Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 31 (2014) 875 – 881

Procedia Computer Science

2nd International Conference on Information Technology and Quantitative Management, ITQM 2014

# Recommendation Algorithm Based On Link Prediction And Domain Knowledge In Retail Transactions

Jing Li[a,b], Lingling Zhang[a,b]*, Fan Meng [a,b], Fenhua Li[a,b]

[a] University of Chinese Academy of Sciences, Beijing, 100190, China
[b] *Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100190, China*

**Abstract**

In this paper, we propose a new recommendation algorithm, which extends the idea of linkage measure to recommendation in bipartite network, and incorporate domain knowledge with topological property in recommendation process. Through calculating domain similarities between products, we weigh the products recommended to potential customer with larger weights, whose categories are more close to the categories meeting with users' preference, so as to improve the recommendation quality. Our preliminary experimental results based on a retail transaction dataset indicate that domain-based link prediction measures achieved better performance than general linkage measures algorithms.



Selection and peer-review under responsibility of the Organizing Committee of ITQM 2014.

*Keywords:* Link prediction, Recommender system, Domain knowledge;

## 1. Introduction

Recommender systems have found various applications of commercial success in e-commerce [1]. At the heart of recommendation systems are recommendation algorithms. Collaborative Filtering (CF) is the most commonly used and successful recommendation algorithm [2, 3]. Despite its success, the performance of collaborative filtering is strongly limited by not only the sparsity of data [3], but also the neglect of accessorial information [4]. Under this circumstance, CF doesn't support valid inference of user neighbors, making it confronted with great challenges of

* Corresponding author. fax: +86-010-8268-0676.
*E-mail address:* zhangll@ucas.ac.cn.


Selection and peer-review under responsibility of the Organizing Committee of ITQM 2014.
doi:10.1016/j.procs.2014.05.339

accuracy. Recently some algorithms based on link prediction are proposed for personalized recommendations, which explore transitive user-item associations by abstracting users' behaviors into a user-object bipartite network, and alleviate the sparsity problem, and it takes an advantage of network structural features to remarkably improve the quality of recommendation.

However, we found that most of network-based recommendation algorithms relied solely on the network structure topological features of the graph, which only takes the transaction data into account without product domain knowledge inherent in certain field, [5] indicates that some accessorial information can be exploited to further improve the algorithmic accuracy, which plays an important role in network-based recommendation as well. Furthermore, there are still few studies on personalized recommendation in traditional fields, such as retail, finance, telecommunication [6].

In this paper, in order to take both topological structure properties and node attributes in bipartite network into account, we proposes a new link prediction algorithms combined with domain knowledge to study the evolution of interactions among consumers and products reflected in real supermarket transactions data, and .apply it into recommender system. In this algorithm, we weigh the products recommended to potential customer with larger weights by calculating domain similarities between products, whose categories are more close to the categories meeting with users' preference.

The rest of the paper is organized as follows. Section 2 briefly discuss the general link prediction approach to recommendation. Section 3 propose a recommendation approach based on domain knowledge and link prediction Section 4 we present a preliminary experimental study comparing the proposed domain knowledge-based link prediction approaches with general link prediction algorithm. Section 5 concludes the study and points out our future work.

## 2. A link prediction approach to recommendation

### *2.1. Description of a customer-product bipartite network*

Complex network, with nodes representing individuals or organization and links denoting the relations or interactions between nodes, is regarded as a new efficient method to represent and study a wide range of complex systems [7-10]. A particular class of network is the bipartite network, whose nodes can be divided into two disjoint sets X and Y, such that only the link between two nodes in different sets is allowed [11, 12].

In a retail transaction, customer-product interactions can be abstracted and represented into a customer-product bipartite network G, in which the nodes are composed of two sets of different classes, product-set $P= \{p_1, p_2,\ldots, p_m\}$ and a customer-set $C= \{c_1, c_2,\cdots, c_n\}$, we therefore can get the $m\times n$ matrix $A= \{a_{ij}\}$, where $a_{ij}=1$, if $c_i$ purchase $p_j$, or otherwise $a_{ij}=0$. Link prediction aims at estimating the likelihood of the existence of a link between two nodes based on the observed links and the attributes of the nodes. Under this graph representation, the problem of effective recommendation can be viewed as a task of selecting unobserved links for each user node, then can be modeled as a link prediction problem [13-15], which achieved better performance than traditional CF, as considering the immediate neighborhood of users in an user-item bipartite graph, other than only the direct neighbors.

