Disparity-Based Measurement on Participation Interest and Competition Recommendation Method

Kaixin Liu, Jianhui Chang *, Huiwen Ren, Hong Zhu
Institute of Mechanical Electronic and Information Engineering, China University of Mining and Technology Beijing, Beijing, China.

* Corresponding author. Tel.: +86 13683032708; email: HellyCla009@gmail.com
Manuscript submitted December 14, 2017; accepted February 10 2018.
doi: 10.17706/jsw.13.2.117-125

Abstract: Most of the popular recommendation algorithms are providing similar recommendations to users based on their ratings. However, in terms of competition, past records of user ratings do not directly translate to users' interests in new competitions. Competitions are challenges, and as such, users are unlikely to choose what they have registered before, but instead, prefer challenges that complement the scope of their present abilities. Thus measurements of a user's interest in competitions should be based on the differences, rather than similarities, in user's past registration data. In this paper, we propose an alternative recommendation algorithm that measures users' interests in competitions based on these differences. First, competition differences, such as registrations, stars and browsers records are modeled and calculated. Then, the peak values and the range of users' interests are attained through such differences. Finally, recommendations of competitions are made if they fall within the range radius. The proposed algorithm proves to be more effective and efficient than conventional recommendation algorithms due to its consideration of competition's features as well as the user's psychology.

Key words: Competition recommendation method, disparity measurement on participation interest, disparity

1. Introduction

With an increase in demand for self-competency improvement, users spend a lengthy amount of time selecting a competition that will allow users to grow and develop their skills. Thus, recommendation algorithms are widely and successfully utilized in music [1], movies [2] and other fields [3], [4], which helps save time for its users.

Balabanovia et al. classify recommendation algorithms into 3 categories: Content-Based Filtering(CBF), Collaborative Filtering (CF) and Hybrid Recommendation(HR) [5].

Content-Based Filtering recommends items to a user by analyzing the competitive features the user likes and then recommending an item which has similar features but is new to the user [6], [7]. This algorithm does not refer to other user's information, so there are subtle issues when it comes to data sparsity and new-item problem. However, in the field of competition, peoples' opinions are subjective and thus there lies an inconsistency in how users are matched with the appropriate competitions. With such inconsistency, it is difficult for tagging competitions in a way that will be suitable for every individual's preferences. Suppose we find a suitable tag for the competition and a user has rated it a "like", users' challenge psychology, the psychological drive for the challenge, would lead him/her to choose the one which will be new rather than

familiarity-based.

Adomavicius *et al*. [8] simplified the Collaborative Filtering algorithm to user's estimation of unknown items. Calculating the similarity between two items by other users' ratings, referring the user's ratings on other items to estimate his rating on an item, choose an item with the highest score to recommend [9]. This algorithm resolves the tag problem of Content-Based Filtering mentioned above, but still, has the following problems:

The new-item problem is of important concern in the field of competitions. User ratings are obtained at a slower rate compared to other field ratings such as movies and cuisines due to the extensive time duration from the initial encounter of the product to the final review of the product. Competitions aren't freely obtainable to be rated and could only be rated once the user has participated in it at a set time thus requiring a lengthy amount of time before a rating can be obtained.

Users rating do not objectively reflect the user's interests in a competition. Users' ratings are not only subjective reflections of the competition itself, but are also influenced by the awards, attitudes of the staff, the partiality of judges and many other factors that users' may find unsatisfactory with. Thus user ratings cannot accurately measure users' interests in a competition.

The Hybrid Recommendation algorithm is the hybrid of the two aforementioned algorithms, yet it still fails to solve the problems mentioned above [10], [11].

In order to solve these issues, we propose an alternative recommendation algorithm that can measure users' interests by quantifying the users' interests, finding the peak of users' interest and giving a radius of users' interests, which will effectively improve the accuracy of competition recommendation made.

