http://ansinet.com/itj

ITJ

ISSN 1812-5638

# INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 13 (5): 978-983, 2014
ISSN 1812-5638 / DOI: 10.3923/itj.2014.978.983


# A Network-based Recommendation Algorithm via Improved Similarity Model

[1]Yuan Wen, [1]Yun Liu and [2]Wei Cao
[1]Key Laboratory of Communication and Information Technology,
Beijing Municipal Commission of Education, Beijing Jiao Tong University,
Beijing, 100044, People's Republic of China
[2]China Information Technology Security Evaluation Center, Beijing,
100085, People's Republic of China

**Abstract:** Many researchers have devoted their works toward improving the effect of recommendation algorithms. Here a new method is introduced to recommend information to users based on the Improved Similarity Model (ISM). Through the use of the well-known data set MovieLens as test data, the experiments testify that this method has a good effect on recommendation than other methods. The method can achieve the optimal value when the parameter in the ISM formula equals to a special value. By comparing the ISM model with several traditional models, the results show that the ISM model always has best recommendation effect in different test criteria fields. This model can significantly outperforms traditional models by not only enhancing recommendation accuracy but also improving recommendation diversity and giving more personalized recommendations.

**Key words:** Bipartite networks, user-object links, personalized recommendation, similarity model, link prediction, collaborative filtering

## INTRODUCTION

With the further development of the internet, lots of applications appear and the network becomes more and more intelligent (Albert and Barabasi, 2002; Newman, 2003; Boccaletti *et al*., 2006; Zhang and Zeng, 2012). Many applications online can recommend objects to users according to the past browsing behavior made by users (Lu and Zhou, 2011; Sarukkai, 2000; Liben-Nowell and Kleinberg, 2007; Clauset *et al*., 2008). For example, the biggest online seller Amazon, who can now not only sell kinds of commodities for users, but also recommend some goods to potential buyer (Billsus *et al*., 2002). When people browse things on Amazon's website, it can record the browsing data of every user. Then it recommends things to users according to some special algorithms. Sometimes the recommendations are not accurate, but usually users can get recommended things which are they just want.

There are manly three approaches for recommendation: Content-based recommendation, physical filtering recommendation and collaborative filtering recommendation. Content-based recommendation does not need large number of past preference information of users. It only relates with similar content of objects, such as similar topics between two movies (Pazzani and Billsus, 2007). However, this approach cannot supply plenty of personalized recommendations for users. Therefore, recently several efforts have been devoted to designing better recommendation algorithms based on physical filtering recommendations (Zhou *et al*., 2007; Jia *et al*., 2008; Liu and Deng, 2009). Some researchers use heat spreading from physics to simulate information spreading between users and objects (Qiu *et al*., 2011). Especially Zhou *et al*. (2008) do a lot of work and propose Network-Based Inference (NBI) model that can improve the recommendation effect in physical filtering.

Most widely used recommendation method in today's online application is collaborative filtering recommendation (Herlocker *et al*., 2004; Huang *et al*., 2004). The principle of a collaborative filtering approach is the use of historical preferences, which are retrieved from the browsing records or rating records of users. According to a serial of links between users and objects, a user-based collaborative approach finally can generate a recommendation list of objects, which are sorted descending by the sum of other users' ratings (Wilson *et al*., 2003; Sarwar *et al*., 2001). The approach of collaborative filtering is originally widely used in commercial commodity recommendation such as Amazon online, movie recommendation such as Netflix and video

**Corresponding Author:** Yun Liu, Key Laboratory of Communication and Information systems,
Beijing Municipal Commission of Education, Beijing Jiao Tong University,
Beijing, 100044, People's Republic of China

recommendation such as YouTube (Linden *et al*., 2003). The core of collaborative filtering is similarity calculation. Here a new model called Improved Similarity Model (ISM) is proposed. Through widely testing this method by numbers of times of tests based on ten group data that randomly sampled from Movie Len data set, the parameter in the ISM formula can be fixed at an optimal value for the best effect of recommendation. The experimental results then demonstrate that the method based on ISM model can remarkably improve the effects of recommendation over other methods.

## MATERIALS AND METHODS

**Introduction of bipartite network:** In a recommendation system, the users, objects and links can form a network usually called a bipartite network. Bipartite network is a specialkind of networks, which can be seen as a graph G (u, o, e). The graph contains users $u = \{u_1, u_2,..., u_m\}$, objects $o = \{o_1, o_2, ..., o_n\}$ and links $e = \{e_{ij}: u_i \in u, o_j \in o\}$. A link is drawn between $u_i$ and $o_j$ if user i has collected object j (when the rating is no less than 3 if the rating scale is from 1-5). For example, a bipartite network with three users and four objects is shown as follows.