Fig. 1 shows prediction of links in a customer-product network. The solid line denotes the fact that the customer has already purchased the product. The dashed line represents the prediction that the customer purchase the product. The main task of link prediction is to predict the possible links (dashed lines) in the customer-product network based on the link information (solid lines) observed.

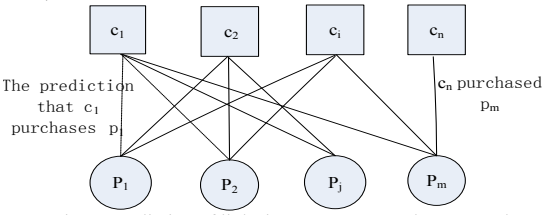


Fig. 1. Prediction of links in a customer-product network.

### 2.2. Linkage measures in a customer-product bipartite network

Linkage measures are defined to infer the potential for a future link to appear based on the topology feature of the graph. In the application of biology, the pair-wise topological similarity among bait proteins is computed according to the number of commonly shared neighbours [16]. When making recommendations for a particular customer on the basis of the topological feature in bipartite network, the very basic approach is to select the products recommended to target customer, which commonly shared neighbours with the products purchased by the target customer. In consideration of this point, Zan Huang et al. [5] revised a wide range of linkage measures proposed for unipartite graphs, where links are allowed between any pair of nodes. One of the adapted linkage measure is adopted here so as to compute linking possibility between customers and products in the bipartite network.

Based on the topology feature of G, we compute certain linkage measure for each customer-product pair (p, c) that haven't been purchased by the target customer (dashed lines) in the bipartite network. Afterwards those measures serve as scores to assess the possibility of a link connecting p and c.

For a node x, $\Gamma(x)$ is defined as the set of neighbours of x, so the set of neighbors of x's neighbors is defined as:

$$\hat{\Gamma}(x) = \bigcup_{c \in \Gamma(x)} \Gamma(c) \tag{1}$$

In a unipartite graph, the number of common neighbours of x and y is defined as

$$CN = \left|\Gamma(x) \cap \Gamma(y)\right| \tag{2}$$

which could represent the topological similarity of nodes. In our study, a customer node is only connected with product nodes in customer-product graph, thus CN will always be zero if x and y form a customer-product pair. The adapted common neighbour measure ( $B\_CN$ ) within a bipartite graph in [5] is

$$B\_CN = \left|\Gamma(x) \cap \hat{\Gamma}(y)\right| \tag{3}$$

Jaccard's Coefficient [16] is defined as

$$\text{Jaccard's Coefficient} = \left|\frac{\Gamma(x) \cap \Gamma(y)}{\Gamma(x) \cup \Gamma(y)}\right| \tag{4}$$

which measures the number of neighbours of both x and y compared to the number of nodes that are either x's or y's neighbours. In [5], In order to accommodate for the bipartite characteristics, the adapted Jaccard's Coefficient ( B_Jaccard's Coefficient ) is defined as

$$\text{B\_Jaccard's Coefficient} = \left|\frac{\Gamma(x) \cap \hat{\Gamma}(y)}{\Gamma(x) \cup \hat{\Gamma}(y)}\right| \tag{5}$$

where $\left|\Gamma(x) \cap \hat{\Gamma}(y)\right|$ denotes the number of common neighbours in customer-product bipartite network, while $\left|\Gamma(x) \cup \hat{\Gamma}(y)\right|$ represent popularity of the product and activeness of the customer.

When the task is to recommend items which are likely to be appreciated by a particular user, it is usually most effective to recommend popular and highly rated items. Such recommendation, however, has very little value for the customers because popular products are easy to find without a recommender system. Under this situation, we choose the B_Jaccard's Coefficient as the main link prediction approach for our task of recommendation, as it lowers the possibility of popular products to be recommended by taking $\left|\Gamma(x) \cup \hat{\Gamma}(y)\right|$ as the penalty factor.