2. The Measurement Method of Competition

Traditional recommendation algorithms provide recommendations based on similarity [12], [14], but the competition itself is a challenge and there is an inverse relationship between similarity of recommendations and the possibility of choosing such recommendations. Therefore, we propose the disparity as the measurement of the distance between two competitions and use it to measure user's interests in competitions. The disparity obtained can better reflect a user's challenge psychology better than similarity can.

2.1. The Disparity between Competitions

Most recommendation algorithms are usually measuring based on users' comments and stars [15]. In the competition field, it is difficult to collect comments for measurement due to the new-time problem that arises. Furthermore, users' comments alone cannot impartially reflect the competition, so any measurement through this alone will be inaccurate. Thus in the proposed algorithm, we will use users' comments data, registers data and browsers data as the measurements.

Define a set *X = {register, star, browse}*, $N_x(i)$ represents the number of users who has relationship with competition i in property x, $N_x(i \wedge \mathrm{j})$ represents the number of users who has relationships with both competition i and j in property x, $d_x(i,j)$ represents the disparity of competition i and j in property x.

$$d_x(i,j) = 1 - \frac{N_x(i \wedge \mathrm{j})}{N_x(i)+N_x(j)-N_x(i \wedge \mathrm{j})}, x \in X \tag{1}$$

The range of $d_x(i,j)$ is [0, 1], larger range indicates higher disparity.

Different properties have different performance when reflecting users' interests in competitions; we give each property a weight according to their performance. Define a set W = {0.5， 0.3， 0.2}, in which "register" is 0.5, "star" is 0.3 and "browse" is 0.2. Considering all of these three properties, the final disparity

between competition i and j is

$$D(i,j) = \sum_{k=1}^{3} \varpi_k d_x(i,j), \quad \varpi_k \in W, x_k \in X \tag{2}$$

In the formula, k represents the position of the set; The range of $D(i,j)$ is [0, 1], the bigger it is, the higher the disparity is.

In the calculation of disparity, user's register data has the highest impact, which makes the result more accurate; user's star data and browse data can be collected during a short period of time, which bypasses the new-item problem.

2.2. Relationship between User-Competition Disparity and User's Interests in Competition

The competition being recommended should have a specific disparity with what they have participated in before, we call it User-Competition Disparity and use it to measure user's interests in competitions.

Define a registered competition set of user u as $C_u = \{c_1, c_2, \ldots, c_n\}$, calculate the mean of disparities between the competition i which user hasn't registered and each element in C_u, the result is what we called "User-Competition Disparity ", marked as $\overline{D_u}(i)$

$$\overline{D_u}(i) = \frac{1}{n}\sum_{k=1}^{n} D(i, c_k) \tag{3}$$

In Competition field, the main reason of why competitions with high similarity are not likely to be chosen is due to the user's challenge psychology. However, if a competition's demands are beyond the user's ability to complete competently, they will not be chosen as well.

The relationship between User-Competition Disparity and user's interests is shown in Figure 1. Starting from the origin, an increase in User-Competition Disparity corresponds with an increase in user's interests until it reaches α, optimal user-competition disparity, where a user's interests are maximized. Any further disparity will lead to a negative association between the two variables.

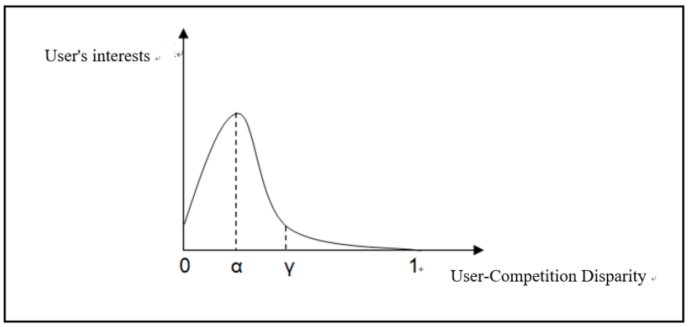

Fig. 1. Relationship between user-competition disparity and participation interest.

