Who Will Follow Your Shop? Exploiting Multiple Information Sources in Finding Followers

Liang Wu[2], Alvin Chin[1], Guandong Xu[3], Liang Du[5],
Xia Wang[4], Kangjian Meng[4], Yonggang Guo[4], and Yuanchun Zhou[2]

[1] Xpress Internet Services, Nokia, Beijing, China
[2] Computer Network Information Center, Chinese Academy of Sciences
[3] Advanced Analytics Institute, University of Technology Sydney, Australia
[4] Beijing NoyaXe Technologies Co. Ltd., China
[5] Institute of Software, Chinese Academy of Sciences, China
alvin.chin@nokia.com, {wuliang,zyc}@cnic.cn
guandong.xu@uts.edu.au, xia_s_wang@yahoo.com, duliang@ios.ac.cn

Abstract. WuXianGouXiang is an O2O(offline to online and vice versa)-based mobile application that recommends the nearby coupons and deals for users, by which users can also follow the shops they are interested in. If the potential followers of a shop can be discovered, the merchant's targeted advertising can be more effective and the recommendations for users will also be improved. In this paper, we propose to predict the link relations between users and shops based on the following behavior. In order to better model the characteristics of the shops, we first adopt Topic Modeling to analyze the semantics of their descriptions and then propose a novel approach, named INtent Induced Topic Search (INITS) to update the hidden topics of the shops with and without a description. In addition, we leverage the user logs and search engine results to get the similarity between users and shops. Then we adopt the latent factor model to calculate the similarity between users and shops, in which we use the multiple information sources to regularize the factorization. The experimental results demonstrate that the proposed approach is effective for detecting followers of the shops and the INITS model is useful for shop topic inference.

Keywords: User Behavior, Location Based Services, Matrix Factorization.

1 Introduction

The growth of intelligent mobile devices like smart phones and tablets have contributed to the popularity of location-based services. The convenience of mobile devices with GPS and wireless technologies has now enabled the linking between offline physical location and online access. A hot business model for mobile service providers to create profit is O2O(offline to online and vice versa) commerce, which helps the shops to find new customers by guiding the online users to the real stores in the physical world.

W. Meng et al. (Eds.): DASFAA 2013, Part II, LNCS 7826, pp. 401–415, 2013.

In this paper, we aim to predict the follow relationship between users and shops, which can be formalized as a relational prediction task. However, in real applications there exists many difficulties for modeling the relationships due to the following reasons: 1)*Heavily-tailed distribution:* The distribution of the relationships between users and shops are heavy-tailed. 2)*Sparsity*: The relationships between users and shops are sparse. 3)*Incompleteness*: The information of the merchants is incomplete. The descriptions and even the names for many shops are lost. A basic intuition to tackle these problems is to use some complementary data. Thus, we propose to exploit some useful information sources from multiple heterogeneous domains. In addition, based on the link analysis algorithm Hyperlink-Induced Topic Search (HITS) [4], we propose a novel approach called INtent Induced Topic Search (INITS), which exploits the user intent to predict the hidden topics of a shop by analyzing the contextual information. The contributions of this paper lie in four aspects: 1)Based on our mobile application WuXianGouXiang, we put forward a new problem for discovering the followers of a shop, by which we can offer the better recommendations for users and help merchants improve the effectiveness of targeted advertising. 2)We design a novel approach called INITS to predict the hidden semantics of the shops. By analyzing the contextual information of user behaviors, the correlations between the topics and each of the shops are discovered to tackle the incompleteness of shop description and reveal the in-depth topic distribution of shops. 3)We exploit several useful auxiliary data to tackle the problems of sparsity and cold-start and matrix factorization is adopted to combine the heterogeneous information sources in a unified manner to solve the optimization problem in global scale. 4)We evaluate our method using real-world data collected by WuXianGouXiang.

The remaining sections of the paper are organized as follows: In Section 2, we present some related work. Section 3 describes the application, the motivations and the architecture of our proposed approach. The shop topic modeling and the INITS algorithm are discussed in Section 4 and Section 5. In Section 6, we present the proposed algorithm to predict the followers of a shop. The experiments are shown in Section 7. Section 8 concludes the paper.