The top-N products can be obtained by sorting the final scores of linkage measures in descending order, so we can predict the links of the target customer between those n products, and make recommendations for them according to the result of descending order.

Consider the case that customer $c_1$ purchased $\{p_3, p_4\}$, while $c_2$ purchased $\{p_1, p_2, p_3\}$, $c_3$ purchased $\{p_1, p_2, p_4\}$, and $c_4$ purchased $\{p_4\}$. Fig 2 describes the purchasing behavior between 4 customers and 4 products, those products including two categories, $\{p_1\}$ belongs to Class1 and $\{p_2, p_3, p_4\}$ belongs to Class2. The result is that $p_1$ and $p_2$ have the same probability to be recommended to c1. In real life, the user's purchase behavior in the past to a certain extent reflect the user's current preferences. As you can see from this example, $c_1$ is used to purchasing products in class2, so the probability of $p_2$ to be recommended should larger in fact than $p_1$, which is contrast to the real life.

In a word, it is necessary to take into account notes information effects when predicting links in a customer-product network. However, the linkage measures described above simply represents whether the customer purchases

the product or not, and the network does not fully describe the information of the products which customer purchased. It is a new problem for us to analyze the customer-product network integrated with domain knowledge.

## 3. Recommendation algorithm based on domain knowledge and link prediction

### *3.1. Analysis of domain knowledge in a customer-product bipartite network*

Domain knowledge is a special knowledge which indicates the interrelation of the concept within a certain domain and among the concepts as well as the constraint integration of the relevant concept. Zhang et al. integrate domain knowledge in the whole process of personalized recommendation to overcome problems of traditional recommendation method [17]. Domain ontology takes a specific domain as describing objects, providing the definition and the relationship between the concepts in a particular field. [18]develops a mechanism for domain knowledge adaptation for personalized knowledge search and recommendation to adapt a suitable domain ontology according to the previous browsing and reading behavior of users (i.e., usage history log).

Generally speaking, the topological structure characteristics in different network are calculated by some common linkage measures, however, the nodes attributes relied on a particular field of the network. For example, attributes of the movie nodes in the network of online movie recommendation, which serve as the key domain knowledge in movie area, such as actor, director, film type and so on. Graph-based attribute-aware method [19] takes into account item attributes, which are defined by certain domain experts.

In retail transaction, as for a customer, the more information related the customer's selecting behavior, the higher the predictive power the system can provide. In general, customer preferences is the guide-lines for the customer's shopping behavior. Those product category paid more attention shall be reflected in the product that customer past purchased, which plays an important role in decision-making while shopping. Therefore, as for the retail industry, the most important domain knowledge is the product category, which reflects the category of the product itself and the relationships between categories. Therefore, if link Prediction algorithms solely on customer's consumption information and ignore the category information of the purchased products, it will lead to the poor recommendation of above cases.

In order to make the recommended product' category more relevant to the category he/she favorites. Therefore, this paper firstly serve product domain knowledge in the retail area as product node attributes, then attaches importance on the establishment of product category hierarchy and its integration with link prediction algorithms. In this case, this article will use the established hierarchical classification of retail products to improve the link prediction algorithm based on the topological structure characteristics of the bipartite network.

### *3.2. Linkage prediction approach integrated with product domain knowledge*

According to the existing FoodMart supermarket data set, we built a product classification hierarchy, with six levels, and covering 1560 different kinds of products, which reflect the subordinate relationship among parent class, subclass, and the entities. Through the product classification hierarchy, we calculate the semantic similarity between different products, which reflect the proximity of product notes to another products notes in bipartite network. A new proposed algorithm in [20] combine topological features and semantic similarities between proteins to discover protein complexes in TAP-MS PPI networks. Inspired by this study, in this paper, we incorporate domain-based semantic similarity with the topology measures of bipartite network to improve the link prediction algorithm based on the structure of the network.

There are many possible semantic similarity functions. They fall broadly into two categories: 1.Semantic similarity calculation based on distance [21, 22]; 2. Semantic similarity calculation based on the information content [23-26].

