Influential Seed Items Recommendation

Qi Liu[1], Biao Xiang[1],Enhong Chen[1], Yong Ge[2], Hui Xiong[2], Tengfei Bao[1], Yi Zheng[1]

[1]School of Computer Science and Technology, University of Science and Technology of China
E-mail: {feiniaol,bxiang,tfbao92,xiaoe}@mail.ustc.edu.cn, cheneh@ustc.edu.cn

[2]Rutgers Business School, Rutgers University
E-mail: yongge@pegasus.rutgers.edu, hxiong@rutgers.edu

ABSTRACT

In this paper, we present a systematic perspective study on choosing and evaluating the initial seed items that will be recommended to the cold start users. We first construct an item consumption correlation network to capture the existing users' general consumption behaviors. Then, we formalize initial items recommendation as the influential seed set selection problem. Along this line, we present several methods, each of which selects seed items according to different rules. Finally, the experimental results on two real-world data sets verify that with different seed items, the users' consumption numbers will be quite different. Meanwhile, the results also provide many deep insights into these selection methods and their recommended seed items.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms

Algorithms, Experimentation

Keywords

Seed Items, Item Network, Influential, Popularity

1. INTRODUCTION

Whenever a fresh user comes, there is no or few personal information available, thus how to make recommendations for these cold start users becomes a huge and urgent problem [2]. Actually, many techniques have been proposed to address this kind of problem [2, 3, 6, 7, 8, 10, 11]. For example, one naive method is to recommend the most popular items. Moreover, some methods alleviate the cold start problem by understanding users with their input information [6, 8] or leveraging the meta data of items [11]. Since the user may not want to input her real information for some reasons (e.g., privacy issue) and the meta data is not always available, as an alternative, some other methods try to ask the user to provide a set of ratings for their elaborately selected *seed items* (i.e., the first several items recommended to a cold start user), so as to collect the user's explicit rating preferences [2, 3, 7, 10] with a few interventions.

However, these related techniques focus on the way of choosing representative items (e.g., with high rating variance or entropy) to predict the preferences of the new user for each specific item. In some real applications, in contrast to this individual perspective strategy, the commercial systems often directly offer some discounted or free items to the new users so as to elicit them to consume as many items as possible. Thus, the profit of the system will increase.

To that end, in the following, we mainly focus on the way of recommending and evaluating seed items from this systematic and marketing perspective. Specifically, we aim to automatically find out the seed items that can bring in more consumptions (i.e., *influential* seed items), without any extra interventions from the users. Along this line, we first construct a weighted and directed item network for capturing the users' general consumption behaviors and the items' consumption correlations. Then, the problem can be formalized as selecting the most influential seed items from this network. Next, we propose several methods (e.g., PageRank based methods), each of which selects influential seed items according to different rules and capturing different information. Finally, we evaluate the presented methods on two real-world data sets Flixster and Douban. The experimental results verify that with different seed items, the users' consumption numbers will be quite different. They also demonstrate that just recommending popular items is not a very effective way to help the system understand and attract users, and better results can be achieved by the algorithms exploiting items' influence or correlations. The main contributions of this paper can be summarized as follows.

- We propose the idea of selecting influential seed items from the systematic and marketing perspective for dealing with the cold start user problem. Along this line, we formalize this problem as the seed items selection for an item network.
- We present many possible methods for seed items selection. The effectiveness of these methods is evaluated on two real-world data sets. The experimental results also provide many deep insights into these selection methods and their recommended seed items.

2. PROBLEM FORMULATION

In the following, we use $U = \{U_1, U_2, ..., U_M\}$ and $I = \{I_1, I_2, ..., I_N\}$ to represent the set of users and the set of items, respectively.

In real applications, when a user registers into a system (e.g., an online movie theater), the system usually offers

RecSys'12, September 9–13, 2012, Dublin, Ireland.