In Fig. 1, the α indicates the peak of user's interests, different users have different α's. The larger the value, the stronger the prediction of user's likelihood of accepting new things. In other words, users are likely to choose competitions that have high disparity with what they registered before. γ is the inflection point of user's interests and if User-Competition Disparity is larger than γ, it will indicate that the level of competition is beyond user's ability. Here the users' interests in it will begin to reach zero. User's interests are higher when the disparity is in the range of [0, γ], and it is approximately symmetrical about α.

2.3. User's Interests Peak and User's Best Disparity

According to Fig. 1, at α, "User's Best Disparity", the corresponding user's interest is at a maximum. It is related to user's ability to accept the things and whether the user is familiar with the competition that is recommended. Furthermore, a user's past competitive records can be utilized to provide data on these two variables, allowing for α to be obtained.

The statistical distribution of disparity between users' registered competitions can be used to describe their interests, and a disparity histogram can be used to display it.

Define a set $C_u = \{c_1, c_2, \dots, c_n\}$ which means user u's registered competitions, calculate the disparity between every two competitions according to the formula (2), put them in a set D_u

$$D_u = \{D(c_i, c_j) | 1 \le i \le n-1, i < j \le n\} \tag{4}$$

In the histogram, the x-axis is disparity intervals $X_u = \{x_0, x_2, \dots, x_k, \dots, x_9\}$, The x_k is defined as the range of disparity with values $[\frac{k}{10}, \frac{k+1}{10})$; The y-axis is the count of elements from D_u falls in very x_k interval, marked as $Num_u(x_k)$. Fig. 2 shows a user u's disparity histogram; it describes u's interests in different disparity intervals objectively. The y-axis shows the number of competitions that a user has registered in every disparity intervals; the bigger it is, the higher interest one has in it.

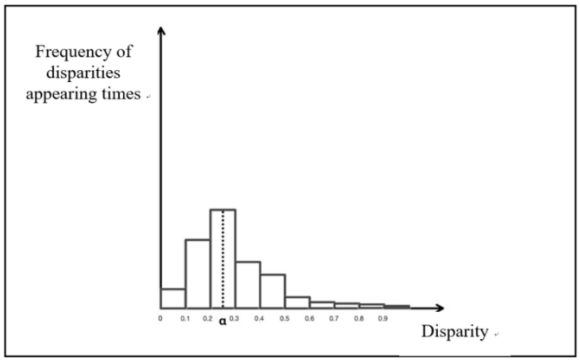

Fig. 2. Disparity histogram.

We define the median value of the interval contains α as User's Best Disparity, mark as α_u

$$\alpha_u = \frac{\frac{k}{10} + \frac{k+1}{10}}{2} = \frac{2k+1}{20} \tag{5}$$

In the formula, $k = \arg\max\ (Num_u\ (x_k)\ | x_k \in X_u)$.

It is sensible to analysis a user's historical records to find out what disparity is best for him.

2.4. The Range of User's Interests

Since user's best disparity is shown just as a single point, even based on the best disparity alone, the number of recommended competitions is extremely limited and may even be zero. Despite the point-introduced disparity interval, the number of recommended competitions is also difficult to meet user's filtering needs.

According to Figure 1, we can infer that the competitions included in the disparity interval have covered most of the user's interests, and are thus proper candidates for recommendation. However, in the vicinity of 0 and γ, interest degree is not high enough, so further narrowing of the range within $[0, \gamma]$ is needed. The range is centered on the user's best disparity α, and the interest radius r can be calculated according to the dispersion degree of disparity among competitions in user's historical entry records. The higher the

dispersion degree is, the larger the radius is, vice versa being true as well. So we can adjust relative parameters based on the above-mentioned fact to guarantee that the quantity of recommended competitions is good enough.

The standard deviation measures of how discrete the sample is as statistics. For the registered competition set of user $C_u = \{c_1, c_2, \ldots, c_n\}$, we define the standard deviation of disparity for all competitions as S_u

$$S_u = \sqrt{\frac{2}{(n+1)(n-2)-2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(D(c_i,c_j)-\bar{D}_u\right)^2} \tag{6}$$

In the formula, $c_i, c_j \in C_u$, $\bar{D}_u = \frac{2}{(n+1)(n-2)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} D(c_i,c_j)$.