The similarity functions referenced in this study was proposed by Ganesan, Garcia-Molina et al, which gained good performance in experiment [17]. The domain similarity between two products ($p_i$, $p_j$) is defined as:

$$Sim(\mathrm{p}_i,\mathrm{p}_j)=\frac{2\times \mathrm{depth}(\mathrm{LCA}(\mathrm{p}_i,\mathrm{p}_j))}{\mathrm{depth}(\mathrm{p}_i)+\mathrm{depth}(\mathrm{p}_j)} \tag{6}$$

$LCA(\mathrm{p}_i,\mathrm{p}_j)$ is the lowest common ancestor between products $\mathrm{p}_i$ and $\mathrm{p}_j$, so $\mathrm{depth}(\mathrm{LCA}(\mathrm{p}_i,\mathrm{p}_j))$ denotes the path length from root node to it.In customer-product bipartite graph the linkage measure of Jaccard's Coefficient is defined as：

$$S\left(\mathrm{p}_i,c_j\right)=\left|\frac{\Gamma(\mathrm{p}_i)\cap\hat{\Gamma}(c_j)}{\Gamma(\mathrm{p}_i)\cup\hat{\Gamma}(c_j)}\right| \tag{7}$$

As Jaccard's Coefficient treat all products' common neighbours equally no matter what the products' position are in product category hierarchy, then we weight initial linkage measure $\mathrm{S}=\left(\mathrm{p}_i,c_j\right)$ with domain similarity $Sim(\mathrm{p}_i,\mathrm{p}_j)$. Linkage measure of Jaccard's Coefficient based on domain knowledge is defined as:

$$D\text{B\_Jaccard's Coefficient}=\frac{\sum_{\mathrm{p}_m\in\mathrm{P}'}^{\left|\Gamma(\mathrm{p}_i)\cap\Gamma_{P\in\Gamma(c_j)}(\mathrm{P})\right|}sim(\mathrm{p}_i,\mathrm{p}_m)}{\left|\Gamma(\mathrm{p}_i)\cup\Gamma_{P\in\Gamma(c_j)}(\mathrm{P})\right|} \tag{8}$$

$\Gamma(\mathrm{p}_i)$ is defined as the set of neighbours of $\mathrm{p}_i$, $\Gamma_{P\in\Gamma(c_j)}(\mathrm{P})$ is the set of neighbours of $c_j$ 's neighbours, In the formula of DS, the denominator $\left|\Gamma(\mathrm{p}_i)\cup\Gamma_{P\in\Gamma(c_j)}(\mathrm{P})\right|$ represent the penalty factor on popular products, which will result in appropriate reduction of those popular products' recommendation power, and the numerator indicate the sum of domain similarity of common neighbours between $\Gamma(\mathrm{p}_i)$ and $\Gamma_{P\in\Gamma(c_j)}(\mathrm{P})$. $\mathrm{p}'$ is the collection that each note in it has the common neighbour with $\mathrm{p}_i$, then we weight each common neighbour between note $\mathrm{p}_i$ and note $\mathrm{p}_m$ in $\mathrm{p}'$ collection with domain similarity $sim(\mathrm{p}_i,\mathrm{p}_m)$. So the Jaccard's Coefficient based on domain $DS\left(\mathrm{p}_i,c_j\right)$ measures the weighted sum of domain similarity $sim(\mathrm{p}_i,\mathrm{p}_m)$ compared to the number of nodes that are either $\mathrm{p}_i$ 's or $c_j$ 's neighbours', instead of employ equal-weight to count the number of common neighbours, and then compare it to the number of nodes that are either $\mathrm{p}_i$ 's or $c_j$ 's neighbours'.

According to *D*B_Jaccard's Coefficient , we calculate the probability of links between target customer note and nodes with the exception of the products purchased by the target customer. In product recommendations stage, we descend probability of the links, and recommended the top-n products to the corresponding customers. Under this circumstance, in above case, the possibility of recommending product $p_2$ to customer $\mathrm{c}_1$ is larger than the product $p_1$, leading the result of the recommendation meet with customers' requirements.

In conclusion, both users' past purchasing records and the categories of the purchased product are important part of the input data, which is beneficial to improve the accuracy of the recommendation.