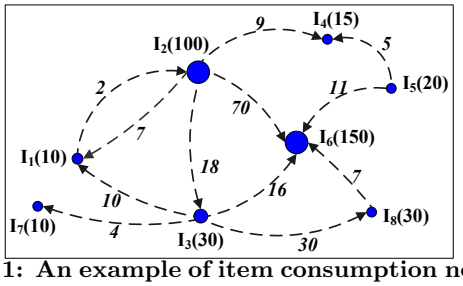

Figure 1: An example of item consumption network.

this cold start user some discounted or free items (e.g., some free movies to watch). There are possibly two kinds of tasks for these free items, the first one is to find out each user's unique preferences (individual perspective), and the other is to attract the user for more consumptions (systematic perspective). Having said, current researches mainly focus on the way of fulfilling the the first task, to our best knowledge, the second aim has so far been overlooked. Since the profit of the system goes directly with the number of users' consumptions, to that end, we focus on the seed item recommendation problem from the systematic perspective.

Along this line, there should be mainly two characteristics of the seed items. At first, they should be *popular* rather than representative. This means we should select the items that are generally enjoyed by the previous users so as to get a potential high acceptance rate. Secondly, they should be *influential*, which means the seed item should have the ability to bring in future consumptions. Please note that, in our scenario, influence/influential is more likely used for measuring the ability of items' consumption correlations, and this is a little bit different from that used in social networks [1].

Then, the problem is how to describe items based on the above two characteristics. Straightforwardly, they can be well represented by item consumption network, as shown by the example in Figure 1. In this network $G = \{I, E\}$, the number in each bracket () is the number of users consumed this specific item, and if there are m users who consumed/enjoyed item I_i before I_j, then we add a directed edge (E_{ij}) and the weight (W_{ij}) should be m. In this way, we can formalize the recommendation problem as selecting the set of seed items from the item consumption network.

3. SELECTION METHODS

In this section, we describe several methods for finding the seed items from the item consumption network.

Popularity. The Popularity method presents items ordered by the number of consumptions that they have been given. It is equivalent to ordering by the probability that a user has enjoyed the item. Popularity is a very straightforward method, and is also easy for calculation.

WDegree. However, the most popular items may not be the most influential. Thus, an alternative approach is to chooses items according to their direct influence abilities, i.e., the number of following consumptions induced by the chosen item. Formally, the WDegree value for each item can be defined as: $WDegree(I_i) = \sum_{j\in[1,N]} W_{ij}$.

Let's take the network in Figure 1 as an example, and suppose we want to select one seed item. According to Popularity, we will choose I_6, which has been consumed 150 times. Instead, according to WDegree, we will select item I_2. We can see that, these two methods directly capture each of the two characteristics of the seed items, respectively.

Intuitively, for better describing each item, the two characteristics should be made use of simultaneously. The idea of PageRank can be well fitted to address this issue. PageRank, introduced by Page et al.[9], was first used for objectively measuring the importance (authority) of web pages.

Though we take advantage of the idea of PageRank, there are still some differences that should be noticed. In PageRank, the *quality* value for each web page is computed recursively by combining the values of the pages that link to this specific page. Thus, a page that is linked to by many pages with high PageRank receives a high rank itself. In contrast, in our situation, the out-edges are used for measuring each item's importance, and if an item links to many items it should receive a high rank value. This can be summarized as: The original PageRank algorithm is used to find out where is the information going, in contrast we want to figure out where is the information from. In the following, based on the idea of PageRank, we illustrate two related methods that can be used for seed set selection.

SPageRank. This is short for Simple PageRank. In this method, each item shares the same initial PageRank value and the same decay factor d. For each item I_i, its PageRank value can be formalized as the following equation:

$$\begin{cases} PR(I_i)^{(0)} = \frac{1}{N} \\ PR(I_i)^{(s+1)} = d \sum\limits_{j,W_{ij}>0} \frac{PR(I_j)^{(s)}}{InDe(I_j)} + (1-d)PR(I_i)^{(0)} \end{cases}$$

where N is the number of items and $InDe(I_j)$ is the number of edges link to I_j (in-degree). $PR(I_i)^{(s)}$ is the PageRank value for I_i after s (e.g.,30) steps iteration.

WPageRank. In contrast to Simple PageRank, weighted PageRank takes the observed consumptions as weight to learn the PageRank values. For WPageRank, item I_i's PageRank value can be formalized as the following equation:

$$\begin{cases} PR(I_i)^{(0)} = \frac{Popularity(I_i)}{\sum_j Popularity(I_j)} \\ PR(I_i)^{(s+1)} = d_i \sum\limits_{j,W_{ij}>0} \frac{W_{ij}PR(I_j)^{(s)}}{InWe(I_j)} + (1-d_i)PR(I_i)^{(0)} \end{cases}$$

where $d_i = \frac{max(W_{ij})}{Popularity(I_i)}$, is the observed decay factor of I_i, and $InWe(I_j)$ is the total weight of edges link to I_j.

