig. 7** is a Recall-Precision Graph according to the number of prediction places (3, 5, 7, 9). According to the average result of comparison with the conventional method, performance is improved, as shown in **Table 6**. As the number of places to predict increased, there was no significant difference from the traditional predict method.

However, when the number of places to predict was small, the prediction accuracy was almost twice as high. In other words, we confirmed that the proposed RCM shows high performance when the number of predictions is small.

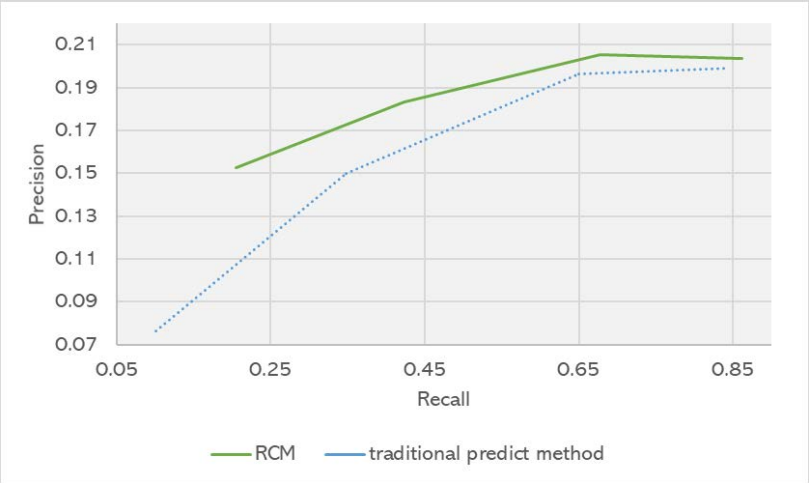

Fig. 7. Recall-Precision graph according to the number of prediction places

Table 6. An average result of performance comparison

Method	Precision	Recall	F-measure
Traditional predict method	0.1555	0.4852	0.2290
RCM	0.1863	0.5417	0.2676

5. Conclusion

In this paper, we propose Rating and Comments Mining (RCM) for improved priority rating. The RCM was composed of a Data Collector, a User Priority Extractor, a Rating Calculator, and a Predictor. The proposed method has value because it accurately identified user priority and reflected it differently according to users, contrary to the traditional recommendation method that calculates the similarity of rating among users.

We have systemized the priority by dividing it into feature and region of the place. The priority for the place's features were extracted by TF-IDF analysis based on the DS. We extracted the priorities of regions by statistical analysis of the frequency of visited places. Also, existing rating was materialized by SO-PMI-based emotion analysis of the comments. We obtain user similarity based on feature, region, and materialized user rating, and derive an improved rating for prediction by weighting on what is more important to the user in terms of features and region.

Ratings and comments data were collected from Google Map; we applied them into RCM. Its experimental performance was compared with the rating-based traditional prediction method. As a result, we confirmed that the RCM improves the accuracy of prediction and the recommended performance increases.

For the limitation of this study, it did not apply the exceptional process since it is difficult to identify user priority regarding data only with rating without any comments. Moreover, the proposed system focused on accuracy. Since there are many calculations, massive amounts of data—more than the amounts used in this study—can trigger low computational speed. Future

studies need to improve performance in terms of speed. Although applied our method only to "place" in this paper, it can be expanded to other fields. Therefore, we expect that it will be able to provide recommendation services to users more efficiently by automatically extracting priority categories and providing appropriate recommendations.

Acknowledgment

This work has supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. NRF-2017R1A2B4008886).

References

[1] Lee, O., and You, E. S., "Predictive clustering-based collaborative filtering technique for performance-stability of recommendation system," *Journal of Intelligence and Information Systems*, vol. 21, no. 1, pp. 119-142, 2015. Article (CrossRef Link).

[2] Amatriain, X., and Pujol, J. M., "Data mining methods for recommender systems," *Recommender systems handbook, Springer, Boston, MA*, pp. 227-262, 2015. Article (CrossRef Link).

[3] Yu, F., Che, N., Li, Z., Li, K. and Jiang, S., "Friend recommendation considering preference coverage in location-based social networks," in *Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Cham*, pp. 91-105, 2017. Article (CrossRef Link).

[4] Lee, S. and Moon, N., "Location recognition system using random forest," *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 4, pp. 1191-1196, 2018. Article (CrossRef Link).

[5] Song, H. and Moon, N., "A preference based recommendation system design through eye-tracking and social behavior analysis," *Advances in Computer Science and Ubiquitous Computing, Springer, Singapore*, pp. 1014-1019, 2017. Article (CrossRef Link).

[6] Han, H. and Park, S., "Traffic information service model considering personal driving trajectories," *Journal of Information Processing Systems*, vol. 13, no. 4, pp. 951-969, 2017. Article (CrossRef Link).

[7] De Maio, C., Fenza, G., Gallo, M., Loia, V. and Parente, M., "Social media marketing through time-aware collaborative filtering," *Concurrency and Computation: Practice and Experience*, vol. 30, no. 1, e4098, 2018. Article (CrossRef Link).

[8] Mitik, M., Korkmaz, O., Karagoz, P., Toroslu, I. H. and Yucel, F., "Data mining approach for direct marketing of banking products with profit/cost analysis," *The Review of Socionetwork Strategies*, vol. 11, no. 1, pp. 17-31, 2017. Article (CrossRef Link).