2 Related Work

In this section, we will describe some related work on the location-based recommendation, the link prediction in social networks and the collective link prediction.

Location-Based Recommendation: As the mobile applications become popular among users, more information like the location, time and the user behaviors can be collected from the intelligent devices. The recommender system on mobile devices can now provide help based on the contextual information. In [16,17], Zheng et al. propose novel approaches to model the user trajectories and the locations, they recommend the locations and the travel package based on the users' GPS logs. Park et al. propose to recommend the restaurants by taking the user preferences and the location contexts[19]. Bayesian learning is incorporated to compute a score for each shop thus providing recommendations.

Link Prediction in Social Networks: When a user follows a shop, the user can be regarded as a follower or a fan of the shop, which is very similar to the social network like Facebook and Twitter. So there exist many related work which also focus on link prediction in social networks. In [1], Backstrom et al. leverage the training data and the random walk model to predict the relational links. The polarity of the relational link is also considered in [5], where a logistic regression model is used to predict the sign of the edge based on the social sciences. More approaches can be found in [13,8,10] and a survey [7] reviews some of them. Different from these methods, our model is designed for recommending the shop following behavior. Since our problem is obtained from a real world application, the information is very sparse and incomplete, so we leverage other data sources to tackle this while the methods mentioned above model the relationship between the user and the item individually.

Collective Link Prediction: In this paper, we leverage several useful information sources to solve the problem of sparsity, cold-start and data incompleteness. Thus, our work is similar to the multi-relational learning problem. In [11], Singh et al. propose a framework to collectively factorize the matrices, where the same entities in different relationships share the same coordinates in latent spaces. In [14], Xu et al. extend such collective matrix factorization models to a Gaussian processes-based nonparametric Bayesian framework. In [15,18], the tensor is composed to provide both the location recommendation and the activity recommendation simultaneously. Unlike the approaches which tackle multiple tasks at the same time, our approach factors only one matrix, the user-shop matrix, i.e., we only predict the links between users and shops. Another similar work is [6] where Li et al. propose to improve One-Class Collaborative Filtering by exploiting the rich user information. However, the difference in our approach is that besides the user information, we also use the shop information from the search engine, the topic model and the INITS model.

3 Overview

In this section, we present a brief introduction of our mobile application WuXianGouXiang and the architecture of the proposed approach.

3.1 WuXianGouXiang

Nokia Research Center developed a mobile application called WuXianGouXiang in April 2011, which is an O2O-based mobile application, with over 20,000 registered users as of April, 2012. The application can guide the online users to the real shops by offering users the deals and coupons of nearby merchants. The users can download the coupons they like and follow the shops to know their latest deals. WuXianGouXiang has several versions including Java, Symbian and Android, and can be downloaded from http://www.gouxiang.com.

3.2 System Architecture

Figure 1 illustrates the framework of our proposed approach. In order to solve the cold-start and sparsity problems, we propose to leverage heterogeneous information sources from other domains. Firstly, we extract the data of the most frequent behaviors, including downloading coupons and clicking products, from the user logs archived in the application server to measure the user-user similarity. Secondly, we used the shop names as queries and use the aggregated search results to compute the shop-shop similarity. Thirdly, we adopt the topic modeling technique to mine the topics of the shops. To alleviate the data incompleteness, we propose the INITS algorithm to estimate the topic distribution of the shops. The proposed model incorporates the user's intents into the HITS link analysis algorithm, and infers the topics by analyzing the contextual information. Based on the useful information sources, Matrix Factorization is used to decompose them and predict the relational link between the shops and the users. All the above steps will be introduced in the following sections.