According to the above result S_u , define the personalized user interests radius r_u

$$r_u = \min\{S_u, \quad \alpha_u\} \tag{7}$$

3. Disparity-Based Measurement on Participation Internet and Competition Recommendation Algorithm

3.1. User's Participation Interests Range and Recommendation Principe

Taking user's best disparity α as a center and interest radius r as radius, the interval $[\alpha - r, \alpha + r]$ represents the best disparity range in user's acceptable competitions. What's more, the sum of the participation interest in this interval is far greater than the sum at all other intervals.

For any competition i that user u don't participate in, i is supposed to satisfy following recommendation rules

$$T = \begin{cases} 1, & \alpha_u - r_u \le \bar{D}_u(i) \le \alpha_u + r_u \\ 0, & \text{other} \end{cases} \tag{8}$$

If T equals 1, the competition i should be recommended to the user, else not recommended.

3.2. Disparity-Based Competition Recommendation Algorithm

In view of the very performance of competition and the problems when using traditional recommendation algorithms in competition field, this paper puts forward a disparity-based algorithm for competition recommendation, whose main process is shown as follows:

1) According to the formula (4), construct a previously-attended competition disparity set D_u for user u, and accurately count the number of disparity in each interval, marked as $Num_u(x_k)(0 \le k \le 9)$;
2) According to the formula (5), calculate user's best disparity α_u;
3) According to the formula (7), calculate user's interest radius r_u;
4) Define a competition set I_u, which represents the competitions that are still available for registration and haven't been registered by the user yet, and $I_u = \{i | i \notin C_u \text{ and } i \text{ is available}\}$. According to the formula (3), calculate each competition disparity value $\overline{D_u}(i)$ corresponding to each i in the set I_u.
5) Choose the competitions in set I_u which satisfy the formula (8), and recommend them to the user.

4. Algorithm Validation and Experiment Results

Experimental data utilized in this research were obtained from a mobile application named "Competition Alliance", including: (n=300) related competitions information, (n=621) users' personal information, (n=1316) corresponding registration data of competition participation, (n=5845) users' interest information, (n=12390) users' browsing information. The experimental steps are as follows:

(1) Look up, via user ID, information from the user's registration form, interest information form and browsing data form, to obtain the competitions' information that the user has registered, browsed or followed with interest. Next, information is processed and redundancies are filtered out, priority given to registration form, then interest form, and finally browsing form. For example, if a contest appears in the registration form, the interest form and the browsing form at the same time, the data from The registration form is utilized while data from the other two forms are disregarded.

(2) 10% of the competition data are removed from both user's registration form and the interest form respectively and stored in the same y_true list after duplicates from the removed data are excluded. The data in above list is used as the test set which includes the user id and competition id;

(3) The remaining 90% of the data are used as training data which are based on the content filtering recommendation algorithm, the collaborative filtering recommendation algorithm and the algorithm proposed in this paper.

(4) Make recommendations to each user with the three aforementioned recommendation algorithms. Since the competitions in the y_true list comes from all users??, it may overlap with the competitions in registration or interest form of the user to be recommended. Therefore, before making any recommendations to a user, overlaps of the two lists must be removed in order to ensure that users are not recommended redundant competitions. The binary list of recommended results, y_predict, which takes into consideration the parameters of user id and competition id, indicates whether or not the competition should be recommended: 0 and 1 for not recommended and recommended, respectively. Subsequently, a comparison between y_predict and y_true is made given that the parameters of user id and competitions id are identical. The values of 0 and 1 between the two lists are compared and if the values are identical, a new binary list, $y_{expected}$, is created. $Y_{expected}$, which informs us of whether the training of the algorithm was effective. If values from the compared lists are equal then the value of $y_{expected}$ is 1, however if the values are different then the $y_{expected}$ value is 0. The experimental results are saved to the csv file, which shows performance of the three algorithms. The drawing based on local data is shown in Fig. 3 as following.