[9] Zhang, X., Zhang, Y., Wang, S., Yao, Y., Fang, B. and Philip, S. Y., "Improving stock market prediction via heterogeneous information fusion," *Knowledge-Based Systems*, vol. 143, pp. 236-247, 2018. Article (CrossRef Link).

[10] Kim, J. and Moon, N., "Design of consumer behavior analysis by region through reflecting social atmosphere based on SNS," *Advances in Computer Science and Ubiquitous Computing, Springer, Singapore*, pp. 1020-1025, 2017. Article (CrossRef Link).

[11] Huang, S. H., Tsai, C. Y. and Lo, C. C., "A multi-data mining approach for shelf space optimization: Considering customer behaviour," in *Proc. of 2014 11th International Conference on e-Business (ICE-B)*, pp. 89-95, 2014. Article (CrossRef Link).

[12] KG, S., Sadasivam, GS., "Modified heuristic similarity measure for personalization using collaborative filtering technique," *Appl. Math. Inf. Sci.[internet]*, vol. 11, no. 1, pp. 307-315, 2017. Article (CrossRef Link).

[13] Kim, S., Seo, Y. and Baik, D. K., "Tweet entity linking method based on user similarity for entity disambiguation," *Journal of Korean Institute of Information Scientists and Engineers (KIISE)*, vol. 43, no. 9, pp. 1043-1051, 2016. Article (CrossRef Link).

[14] Yun, Y., Hooshyar, D., Jo, J. and Lim, H., "Developing a hybrid collaborative filtering recommendation system with opinion mining on purchase review," *Journal of Information Science*, vol. 44, no. 3, pp. 331-344, 2018. Article (CrossRef Link).
[15] Ishanka, U. A. and Yukawa, T., "The prefiltering techniques in emotion based place recommendation derived by user reviews," *Applied Computational Intelligence and Soft Computing*, 2017 Article (CrossRef Link).
[16] Jeon, B., and Ahn, H., "A Collaborative Filtering System Combined with Users' Review Mining: Application to the Recommendation of Smartphone Apps," *Journal of Intelligence and Information Systems*, vol. 21, no. 2, pp. 1-18, 2015. Article (CrossRef Link).
[17] Rosa, R. L., Rodriguez, D. Z. and Bressan, G., "Music recommendation system based on user's sentiments extracted from social networks," *IEEE Transactions on Consumer Electronics*, vol. 61, no. 3, pp. 359-367, 2015. Article (CrossRef Link).
[18] Chao, August FY and Cheng-Yu Lai., "SNS opinion-based recommendation for eTourism: A Taipei restaurant example," in *Proc. of International Conference on Multidisciplinary Social Networks Research, Springer Berlin Heidelberg*, pp. 393-403, 2015. Article (CrossRef Link).
[19] Hung, Y. H., Ou, Y. Y., Kuan, T. W., Cheng, C. H., Wang, J. F. and Wu, J. S., "An emotional feedback system based on a regulation process model for happiness improvement," in *Proc. of 2014 IEEE International Conference on Orange Technologies (ICOT)*, pp. 205-208, 2014. Article (CrossRef Link).
[20] Choi, E. J., and Kim, D. K., "Location Recommendation Customize System Using Opinion Mining," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 21, no. 11, pp. 2043-2051, 2017. Article (CrossRef Link).
[21] Pantelidis, I. S., "Electronic meal experience: A content analysis of online restaurant comments," *Cornell Hospitality Quarterly*, vol. 51, no. 4, pp. 483-491, 2010. Article (CrossRef Link).
[22] Jurafsky, D., Chahuneau, V., Routledge, B. R. and Smith, N. A., "Narrative framing of consumer sentiment in online restaurant reviews," *First Monday*, vol. 19, no. 4, 2014. Article (CrossRef Link).
[23] Chaves, M. S., Laurel, A., Sacramento, N. and Pedron, C. D., "Fine-grained analysis of aspects, sentiments and types of attitudes in restaurant reviews," *Tourism & Management Studies*, vol. 10, no. 1, pp. 66-72, 2014. Article (CrossRef Link).
[24] Yan, X., Wang, J. and Chau, M., "Customer revisit intention to restaurants: Evidence from online reviews," *Information Systems Frontiers*, vol. 17, no. 3, pp. 645-657, 2015. Article (CrossRef Link).
[25] Mubarok, M. S., Adiwijaya and Aldhi, M. D., "Aspect-based sentiment analysis to review products using Naïve Bayes," in *Proc. of AIP Conference, AIP Publishing*, vol. 1867, no. 1, p. 020060, 2017. Article (CrossRef Link).

Jinah Kim received B.S. and M.S. degrees in School of Computer Science and Engineering from Hoseo University. She is currently pursuing the Ph.D. degree in Department of Computer Engineering with Hoseo University, Korea. Her research interests include Smart Service, Machine Learning, Big Data Processing and Analysis.

Nammee Moon received B.S., M.S. and Ph.D degree in School of Computer Science and Engineering from Ewha Womans University in 1985, 1987 and 1998. She served as an assistant professor at Ewha Womans University from 1999 to 2003, From 2003 to 2008: Professor of Digital Media, Graduate School of Seoul Venture Information. Since 2008, She is currently a professor of computer science at Hoseo University. She is current research interests include Social Learning, HCI and User Centric Data, Big data Processing and Analysis.