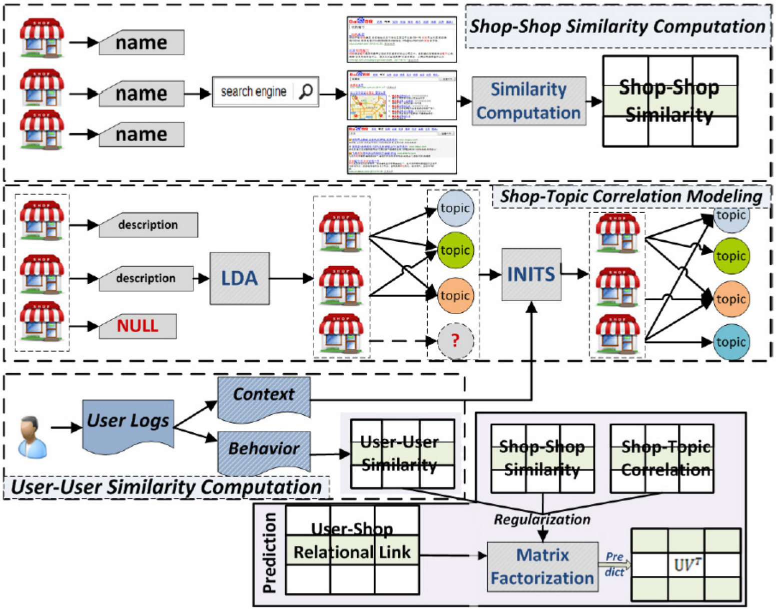

Fig. 1. The framework of the proposed approach on relational link prediction

4 Shop Topic Modeling

When using WuXianGouXiang, the merchants can write a description for their shops. The descriptions contain rich information about the services they provide. The features of the shops are characterized by the words of the descriptions and also by the hidden topics underlying the bag of words. Thus, we adopt the Latent Dirichlet Allocation (LDA)[3] to discover the hidden topics of the descriptions to better model the shops, which is a probabilistic topic model that has been proven to be useful for extracting the latent semantics of documents. Table 1 depicts the high weighted words of some topics. Here we set the number of latent topics k as 10 based on experiments, since some redundant and meaningless topics may be produced given a larger k and some important topics may be neglected when k becomes smaller.

Table 1. The examples of topic-word distribution of experimental results from shop descriptions

Topic 1	ice cream	queen	dairy	instant
Topic 2	italian	spa	luxury	vip
Topic 3	mcdonald	subway	american	hamburger

5 Intent Induced Topic Search

As discussed in Section 4, LDA can discover the topics of the shops by analyzing their descriptions. Since the descriptions are generated by the merchants, there exist many difficulties when analyzing the user generated contents. A major difficulty is that we cannot judge the correctness and quality of the descriptions, many of the merchants even refuse to write one for their shops. If the contents are directly used for the shops' topic inference, noises may be brought in and the features of a shop will be characterized by the shop owner, rather than the customers. Therefore, to circumvent this problem, we propose a novel approach, INtent Induced Topic Search (INITS) to estimate the topic distributions of shops.

Since simply relying on the merchants' descriptions alone will bring in unavoidable bias, we aim to exploit the user ratings to alleviate it. Though there are no explicit ratings between users and shops, the user behaviors can be regarded as implicit feedbacks. In particular, if a shop is visited by many experienced users online, it is highly possible to be a good shop; on the other hand, if a user visits many good shops, she is very likely to be a shopping expert, i.e., there is a mutual enforcement relation between the users and shops. This indicates that the Web page ranking algorithm (HITS) [4] is applicable here. To better study the relationship between users and shops, we construct a bipartite graph, where the users and shops are the two sets of vertices and a directed link is built if a user visits the deals of a shop. As all the edges are pointing from the users to the shops, every shop will get an authority score after applying the HITS algorithm.

Based on the authority scores, we can get the overall popularity of each shop. To further discover the authority scores of each topic of a shop, we propose a novel approach, named INtent Induced Topic Search (INITS) algorithm based on the user contexts. The basic intuitions are as follows: When a user visits a shop online (e.g., downloading the coupons of a shop), she may be interested in a hidden topic the shop bears and the timestamp when the user visits the shop can represent the user's intent, for instance, if a user viewed McDonalds at 11:00 and visited KFC at 11:20, the intents of the user for viewing them are probably the same, which most likely is the user wants to find something to eat. If McDonalds has a latent topic which is "fast food", but KFC does not, then we can possibly infer that KFC is also fast food based on the view and visit.