Circuit. This is a social influence model based on electrical circuit theory to simulate the information propagation process [5] in social networks, which we proposed in [12]. Different from the previous methods which do not consider the possible influence overlaps between seed items and select all the items simultaneously, the Circuit method identifies the independent influence of each item and selects the seeds one by one following a greedy strategy. Specifically, the number of consumptions (popularity) on each item is used for measuring the probability that an event happens on the given item, and the weight of each edge stands for the probability that the information of this event may propagate through this edge. In this way, the seed item set can be selected by solving the social influence maximization problem [5], for more detailed information please refer to [12].

4. EXPERIMENTAL RESULTS

All the experiments were performed on two real-world rating (in the 1-to-5 scale) data sets: Flixster [4] and Douban [1], since the rating records in both of them last for more than 5 years and there are strong correlations between rating orders and the users' consumption (watch) orders [2]. The detailed information is described in Table 1.

[1]We collected from the douban.com.

[2]There may be a few exceptions.

Table 1: The description of two data sets.

Data Set	Domain	#User(M)	#Item(N)	#Record	#Edge(|E|)
Flixster	Movie	132,774	28,900	1,928,610	13,540,172
DouBan	Book	25,660	212,647	2,545,054	18,037,755

Since we focus on the positive rating records, we removed the records with rating value lower than 3, and removed the users who gave more than 500 ratings because the correlations between their ratings will be very weak. In Figure 2, we take Flixster data as an example and show the power law distributions of the items' popularity and weighted out-degree. From Figure 2(c) we can see that, though positive correlations between each item's popularity and it's out-degree can be observed, there are still some fluctuations. Thus, most popular items may not be most influential and vice versa.

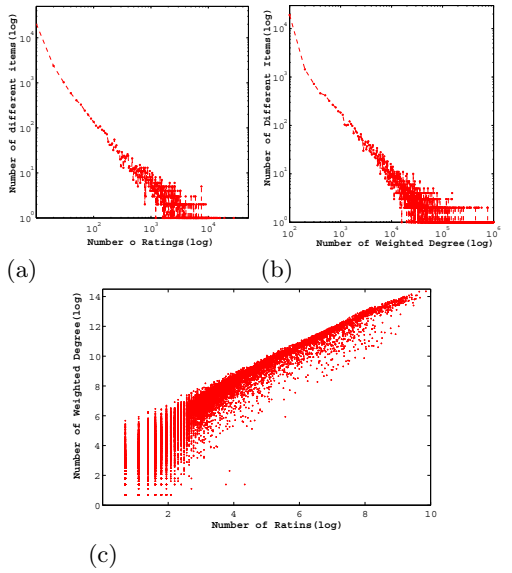

Figure 2: The distributions of Flixster items' popularity (a), weighted out-degree (b) and the relationships between popularity and out-degree (c).

4.1 Performance Comparison

In this section, we present a performance comparison of both effectiveness and efficiency between each method. We fix d equal to 0.85 for PageRank based methods [3].

Effectiveness. To evaluate the selected seed items and the effectiveness of each algorithm, we use *Average-Consumption Number-After Given Item* (AveNum) as a metric, which directly measures how many items will be consumed by the test user after the seed item(s). It is formally defined as:

$$AveNum(k) = \frac{1}{|U^k|} \sum_{U_i \in U^k} \frac{\sum_{j \in [1,k]} RN_{U_i S_j}}{k}, \quad k > 0 \qquad (1)$$

Where U^k is the set of users in the test set who have consumed k items from the selected item set S. Thus, $RN_{U_i S_j}$ is the number of U_i's consumptions after she consumed seed item S_j. If $k = 0$, then we consider all of U_i's consumptions(this value can be also viewed as baseline). In the experiments, each time 4/5 users and their ratings are used for training, i.e., constructing the consumption network and selecting the seed items S, and then the AveNum value of the remaining 1/5 users are computed for testing. At last, we give the average result of these five testing splits.

We run each method and use their selected Top-10 items as seeds to evaluate the performances. Figure 3 illustrates

[3]Different values of d contribute little impact on the results.