In Fig. 1 user $U_1$ has linked objects $O_1$ and $O_3$, $U_2$ has linked objects $O_2$ and $O_3$ and $U_3$ has linked objects $O_1$ and $O_4$. The topology of network between users and objects can be described as a matrix:

$$M = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

This matrix M can describe the relation between users and objects. The goal of a recommendation algorithm is to "guess" whether a user likes an object or not. This guess is also called prediction, which can be calculated by the topology matrix between users and objects. Thus the matrix M can help to calculate pair wise similarity between users. After calculating the similarity based on similarity model, the recommendation of objects can be got by most similar users. The process of recommendation can be summarized as three steps listed below:

- **Step 1:** To decide which pair of users has the largest similarity, firstly calculate the similarity between users. So, if there are n users, it will have $c_n^2$ user pairs and need to calculate t so many times. This can mathematically calculated to similarity model
- **Step 2:** To recommend objects for any user, we need to choose other users with similarity above the threshold. The threshold is calculated by the average value in similarity matrix S. Then a list of objects linked can be got by these users

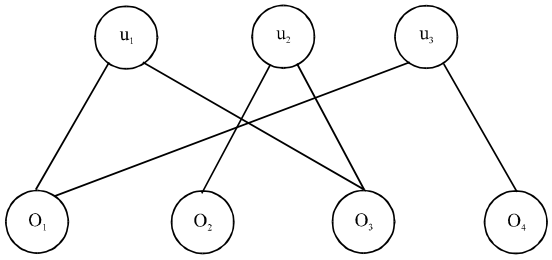


Fig. 1: Links between users and objects in a bipartite network

- **Step 3:** After wiping off the objects in the list linked by the target user (who is recommended), top L objects from the list with descending order are chosen and can be recommended to this user. These objects are called recommendation list

**Traditional similarity formulas:** Firstly choose out nine most widely used similarity formulas, then define x and y as two arbitrary users in a bipartite network. Then set C(x) and C(y) as the sets of objects linked by x and y respectively. k(x) and k(y) denote the degrees of user x and y, respectively. Nine similarity formulas are showed as follows:

**Common Neighbors Model (CN):**

$$Sim_{CN}(x,y) = |C(x) \cap C(y)| \quad (2)$$

Common neighbors' model is an easy way to calculate the similarity between two users. It is motivated by the idea that if user x and user y have more common linked objects, they may have more similarity. Above $|C(x) \cap C(y)|$ measures the number of the overlap of objects linked by user x and user y. The more number of common linked objects by two users, the larger the similarity between them.

**Cosine similarity model (cosine):**

$$Sim_{Cosine}(x,y) = \frac{|C(x) \cap C(y)|}{\sqrt{k(x) \times k(y)}} \quad (3)$$

This also is called Salton index in some studys. Here k(x) and k(y) indicate the degrees of user x and user y. Similarity between them strongly depends on the number of commonly linked objects by two users, but is inversely proportional to the degrees of two users. This similarity method is mostly used in many fields such as classifying or clustering of documents data.

**Jaccard similarity model (Jaccard):** The Jaccard model measures similarity between sample sets and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$Sim_{Jaccard}(x,y)=\frac{|C(x)\cap C(y)|}{|C(x)\cup C(y)|} \quad (4)$$

**Sorensen similarity model (sorensen):** This model is usually used for ecological similarity. It is defined as:

$$Sim_{Sorensen}(x,y)=\frac{|C(x)\cap C(y)|}{k(x)+k(y)} \quad (5)$$

**Hub Promoted model (HP):**

$$Sim_{HP}(x,y)=\frac{|C(x)\cap C(y)|}{\min\{k(x),k(y)\}} \quad (6)$$

The HP method is similar with HD method. The only difference is that this method considers the minimum of degrees in the denominator. So this method can get higher value than HD method.