Theoretically, each shop has k hidden topics. The INITS model assigns the shop's authority score to the hidden topics based on the user's intent. In the HITS algorithm, a shop's authority score is the sum of the hub scores of all the users which have visited it. Given that a user visited a shop h times, then we say the user contributes $\frac{1}{h}$ of its hub score to the shop for each visit. The INITS algorithm assigns the authority score of the shop to each of the shop's latent topic according to the following intuitions: 1): Given a user's log data, when a user visits from one shop to another, if the $\Delta time$ is within a certain threshold, we say that the intents of the user do not change, and the similar topics of the two shops gain a larger share of the $\frac{1}{h}$ of the hub score, e.g., if a user visits McDonalds and KFC successively, and the topic value of "fast food" is similar for both the shops, then the topic of "fast food" of the KFC store will get a higher authority score. 2): If the intents of the user change, i.e., the user may seek for different topics, therefore, the different topics should get a larger share. 3): The topic with a richer value should get a higher authority score, e.g., if a store is famous for the fast food, the user should be more interested in the fast food of it.

We show an example of the INITS model below:

Given that a user first visits shop $shop_i$ and then visits shop $shop_j$, the example shows how much authority score a hidden topic $p \in \{1, \ldots, k\}$ of shop $shop_j$ gains for this visit.

Notations:

k: The number of latent topics.

S^p: The share of the authority score that topic p gets. If the topic is to get a larger share of the authority score, it should have a larger S^p. $\frac{S_p}{\sum_{q=1}^{|k|} S_q}$ is the ratio of the authority score the topic gets.

c_i^p: The correlations between $shop_i$ and the hidden topic p. For the shops which have no descriptions, each topic will be assigned with a default value of $\frac{1}{k}$.

T_i: The time when the user visits shop $shop_i$.

T_{thrd}: The time threshold. If the time between the two visits exceeds the threshold, we say the user's intent changes.

$T_{i \cdot hour}$: The hour of T_i, e.g., if T_i is 13:00, $T_{i \cdot hour}$ equals 13.

hub: The hub score of the user.

$AuthorityGain_j^p$: The gain of the authority score of $shop_j$'s topic p.

N_j: The number of times that the user visits $shop_j$.

Equation 1 computes the dissimilarity of the authority score of topic p between the two shops. For some shops which have a same value of a topic, the dissimilarity will be 0, which will cause the division by 0 in equation 4. Thus, μ is used to avoid this. On the other hand, the S_p cannot be larger than 1, so the μ should be small enough. In this paper, we fix it as 0.01 for simplicity.

$$S_p = |c_i^p - c_j^p| + \mu \tag{1}$$

Equation 2 calculates the time interval between the two visits.

$$\varDelta T = T_j - T_i \tag{2}$$

$$ind = \begin{cases} 1, \varDelta T > T_{thrd} \quad AND \quad T_{i \cdot Hour} \neq T_{j \cdot Hour} \\ -1, \varDelta T < T_{thrd} \quad OR \quad T_{i \cdot Hour} = T_{j \cdot Hour} \end{cases} \tag{3}$$

In Equation 3, the indicator *ind* indicates whether the user intent has changed, that is, if the time span between two visits is within the time threshold, it is highly probable that the user intents are similar. For some users who do not use the application frequently, they may visit one shop and then another shop after one or several days, though the time span exceeds the threshold, their intents may hold. Thus, if the hour of day of the two visits are the same, we say the intents do not change.

$$S_p^{'} = (S_p)^{ind} \times c_j^p \tag{4}$$

$$AuthorityGain_j^p = \frac{1}{N_j} \times hub \times \frac{S_p^{'}}{\sum_{q=1}^{|k|} S_q^{'}} \tag{5}$$

Equation 5 computes the authority score that is assigned to the topic p, where $S_p^{'}$ is the ratio the topic occupies and $\sum_{q=1}^{|k|} S_q^{'}$ is the sum of all the topics' share. Equation 4 computes the share of the topic p. *ind* is used for implementing the second intuition: if the user's intent changes, *ind* will make the share of the more different topics bigger, and vice versa. c_j^p is used to implement Intuition III, which makes the share of the topics with a larger value bigger.