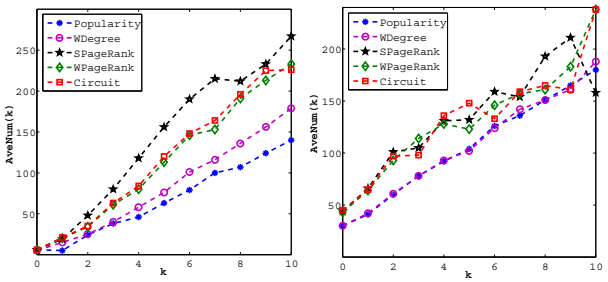

Figure 3: The effectiveness comparison.

Table 2: Comparison of the execution time.

Flixster		Douban	
Method	Time(sec.)	Method	Time(sec.)
Popularity	0.015	Popularity	0.58
WDegree	0.075	WDegree	0.93
SPageRank	32.1	SPageRank	56.40
WPageRank	34.8	WPageRank	58.75
Circuit	490	Circuit	470

the AveNum result that we get by the 5 methods with respect to different number of consumed seeds (k). Generally, the more seed items the test users have consumed, the more items they will consume in the future. Since SPageRank, WPageRank and Circuit methods capture both of the two characteristics of seed items, they usually perform better than Popularity and WDegree. However, the WPageRank and Circuit methods, which perform similar with each other, do not perform better than SPageRank, and we think the reasons are the following: Different from SPageRank which only exploits the structure of the item graph, WPageRank and Circuit also rely on the exact observations, and based on these observations, the importance of items' popularity is manually raised or reduced. Thus, these two methods' performance will be impacted if there exists bias in observations, and this manually weight adjustment may break the balance between popularity and influence.

Efficiency. All the methods are performed on the same platform, and Table 2 shows their average execution time [4]. Without a surprise, on both data sets Popularity and WDegree cost the least time due to their simplicity, and since Circuit model has to choose the seed items one by one following a greedy algorithm, it performs the worst.

In summary, Popularity and WDegree are easy for implementation, but this also brings in their poor performance. Considering both effectiveness and efficiency, SPageRank seems to be the most suitable seed items selection method. However, many more deep understandings should be provided before we make this conclusion.

4.2 Understanding Selections

In this section, we give an analysis of the selection results and they provide more insights into each method.

At first, Figure 4 shows the Jaccard similarity coefficient of the seed items (Top-10 for each data split) chosen by the selection methods. We can see that similar coefficient results can be observed from Flixster and Douban. First, the result got by SPageRank is most different from others. Second, since Popularity and WDegree capture different characters of the items, their choosing items are also different from each other. Third, similar to the observations in Figure 1, the output of WPageRank and Circuit have very

[4]Note that the time consuming on constructing item network is not included, which are similar for each method.

Table 3: Number of Douban users with different k.

Method	$\lvert U^k \rvert$		
	k=1	k=5	k=10
Popularity	4,605	1,314	59
WDegree	4,442	1,267	68
SPageRank	4,842	526	4
WPageRank	4,540	641	2
Circuit	4,558	638	2

strong correlations. We believe the reason lies in the close relation between Circuit and WPageRank [12].

Then, we compare the shortest distances between the seed items. We set the distance among two neighbor vertexes (e.g., I_i,I_j) as $1 - \frac{W_{ij}}{WDegree(I_i)}$, and after running Dijkstra's algorithm for each seed item set we find that the seed items output by Popularity and WDegree are usually very close to each other(with average distance less than 1). In contrast, the seeds output by SPageRank, WPageRank and Circuit often locate in different subnetworks, and thus no path can be found [5]. As example, Table 3 lists the number of Douban users who consumed k (1, 5 and 10) of the seed items with respect to different methods. From Table 3 we can see that when k equals to 1, the $|U^k|$ are similar for each method, and this also reveal that there is no big difference between the average consumption numbers (popularity) of each single seed item selected by different methods. However, much less users consumed all the 10 seed items that are selected by SPageRank, WPageRank and Circuit, respectively.

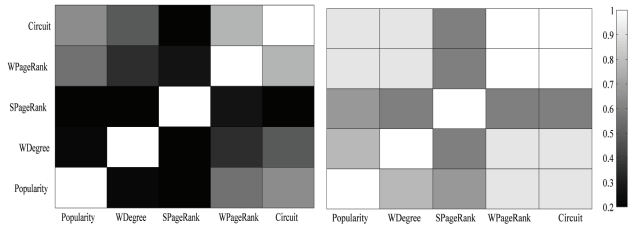

Figure 4: The coefficient of the seed item set selected by each method on Flixster(L) and Douban(R).