**Hub Depressed model (HD):** Analogously to the above HP method, use this method for comparing with HP method. It is defined as:

$$Sim_{HD}(x,y)=\frac{|C(x)\cap C(y)|}{\max\{k(x),k(y)\}} \quad (7)$$

**Leicht-Holme-Newman model (LHN):**

$$Sim_{LHN}(x,y)=\frac{|C(x)\cap C(y)|}{k(x)\times k(y)} \quad (8)$$

**Adamic-Adar model (AA):**

$$Sim_{AA}(x,y)=\sum_{z\in C(x)\cap C(y)}\frac{1}{\log k_z} \quad (9)$$

In AA model and RA model above, z is the common neighbors linked by user x and user y. $K_z$ is the degree of object z. The AA model and RA model have very similar form. They both depress the influence of the high-degree common neighbors. It means that at the same time increase the effect of low-degree objects. The difference between AA method and RA method is that the former depresses the objects of high degree by $\log k_z$ but the latter d by $k_z$.

**Resource Allocation model (RA):**

$$Sim_{RA}(x,y)=\sum_{z\in C(x)\cap C(y)}\frac{1}{k_z} \quad (10)$$

**Improved similarity model:** The Improved Similarity Model (ISM) is motivated by the idea that if we can calculate a similarity between two users, we may focus on not only common linked objects but also comprehensive factors:

$$Sim(x,y)=\frac{C(x)\cap C(y)}{\Gamma(x,y)}\cdot\frac{\log^{\gamma}\Gamma(x,y)}{k(x)\times k(y)} \quad (11)$$

Above $|C(x)\cap C(y)|$ measures the number of the overlap of objects linked by user x and user y. $\Gamma(x, y)$ stands for the number of totally linked objects by user x and user y. The fraction $|C(x)\Gamma C(y)|$ divided by $\Gamma(x,y)$ refers to how many percent of links in totally linked objects. At the same time, Sim (x, y) is also proportional to $\log^{\gamma}\Gamma(x, y)$, but inversely proportional to the degrees of two users. Inside it $\gamma$ can be fixed for an optimal value by experiments.

## RESULTS AND DISCUSSION

**Dataset:** In order to test the proposed method, a well-known data set called MovieLens is used. It is placed online and can be downloaded from the following website: http://www.grouplens.org. MovieLens has collected lots of movies (which can be treated as objects in the bipartite networks) and the ratings of them by a lot of users. In the dataset each user gives any movie a rating from 1-5. If the rating is not less than 3, then we can draw a link between the user and the movie. All links in the dataset between users and objects can form a bipartite network. This data set contains 82,520 links between 943 users and 1682 objects.

Ten groups of data are randomly chosen from MovieLens. In the dataset, each user chooses at least 20 objects. Each group of test data is 90/10% split into training set and test set. The length of recommendation list is set to be L=10. Table 1 illustrates the basic statistics of the data sets.

Each test data can be seen as a bipartite network. The average network sparsity of the data sets is 5.83%.

Table 1: Basic statistics of the data sets

| Data sets | Users | Objects | Links | Network sparsity (%) |
|---|---|---|---|---|
| Test Date 1 | 454 | 1554 | 37598 | 5.33 |
| Test Date 2 | 177 | 1393 | 14866 | 6.03 |
| Test Date 3 | 235 | 1497 | 20622 | 5.86 |
| Test Date 4 | 129 | 1298 | 10949 | 6.54 |
| Test Date 5 | 320 | 1579 | 27120 | 5.37 |
| Test Date 6 | 402 | 1555 | 34848 | 5.57 |
| Test Date 7 | 355 | 1546 | 32726 | 5.96 |
| Test Date 8 | 283 | 1479 | 24944 | 5.96 |
| Test Date 9 | 192 | 1431 | 16791 | 6.11 |
| Test Date 10 | 221 | 1418 | 18138 | 5.79 |

**Criteria of recommendations:** There are many criteria of recommendations. Many indices can be used for measuring recommendation effect, such as precision, recall, F-measure, ranking score and so on. Some of them are used to evaluate the recommendation effect:

- **Precision:** When measuring the precision of recommendation algorithm, we can use average precision, which can usually be defined as follows:

$$P = \frac{1}{m} \cdot \frac{\sum_{i=1}^{m} d_i}{L} \tag{12}$$

where, m means the number of users and L is the recommendation list's length (in the experiment, L = 10) stands for the number of correct recommendations for user i

- **Recall:** The average of recall rate can usually be defined as follows:

$$R = \frac{1}{m} \cdot \sum_{i=1}^{m} \frac{d_i}{N_i} i \tag{13}$$

where, m means the number of users and L is the recommendation list's length (in the experiment, L = 10). $N_i$ stands for the number of recommended objects for user i in test set