Based on the INITS model, the correlations between the hidden topics and the shops which have no descriptions and low-quality descriptions can also be estimated. In addition, the topics of the shops with descriptions will also be updated based on user's intents. Thus, after updated by INITS, the problems of sparsity and incompleteness are relieved.

6 Followers Discovery

In this section, we will introduce the proposed approach for predicting the followers of shops, which leverages several auxiliary information sources to help the prediction.

6.1 User Information Extraction

As introduced in Section 3.1, the users can download and use the digital coupons in offline shops, view the deals and the shops when using WuXianGouXiang. The behaviors of the users can reflect their habits and preferences. If the users have similar preferences, then they tend to follow similar shops. We extract several common online actions to model the user behaviors, including downloading deals, coupons, get the latest deals and expiring deals. The downloaded information like the price, discount, description, the unique ID of the products, deals and the corresponding shops, are obtained from the logs.

Since we have thousands of deals and coupons in WuXianGouXiang, the Cartesian product of the actions and the deals is very huge, which makes the action vectors of users sparse. Thus, we replace the deals and coupons with their shops. Then we build a vector for each of the users as follows:

$$\boldsymbol{user_i} = < \#action_1(user_i), \ldots, \#action_n(user_i) > \qquad (6)$$

where $\#action_j(user_i)$ is the number of the $action_j$ performed by the $user_i$. An action contains a behavior and the object(shop) of the behavior. The similarity of any two users is calculated by cosine measure:

$$CosSim(\boldsymbol{user_i}, \boldsymbol{user_j}) = \frac{\boldsymbol{user_i} \cdot \boldsymbol{user_j}}{||\boldsymbol{user_i}|| \cdot ||\boldsymbol{user_j}||} \qquad (7)$$

6.2 Shop Information Extraction

The users will follow the shops which can meet their needs and preferences. So the correlations between the shops are useful for predicting the missing values based on the training data. A direct way to measure the similarity between the shops is to use the description information. The descriptions, however, are sparse and incomplete as discussed above. One possible solution is to exploit the external data source. Fortunately, the information from the World Wide Web can be used to measure the similarity, which is similar to the approach in [18].

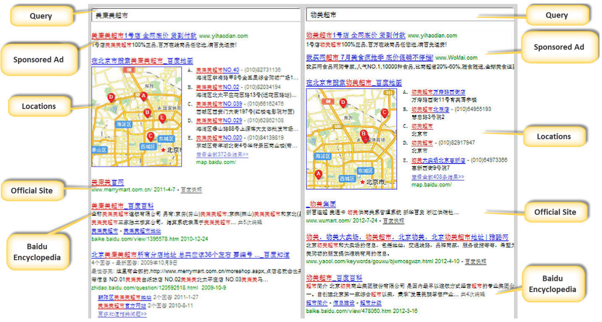

Fig. 2. The results returned by Baidu by using MeiLianMei supermarket and WuMei supermarket as queries

We use the name of the shops as queries and the Chinese search engine Baidu is adopted in this work. The results returned by the search engine are very useful for measuring the similarity. Figure 2 on the previous page illustrates the results when we use the names of two supermarkets as queries. As displayed in the figure, the search results are semi-structured for similar searched entities. That is, for similar kinds of merchants, the search engine has a semi-structured template to display the returned items. The semi-structured results are not proper for modeling the semantics of the shops, but are useful to compute the similarity between shops.

We extract the words of the search results on the first page of each query. The words are then used to describe the shops in a vector space as follows, where $\#word_j(shop_i)$ denotes the word count of $\#word_j$ that appears in $(shop_i)$'s search results.

$$\boldsymbol{shop_i} =< \#word_1(shop_i), \ldots, \#word_n(shop_i) > \tag{8}$$

To filter out the noises from the sponsored advertising, the weighting scheme TF-IDF [2] is adopted to generate a weight for the words of each of the shops. The weighting scheme can avoid the computation to be dominated by the common words.