Based on the observations from shortest distances and Table 3, we conclude that though the seed items selected by SPageRank, WPageRank and Circuit are also popular items, they are more diverse, e.g., these items may be chosen from different user groups or from different categories. In the future, we plan to find some data sets with available meta data for understanding the selected items more deeply.

5. DISCUSSION

In this section, we analyze the advantages and limitations of current influential seed items recommendation, and show the directions for our future work.

From the experimental results we can see that, by considering the general behavior of users, it is reasonable to find many seed items and recommend them to the cold start users, so as to help the system earn more profit. Furthermore, by exploiting the influence of candidate items, we can get a better recommendation result. However, there are still many limitations of current work. At first, we do not deeply consider the utility of each single user and their personalized acceptance. Second, in this paper, we assume each user's consumption are time ordered and correlated, then make use of the order of users' consumptions to measure item *influence*, and the consumption correlations produced between different lengths of time are weighted equally. This simplicity also leads to some shortages, for example the long time ago consumption may have little contribution on user's current purchase. Third, none of the existing factors that account for the consumption orders are included in current ideal recommendation strategy, such as the items' life cycles.

According to these limitations, there are many directions for our future work. At first, we plan to evaluate our methods in the real-world applications. Meanwhile, figuring out the way of combining advantages from both representative [2, 7, 10] and general item selections for making more reasonable recommendations. More importantly, we plan to consider domain knowledge and constraints (e.g., some meta data and the time factor) into the seed selection process.

[5]In this situation, the influence of items are almost independent from each other, and no need for finding overlaps.

6. CONCLUSION

In this paper, we provide a systematic perspective on choosing and evaluating seed items for cold start user recommendation. We first formalize initial item recommendation as the influential seed set selection problem for the item consumption network. Then, we present several influential seed items selection methods. At last, the performance of these methods are evaluated on two real-world data sets. The experimental study delivers encouraging results, and we hope this work could lead to many future work.

Acknowledgment. This research was partially supported by grants from the Natural Science Foundation of China (61073110, 70890082, 71028002), the National Major Special Science & Technology Projects (2011ZX04016-071), Research Fund for the Doctoral Program of Higher Education of China (20113402110024), the Key Program of National Natural Science Foundation of China (60933013), and the National Science Foundation (NSF, CCF-1018151).

7. REFERENCES

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *ACM SIGKDD'08*., pages 7–15, 2008.

[2] M. Crane. The New User Problem in Collaborative Filtering. *University of Otago*, 2011.

[3] N. Golbandi, Y. Koren, and R. Lempel. Adaptive bootstrapping of recommender systems using decision trees. In *WSDM'11*, pages 595–604, 2011.

[4] M. Jamali, and E. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *RecSys'10*, pages 135–142, 2010.

[5] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *ACM SIGKDD'03*, pages 137–146, 2003.

[6] X.N. Lam, T. Vu, T.D. Le, and A.D. Duong. Addressing cold-start problem in recommendation systems. In *ICUIMC'08*, pages 208–211, 2008.

[7] N. Liu, X. Meng, C. Liu, and Q. Yang. Wisdom of the better few: cold start recommendation via representative based rating elicitation. In *RecSys'11*, pages 37–44, 2011.

[8] A. Nguyen, N. Denos, and C. Berrut. Improving new user recommendations with rule-based induction on cold user data. In *RecSys'07*, pages 121–128, 2007.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. *Stanford InfoLab*, 1999.

[10] A. Rashid, I. Albert, D. Cosley, and etc. Getting to know you: learning new user preferences in recommender systems. In *IUI'02*, pages 127–134, 2002.

[11] A.I. Schein, A. Popescul, L.H. Ungar, and D.M. Pennock. Methods and metrics for cold-start recommendations. In *SIGIR'02*, pages 253–260, 2002.

[12] B. Xiang, E. Chen, Q. Liu, H. Xiong, Y. Yang, and J. Xie. A Social Influence Model Based On Circuit Theory. rxiv preprint arXiv:1205.6024,2012.