- **Diversity:** The diversity of recommendation list is a very useful criterion to evaluate the recommendation effect. Because the greater the diversity is, the more personalized the recommendation algorithm gives. In this study the average intra-user diversity is used as the criterion of diversity of recommendation

First of all, calculate the similarity between two objects of bipartite networks by:

$$S_{ij} = \sum_{u=1}^{m} \frac{a_{ui} \cdot a_{uj}}{\sqrt{k_i \cdot k_j}} \tag{14}$$

Above the value of $\alpha$ depends on that whether user linked the object. If user u links object i, then $\alpha$ equals to 1, otherwise $\alpha$ equals to 0. $k_i$ and $k_j$ mean the degree of object i and j, respectively. So, we can get the intra-user diversity of any user by:

$$D_{user} = \frac{1}{L(L-1)} \cdot \sum_{i \neq j} (1 - S_{ij}) \tag{15}$$

Then, we can get the diversity (average inter-user diversity) finally as:

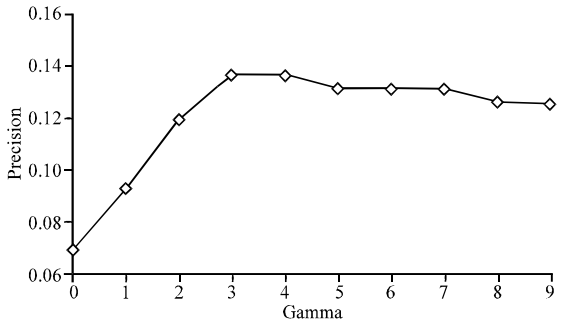


Fig. 2: Precision rate with different gamma

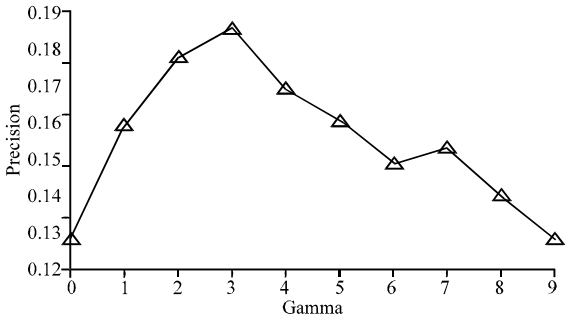


Fig. 3: Recall rate with different gamma

$$D = \text{average}(D_{user}) = \frac{1}{m} \cdot \sum_{user=1}^{m} D_{user} \tag{16}$$

The diversity D reflects the personalized of recommendation result. The greater the diversity is the higher novelty for any user's recommendation.

**Value of parameter in the similarity formula:** To determine the gamma in Eq. 11, many experiments can be tested and then choose the optimal one. The relation between gamma and precision and diversity can be showed below.

Using ISM (improved similarity model) based on 10 groups of data (Table 1), the average of relation can be got between precision and the parameter gamma as shown in Fig. 2. We can also get the relation between recall rate and the parameter gamma as shown in Fig. 3. In this test, the length of recommendation list is 10 and the data is 90/10% split into training set and test set.

The figures above reveal the recommendation effect of different gammas. From Fig. 2, we can see that the largest value of precision when gamma is 3. From Fig. 3, we can see that the largest value of recall rate when gamma is also 3. Thus, the parameter value of gamma is 3 in Eq. 11. Then the recommendation can get the optimal recommendation effect.

Table 2: Comparison of different similarity models

| Models | Precision (%) | Recall(%) | Diversity(%) |
|---|---|---|---|
| CN | 11.4 | 9.9 | 44.1 |
| Cosine | 13.2 | 15.8 | 50.3 |
| Jaccard | 13.6 | 16.9 | 53.1 |
| Sorensen | 13.1 | 16.3 | 52.5 |
| HP | 11.5 | 10.9 | 47.6 |
| HD | 13.7 | 18.0 | 53.5 |
| LHN | 13.5 | 18.0 | 58.3 |
| AA | 30.70 | 50.67 | 22.7 |
| RA | 11.6 | 10.0 | 45.2 |
| ISM | 14.6 | 18.6 | 57.5 |