$$\begin{aligned} IDF(word_j) &= log \frac{D}{|d \in D : word_j \in d|} \\ Weight_{i,j} &= \#word_j(shop_i) \times IDF(word_j) \end{aligned} \tag{9}$$

Equation 9 calculates the weight of $shop_i$'s $word_j$ and we get the weight vector of the shops.

$$\boldsymbol{shop_i^{'}} =< \#weight_1(shop_i), \ldots, \#weight_n(shop_i) > \tag{10}$$

Cosine similarity is adopted here and the similarity between two shops is calculated as follows:

$$CosSim(\boldsymbol{shop_i^{'}}, \boldsymbol{shop_j^{'}}) = \frac{\boldsymbol{shop_i^{'}} \cdot \boldsymbol{shop_j^{'}}}{||\boldsymbol{shop_i^{'}}|| \cdot ||\boldsymbol{shop_j^{'}}||} \tag{11}$$

The shop-shop similarity and user-user similarity information are then leveraged to help the prediction task, which will be introduced in Section 6.3.

6.3 Link Relation Prediction

In order to solve the problems of sparsity, cold-start and data incompleteness, we propose to leverage the useful complementary information sources. As these external sources are all heterogeneous, which cannot be exploited directly, we use matrix factorization to borrow the knowledge by using them as the regularizer.

Definitions:

$F_{m\times n}$: The relationship matrix of the users and the shops, where each entry represents whether the shop is followed by the user. m is the number of users and n is the number of shops.

$U_{m\times k}$: The low rank factor of the users, where k is the number of hidden topics and $k < n$.

$V_{n\times k}$: The low rank factor of the shops.

$C_{m\times m}$: The user-user similarity matrix, which is obtained from the user logs and described in Section 6.1.

I_C: The indicator matrix of $C_{m\times m}$, $I_{C,ij} = 1$ if $C_{i,j}$ is not null.

$M_{n\times n}$: The shop-shop similarity matrix, which is based on the search results and described in Section 6.2.

I_M: The indicator matrix of $M_{m\times m}$, $I_{M,ij} = 1$ if $C_{i,j}$ is not null.

$T_{n\times k}$: The shop-topic matrix based on the descriptions and the INITS model.

$\lambda_1, \lambda_2, \lambda_3, \lambda_4$: The first three parameters are used to control the influence of the complementary information sources and the last controls the regularization over the factorized matrices so as to avoid over-fitting.

The basic intuition of the proposed model is, the users with similar behaviors tend to follow similar shops and the shops with similar topics tend to be followed by similar users. As illustrated in Figure 1, given the relational links $F_{m\times n}$, we decompose it as a product of $U_{m\times k}$ and $V_{n\times k}$. The factorization leverages the auxiliary data sources by sharing the user-topic matrix $U_{m\times k}$ with the user-user matrix $C_{m\times m}$, the shop-topic matrix $V_{n\times k}$ with the shop-shop similarity $M_{n\times n}$ and the shop-topic correlation $T_{n\times k}$. Hence, the objective function is:

$$\begin{aligned} L(U,V) = {} & \frac{1}{2}||F - UV||_F^2 + \frac{\lambda_1}{2}||I_C \circ (C - UU^T)||_F^2 + \\ & \frac{\lambda_2}{2}||I_M \circ (M - VV^T)||_F^2 + \frac{\lambda_3}{2}||V - T||_F^2 + \\ & \frac{\lambda_4}{2}(||U||_F^2 + ||V||_F^2) \end{aligned} \tag{12}$$

where $||\cdot||_F$ is the Frobenius norm and the operator $\circ$ denotes the entry-wise product. The objective function is a non-convex optimization problem. Therefore, we use stochastic gradient descent(SGD) to get the local optimal solution. The gradients (denoted as $\bigtriangledown$) for U and V are as follows:

$$\begin{aligned} & \bigtriangledown_U L = (UV^T - F)V + 2\lambda_1[I_C \circ (UU^T - C)]U + \lambda_4 U \\ & \bigtriangledown_V L = (UV^T - F)^T U + 2\lambda_2[I_M \circ (VV^T - M)]V + \\ & \lambda_3(V - T) + \lambda_4 V \end{aligned} \tag{13}$$

7 Experiments

7.1 Dataset

We conducted experiments with the user logs of WuXianGouXiang from September 2011 to March 2012. We obtained a dataset from the server which contains 998 shops and 681 users.