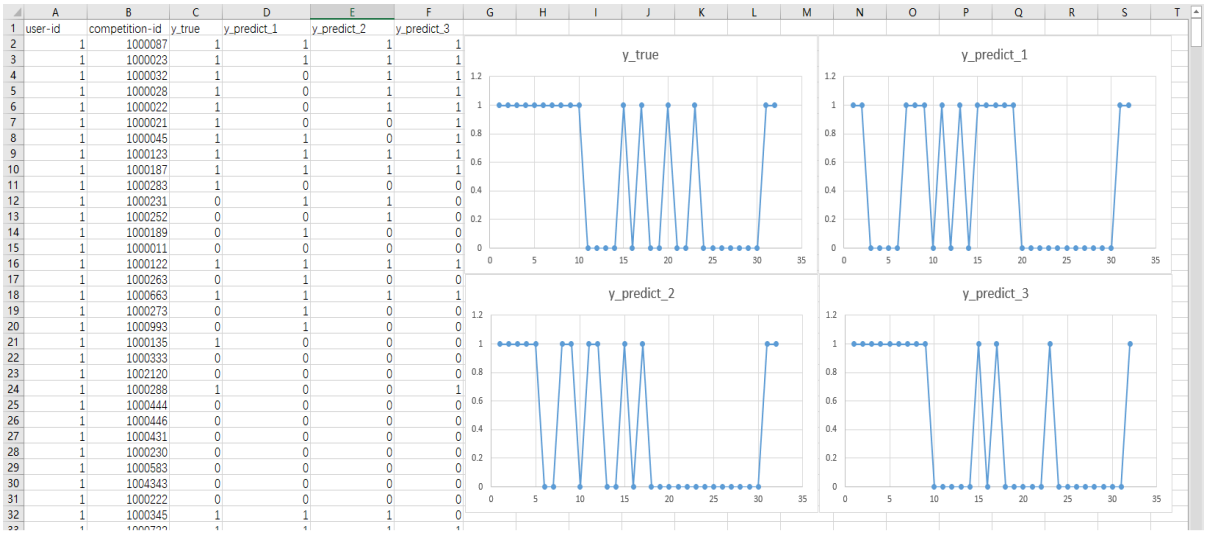

Fig. 3. Local data and recommendation results sample graph.

In Fig. 3, y_predict_1 shows the results of recommendation algorithm based on content filtering, y_predict_2 shows the results of the recommendation algorithm based on collaborative filtering, and y_predict_3 shows the results of the algorithm proposed in this paper.

The prediction accuracy of recommendation algorithm, defined P, can be calculated. TP, the number of adopted recommendations, accurately represents number of competitions that were registered or given attention after being recommended. FP, the virtual recommendation amount, represents the number of competitions that aren't been registered or given any attention despite being recommended. FN, missed recommendation amount, represents the number of competitions that are registered or given attention but aren't recommended. After running the experiment 100 times, the average values of TP, FP, FN are 73, 27, 89 respectively in content-based filtering recommendation algorithm, 80,20,89 respectively in collaborative filtering algorithm, and 73,27,89 respectively in the proposed algorithm in this paper. The graph of TP, FP and FN in the first 20 experiments obtained by the three algorithms is shown in Fig. 4.