Table 3: Comparison of different lengths of recommendation list

| Parameters | Recommendation lists | CN (%) | Cosine (%) | Jaccard (%) | AA (%) | ISM (%) |
|---|---|---|---|---|---|---|
| Precision (%) | L=5 | 13.6 | 14.5 | 15.9 | 6.4 | 16.4 |
| | L=10 | 11.4 | 13.2 | 13.6 | 3.70 | 14.6 |
| | L=20 | 8.35 | 9.89 | 10.2 | 2.20 | 10.75 |
| | L=50 | 5.40 | 7.55 | 7.91 | 1.83 | 8.28 |
| Diversity (%) | L=5 | 41.9 | 45.6 | 49.5 | 13.9 | 49.9 |
| | L=10 | 44.1 | 50.3 | 53.1 | 22.7 | 57.5 |
| | L=20 | 46.7 | 52.8 | 55.2 | 31.5 | 62.2 |
| | L=50 | 52.3 | 54.2 | 59.7 | 39.4 | 69.5 |

**Comparisons:** When the length of recommendation list is 10 with a 90/10% split of the data set, we can test the recommendation effects of the model (ISM) and compare this model with nine traditional models.

Table 2 displays that the ISM model has the best precision and recall rate of recommendation. When focusing on the precision of recommendation, the ISM model increases 6.57% compared with the HD model, which has the second-largest precision. When focusing on the recall rate of recommendation, the ISM model increases 3.3% compared with the HD and LHN models, which have the second largest precision. In the field of diversity, the ISM model also has a good performance and 57.5% diversity, which means that it can provide users with personalized recommendation.

We test the recommendation effect of typical models with the different lengths of recommendation list. As Table 3 shows, the precision decreases when the L increases. That means when recommending in a few list, the results may have good performance; when in a large list, the result decreases because more irrelative objects can be included in the list. The diversity presents the reverse effect: when L increases, the diversity of recommendation also increases.

**Discuss the threshold of similarity:** In step 2 (stated previously), the threshold can be set at an average value of all similarity. With the different threshold, the final recommendation effects may change. If give a user indexed by i, then we can sort the i-th row of the users' similarity matrix S. Then rank them in non-ascending order to obtain the ranks of the other users:

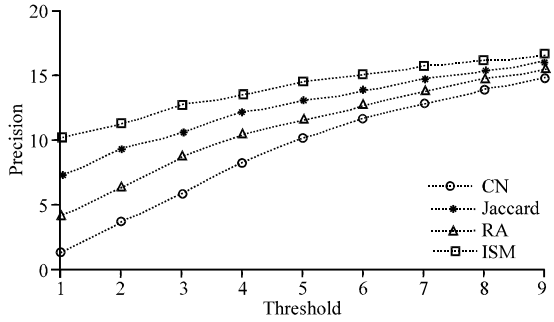


Fig. 4: Effects of different thresholds of similarity on the recommendation precision

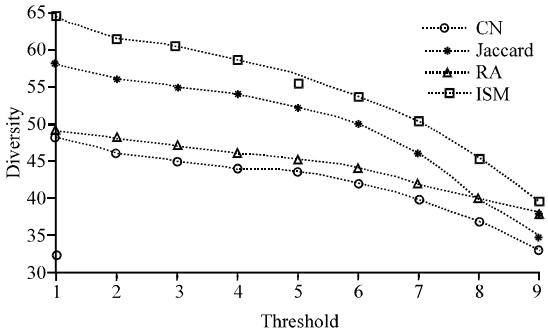


Fig. 5: Effects of different thresholds of similarity on the recommendation diversity

$$R_i = [r_{i1}, r_{i2}, \cdots, r_{im}] \tag{17}$$

Then, a parameter called lambda is introduced and when rank is behind than the value of $\lambda \times m$, this object is disposed. Among them m is the number of users. We can get the new similarity matrix as:

$$S = \begin{cases} S_{ij} & \text{when } r_{ij} < \lambda \times m \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

With different $\lambda$ (threshold of similarity), we can draw Fig. 4 and 5 based on the average values of data.

From Fig. 4, we can clearly see the increase of the recommendation precision when the threshold increases. We also see that ISM always has the highest value when the threshold increases. From the Fig. 5, the diversity is decreasing when the threshold is increasing. In this situation, the ISM model is also the highest among other models. The figures above reveal that ISM has the best recommendation effect in the test of precision and diversity even if with different threshold of similarity.