## 4. An Empirical Study

In order to test the accuracy of the algorithm, we used a retail sales dataset from FoodMart. This dataset covers 2 years of 10281 customers' transactions, involving 1560 products and 269720 transactions. To evaluate recommendation performance, we adopted the top-N recommendation task. For each customer we recommended the top-N products that he/she had not purchased previously ranked by linkage measures based on domain knowledge. For comparison purposes, we included link prediction approach of Jaccard's Coefficient in bipartite network. In fact, customers are usually concerned only with the top part of the recommendation list, so we choose the method evaluating quality of the recommendations that consider the number of a customer's relevant products ranked in the top-N places. This paper employ precision, recall, and F-score to measure the recommendation methods' performance in top 10 recommendations, which are the standard and popular metrics.

In data preprocessing phase, we firstly delete the repeat records of purchasing transaction in original data set. The average number of products most customers purchased is below 50, which is too small. This paper consider that in practical application, supermarket usually promote the hot-sale products by placed them in the mass customer passing position. So it has very little value for the users because popular products are easy to find without a recommender system. In this study, we recommend to the customers whose number of transaction records is more than 50. After data preprocessing, the remaining 970 customers are our recommendation target customer. Finally we get a bipartite network including 970 customer notes, 1560 product notes and 86692 links.

To evaluate recommendation algorithms, the data is randomly divided into two parts: the training set contains 80% of the data and the other 10% as the probe. The training set is treated as known information, while no information from the probe set is allowed to be used for recommendation.

This part compared link prediction approach of B_Jaccard's Coefficient with popular recommendation

Table 1recommendation result: algorithm performance measures

| Algorithm | Precision | Recall | F Measure |
|---|---|---|---|
| B_Jaccard's Coefficient | 0.01116 | 0.0134 | 0.0124 |
| DB_Jaccard's Coefficient | 0.01103 | 0.0318 | 0.0163 |

The recommendation quality measures for both linkage measures are presented in Table 1. We can see that DB_Jaccard's Coefficient obtained the better performance than B_Jaccard's Coefficient by both Recall and F Measure on our data set. Experimental results in this paper strongly suggest using the notes attribute (product domain knowledge in customer-product bipartite network) can improve the quality of recommendations list. It confirms that, in the product purchasing application, the category of a product has a strong influence on users' purchasing.

## 5. Conclusions and Future Directions

In this paper, we propose a new link prediction algorithm combining topological features in customer-product bipartite network and domain-based semantic similarities between products to make recommendation for customers. The proposed algorithm is extended from a previously proposed algorithm [5]. It has been tested on one published retail transactions data. The proposed algorithm inherits the main feature of link prediction approach, which is that it predict links by taking topological features into account from bipartite network data, what's more, result indicates that by integrating domain-based similarity knowledge into link prediction recommendation process, the proposed algorithm outperforms the previously algorithm. Not only a higher accuracy has been achieved, but the proposed algorithm also significantly improves consumer loyalty because recommendation are more in line with customer needs, and promote the sales.

We hope to employ more linkage measures approaches to study recommendation problem, including the following categories [27]: 1.neighbor-based linkage measures: CN, Salton, Sørensen, Adamic/Adar, RA, PA, Hub promoted, Hub depressed, Leicht-Holme-Newman 2.path-based linkage measures: SimRank, Matrix forest index, local random walk etc. Future work may include following aspect, first of all, we will focus on how to preferably incorporate domain knowledge in certain area with some other link prediction approaches, so as to making a better prediction; secondly, we will pay more attention on other industries, such as online e-commence, research co-author network etc.

## Acknowledgements

We gratefully thank the anonymous reviewers for their constructive feedback. We appreciated Professor Yong Shi's assistance in this paper. This work was partially supported by National Natural Science Foundation of China (Grant No. 71071151).

## References

1. J. B. Schafer, J. Konstan, and J. Riedi. E-Commerce Recommendation Applications. Data Mining and Knowledge Discovery, 5(1) (2001) 115-153.