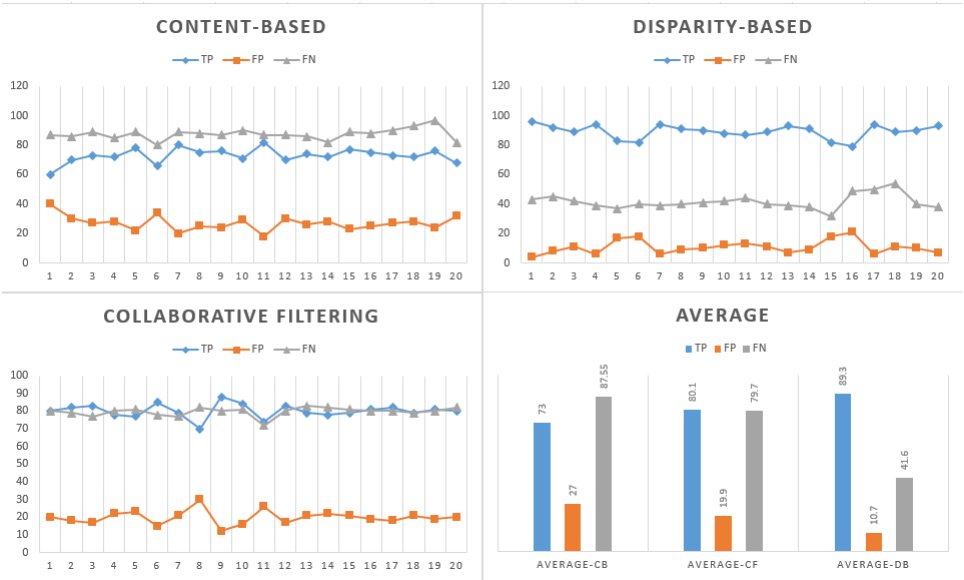

Fig. 4. The comparison chart of TP, FP, FN in three algorithms.

The prediction accuracy of recommendation algorithm P:

$$P = \frac{TP}{TP+FP} \tag{9}$$

And calculate recall rate R:

$$R = \frac{TP}{TP+FN} \tag{10}$$

Due to the fact that virtual recommendations and missed recommendations cannot be properly estimated based on prediction accuracy or recall rate alone, comprehensive evaluation measurement is needed which will consider accuracy, virtual and missed recommendation amounts This is defined as F_β score :

$$F_\beta = (1+\beta^2)\frac{P*R}{(\beta^2*P)+R} \tag{11}$$

Statistically, β 's values are generally take 1, 0.5, 2. When $\beta = 1$, F_β is defined as Harmonic Average of accuracy and recall rate; when $\beta = 0.5$, it results that accuracy has a higher weight than recall rate; and if $\beta = 2$, accuracy will have a lower weight than recall rate. Due to the fact that missed recommendations are undesired, we choose $F_{0.5}$ score to estimate experiment results.

According to the mean of TP, FP and FN after 120 experiments, the accuracy rate *P*, the recall rate R and

the $F_{0.5}$ score were calculated respectively. The experimental results are shown in Table 1.

Table 1. Accuracy, Recall and F-score of Three Algorithms

Algorithm	Accuracy P	Recall R	$F_{0.5}$ (calculated based on P and R)
Content-Based Filtering	0.73	0.45	0.65
Collaborative Filtering	0.80	0.51	0.72
Disparity-Based Algorithm	0.92	0.65	0.85

According to the Table 1, disparity-based algorithm demonstrated higher recommendation accuracy, recall rate, and F-score than those obtained from the two other algorithms. High accuracy demonstrates low virtual recommendation rate, i.e. amount of inaccurate, non-suitable recommendations are small. High recall rate indicates low missed recommendation rate, which means desirable recommendations that were missed are low. Finally, higher F-score further proves that algorithm in this paper demonstrate lower virtual and missed recommendations than those obtained from other algorithms apparently.

The experiment results, when compared to classical algorithms, of disparity-base algorithm evidently demonstrates decreased unsuitable recommendations and misses of suitable ones while providing more accurate recommendations.

5. Conclusion

Disparity-based competition recommendation algorithm demonstrates a promising novel approach in providing accurate and suitable recommendations for users. Data sparsity problem and a new-item problem that are usually associated with the current recommendation algorithms play roles in providing flawed and inaccurate recommendations. However, Disparity-based competition recommendation algorithm takes into consideration of features of competition evaluation, the differences in users' self-ability scope and differences in the individual's challenge psychology, which in the end helps address such issues. By replacing similarity degree of disparity degree for a measure and calculating the most appropriate disparity degree and interest radius for users, the personalized competition recommendation range can be determined for each user.