## CONCLUSION

A network-based recommendation algorithm based on the improved similarity model is proposed. Comparing with other traditional similarity models, the ISM model has the construction of considering not only common linked objects but also comprehensive factors when calculating the similarity between two users. In the ISM model, a parameter denoted as gamma is introduced and the algorithm can get the best recommendation effect when gamma is fixed in an optimal value through a series of experiments. When the parameter is fixed, the model has an increases of 6.57% compared with the traditional HD model when focusing on the precision of recommendation. In the experiments, a parameter denoted as lambda is introduced as the similarity threshold. The threshold is useful to filters out low similarity in the similarity matrix. When irrelative objects are deleted, the model can improve the recommendation precision. Through performing large numbers of experiments and then demonstrate that the ISM model has the best performance over other typical methods focused on the precision, recall and diversity of recommendation.

## ACKNOWLEDGMENTS

This research is supported by National Natural Science Foundation of China (No. 61271308, 61172072), Beijing Natural Science Foundation (No. 4112045), Beijing City Science and Technology Project (No. Z121100000312024) and the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 2010000911000).

## REFERENCES

Albert, R. and A.L. Barabasi, 2002. Statistical mechanics of complex networks. Rev. Modern Phys., 74: 47-97.

Billsus, D., C.A. Brunk, C. Evans, B. Gladish and M. Pazzani, 2002. Adaptive interfaces for ubiquitous web access. Commun. ACM, 45: 34-38.

Boccaletti, S., V. Latora, Y. Moreno, M. Chavez and D.U. Hwang, 2006. Complex networks: Structure and dynamics. Phys. Rep., 424: 175-308.

Clauset, A., C. Moore and M.E.J. Newman, 2008. Hierarchical structure and the prediction of missing links in networks. Nature, 453: 98-101.

Herlocker, J.L., J.A. Konstan, L.G. Terveen and J. Reidl, 2004. Evaluating collaborative filtering recommender systems. ACM Trans. Infrom. Syst., 22: 5-53.

Huang, Z., H. Chen and D. Zeng, 2004. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Trans. Inform. Syst., 22: 116-142.

Jia, C.X., R.R. Liu, D. Sun and B.H. Wang, 2008. A new weighting method in network-based recommendation. Phys. A: Stat. Mechanics Appl., 387: 5887-5891.

Liben-Nowell, D. and J. Kleinberg, 2007. The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol., 58: 1019-1031.

Linden, G., B. Smith and J. York, 2003. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Int. Comput., 7: 76-80.

Liu, J. and G. Deng, 2009. Link prediction in a user-object network based on time-weighted resource allocation. Phys. A: Stat. Mechanics Appl., 388: 3643-3650.

Lu, L. and T. Zhou, 2011. Link prediction in complex networks: A survey. Phys. A: Stat. Mechanics Appl., 390: 1150-1170.

Newman, M.E.J., 2003. The structure and function of complex networks. Soc. Ind. Applied Math. Rev., 45: 167-256.

Pazzani, M.J. and D. Billsus, 2007. Content-Based Recommendation Systems. In: The Adaptive Web: Methods and Strategies of Web Personalization, Brusilovsky, P., A. Kobsa and W. Nejdl (Eds.). Springer, New York, pp: 325-341.

Qiu, T., G. Chen, Z.K. Zhang and T. Zhou, 2011. An item-oriented recommendation algorithm on cold-start problem. EPL, Vol. 95. 10.1209/0295-5075/95/58003

Sarukkai, R.R., 2000. Link prediction and path analysis using Markov chains. Comput. Networks, 33: 377-386.

Sarwar, B., G. Karypis, J. Konstan and J. Reidl, 2001. Item-based collaborative filtering recommendation algorithms. Proceedings of the 10th International Conference on World Wide Web, May 1-5, 2001, Hong Kong, China, pp: 285-295.

Wilson, D.C., B. Smyth and D. O'Sullivan, 2003. Sparsity reduction in collaborative recommendation: A case-based approach. Int. J. Pattern Recognition, 17: 863-884.

Zhang, C.J. and A. Zeng, 2012. Behavior patterns of online users and the effect on information filtering. Phys. A: Stat. Mech. Appl., 391: 1822-1830.

Zhou, T., J. Ren, M. Medo and Y.C. Zhang, 2007. Bipartite network projection and personal recommendation. Phys. Rev. E, Vol. 76. 10.1103/PhysRevE.76.046115

Zhou, T., L.L. Jiang, R.Q. Su and Y.C. Zhang, 2008. Effect of initial configuration on network-based recommendation. EPL (Europhys. Lett.), Vol. 81. 10.1209/0295-5075/81/58004