2. X. Zhang, Y. Li, Use of collaborative recommendations for web search: an exploratory user study, J. Inform. Sci. 34 (2) (2008) 145–161
3. J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1) (2004) 5-53.
4. Zi-Ke Zhang, Tao Zhou and Yi-Cheng Zhang, Personalized Recommendation via Integrated Diffusion on User-Item Tag Tripartite Graphs. Physica A: Statistical Mechanics and its Applications (Impact Factor: 1.68). 01/2009;
5. P.-Y. Chen, S.-Y. Wu, J. Yoon, The impact of online Recommendations and Consumer Feedback on Sales, In: Proceedings of the 25th International Conference on Information Systems, 2004, 711-724.
6. Bingjing Cai, Haiying Wang, Huiru Zheng, Hui Wang. Integrating domain similarity to improve protein complexes identification in TAP-MS data. From IEEE International Conference on Bioinformatics and Biomedicine. Philadelphia, PA, USA. 4-7 October 2012
7. A.-L. Barabási, R. Albert, Emergence of scaling in random network, Science 286 (1999) 509-512.
8. G. Palla, I. Derenyi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005)814-818.
9. G. Palla, A.-L. Barabási, T. Vicsek, Quantifying social group evolution, Nature 446 (2007) 664-667.
10. M. Medo, Distance-dependent connectivity: Yet another approach to the small-world phenomenon, Physica A 360 (2006) 617-628
11. Q. Ou, Y.-D. Jin, T. Zhou, B.-H. Wang, B.-Q. Yin, Power-law strength-degree correlation from resource-allocation dynamics on weighted networks, Physical Review E 75 (2007) 021102.
12. M. Blattner, Y.-C. Zhang, S. Maslov, Exploring an opinion network for taste prediction: An empirical study, Physica A 373 (2007) 753758.
13. D. Liben-Nowell, J.M. Kleinberg, The link-prediction problem for social networks, Journal of the American Society for Information Science and Technology 58 (2007) 1019-1031.
14. T. Zhou, L.-Y. Lu, Y.-C. Zhang, Predicting Missing Links via Local Information, 2009, arXiv: 0901.0553.
15. Zan Huang, Daniel D. Zeng, Hsinchun Chen. Analyzing Consumer-Product Graphs: Empirical Findings and Applications in Recommender Systems. Management science. Vol. 53, No. 7, July 2007, pp. 1146–1164.
16. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin De La Société Vaudoise Des Sciences Naturelles. 1901,37:547-579.
17. Lingling Zhang, Caifeng Hu, Quan Chen, Yibing Chen,Yong Shi. Domain Knowledge Based Personalized Recommendation Model and Its Application in Cross-selling. International Conference on Computational Science, ICCS. Procedia Computer Science 9 (2012) 1314 – 323.
18. Yuh-Jen Chen, Hui-Chuan Chu, Yuh-Min Chen, Chung-Yueh Chao. Adapting domain ontology for personalized knowledge search and recommendation. Information & Management. 50 (2013) 285–303
19. K. Tso and L. Schmidt-Thieme. Attribute-aware collaborative filtering. In: Proceedings of 29th Annual Conference of the German Classification Society, (Magdeburg, Germany, 2005).
20. Bingjing Cai, Haiying Wang, Huiru Zheng, Hui Wang. Integrating domain similarity to improve protein complexes identification in TAP-MS data. IEEE International Conference on Bioinformatics and Biomedicine .Philadelphia, PA, USA. 4-7 October 2012
21. Leacock, C. and M. Chodorow (1998). "Combining local context and WordNet similarity for word sense identification." WordNet: An electronic lexical database 49(2): 265-283.
22. Ganesan, P., H. Garcia-Molina and J. Widom (2003). "Exploiting hierarchical domain structure to compute similarity." ACM Transactions on Information Systems (TOIS) 21(1): 64-93.
23. Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. 14th International Joint Conferenceon Artificial Intelligence. California: 448-453.
24. Resnik, P. (1999). "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language." Journal of Artificial Intelligence Research: 55-130.
25. Formica, A. (2008). "Concept similarity in Formal Concept Analysis: An information content approach." Knowledge-Based Systems 21(1):
26. Pirró, G. (2009). "A semantic similarity metric combining features and intrinsic information content." Data and Knowledge Engineering 68(11): 1289-1308.80-87.
27. Linyuan Lv, Matus Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang，Tao Zhou.recommender systems. Physics Reports，2012，519:1-49