References

[1] Su, J, Yeh, & H, Yu, P. S., *et al*. (2010). Music recommendation using content and context information mining, *25(1)*, 16-26.

[2] Lan, Y., & Cao, F. F. (2007). Research of time weighted collaborative filtering algorithm in movie recommendation. *Computer Science*, *44(4)*, 295-301.

[3] Yoon, V. Y., Hostler, R. E., Guo, Z., *et al*. (2013). Assessing the moderating effect of consumer product knowledge and online shopping experience on using recommendation agents for customer loyalty. *55(4)*, 883-893.

[4] Zhou, X., Liang, H., & Dong, Z. (2017). A personalized recommendation model for online apparel shopping based on Kansei engineering. *International Journal of Clothing Science and Technology*, *29(1)*, 2-13.

[5] M. B., & Y, S., Fab: Content-based, collaborative recommendation. *40(03),* 66-72.

[6] Lang, K. (1995). Newsweeder: Learning to filter netnews: Machine learning. Tahoe City, CA.

[7] Roy, L., & Mooney, R. J., Content-based book recommending using learning for text categorization. *2000 Digital Libraries*, San Antonio, TX.

[8] G., A, & A., T. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering*, *17(6)*, 734-749.

[9] Huifeng, S. (2012). Personalized web recommendation via collaborative filtering. *Beijing University of Posts and Telecommunications*.
[10] Li, X., Cheng, X., Su, S., *et al*. (2017). A hybrid collaborative filtering model for social influence prediction in event-based social networks. *Neurocomputing*, 197-209.
[11] Viana, P., & Soares, M. (2017). A hybrid approach for personalized news recommendation in a mobility scenario using long-short user interest. *International Journal on Artificial Intelligence Tools*.
[12] Liu, C., & Wu, X. (2016). Fast recommendation on latent collaborative relations. *Knowledge-Based Systems*.
[13] Han, J., Jo, J., & Ji, H., *et al*. (2016). A collaborative recommender system for learning courses considering the relevance of a learner's learning skills. *Cluster Computing-The Journal of Networks Software Tools and Applications*.
[14] Aznoli, F., & Navimipour, N. J. (2017). Cloud services recommendation: Reviewing the recent advances and suggesting the future research directions. *Journal of Network and Computer Applications*.
[15] Yu, D., Mu, Y., & Jin, Y. (2017). Rating prediction using review texts with underlying sentiments. *Information Processing Letters*.

Kaixin Liu was born in Inner Mongolia Autonomous Region on August 18th, 1994.

She graduated from China University of Mining and Technology (Beijing) with bachelor's degree in engineering in the field of computer science and technology in June 2017.

She has done test engineer, Android developer and server developer in the past three summer vacation, she is now studying for a master degree in China University of Mining and Technology (Beijing), major in machine learning and big data analysis. Ms. Liu was the champion in computer application development competition in 2015.

Jianhui Chang was born in Hebei province, China on September 1st, 1996. She has been a college student in major of computer science and technology in China University of Mining and Technology (Beijing) since September 2015.

Her research interests include multimedia search and recommendation, 3D face modeling, and artificial intelligence. She involves the study of human facial expression capture as college student innovation project. She won the first price in application development competition for Android in November 2017.

Huiwen Ren has been a college student in major of computer science and technology in China University of Mining and Technology (Beijing) since September 2015.

He is now major in deep learning and android video encoding. He has developed several applications using the theories above.

Hong Zhu received her master degree in Kunming University of Science and Technology and Ph.D. from China University of Mining and Technology (Beijing). She is currently an associate professor and undertaking a master course on advanced computer graphics and bachelor courses on computer networks and compiler theory. She has devoted to the study of computer graphics and artificial intelligence in long terms.