7.2 Evaluation

In order to measure the accuracy of the prediction, we use two methods to evaluate the recommendation performances. The first one is Mean Absolute Error (MAE):

$$MAE = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} |f_{i,j} - f_{i,j}^{p}|}{m \times n} \tag{14}$$

where m and n are the number of the users and the shops, $f_{i,j}$ is the ground truth of whether user i follows shop j and $f_{i,j}^{p}$ is the predicted result. Noticeably, for the predicted results we transform the values which are larger than one as one and transform the values which are less than zero as zero, since the correlations between the users and the shops cannot be negative and will be one at most. Thus, a smaller MAE score means better prediction performance.

Another measure method we employ is the normalized discounted cumulative gain(nDCG)[9]. This measure is useful for computing the quality of search engines as it considers both the returned contents and the rank of the results. To evaluate the quality of ranking list, we rank the shops for each of the users based on the online visit behavior to get the ground truth. That is, given a shop and a user, if the shop is followed by the user, the shop is relevant for the user. The more times the user visits the shop online, the higher the shop ranks for the user. When testing our proposed approach, we generate the recommended list for each user based on the correlations between the user and the shops in UV^T. A higher $nDCG$ represents a better ranking result.

7.3 Settings

In order to investigate the effectiveness of the auxiliary information sources, we experiment on the following methods:

1)*IM*: The proposed integration method which is based on the INITS model.

2)*WU, WS, WT*: The methods that use all the information sources except the user-user similarity(WU), the shop-shop similarity(WS) or the shop-topic information(WT).

3)*CF*: The Collaborative Filtering method which only uses the training data to predict the missing values. Low-rank matrix factorization [12] is adopted to act as the baseline.

To test the impact of the INITS model, we perform another experiment using the following methods:

1)*INITS*: The integrating method which uses the INITS model to generate the shop-topic information

2)*HITS*: The integrating method which uses the HITS model to generate the shop-topic information, where the HITS model computes an authority score for each of the shops. For the shops without a proper description, the authority score is divided equally.

For each experiment, we repeat five rounds by randomly choosing different entries of matrix F as the training data and the rest as the testing data. The average value of MAE and nDCG are used to measure the performance.

7.4 Experimental Results

Figure 3 illustrates the experimental results of the different methods introduced above based on Mean Absolute Error(MAE). It can be observed that the best performance is achieved by our proposed approach. Thus, we can say that the model which combines the useful auxiliary information sources performs better. When we use the WS model which ignores the shop-shop similarity, the experimental result is closest to the best performance, which means the shop-shop similarity contributes the least to the prediction task. This may be caused by the poor quality of the search results: 1) The search results often contain advertisements of the merchants' competitors, which may bring in some noise when compared with other shops and 2) for the shops which are not so famous, the search engine returns very few results. This makes the feature vector of the shop very sparse and the shops which are not famous and are quite different will be judged to be similar. Though some noises may be taken from the search results, it is still proven to be helpful (about 0.04% improvement). When we use the WU model which ignores user-user similarity and the WT model which ignores shop-topic information, the Mean Absolute Error increases significantly. Based on this we can say that: 1) The users who view the similar deals, download the

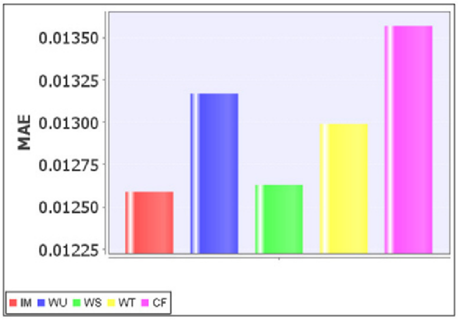

Fig. 3. The overall experimental results on the complementary information sources based on Mean Absolute Error

similar coupons, and will follow similar shops online, and 2) the shops which own similar topics and are viewed at similar times will be followed by the similar group of users, as the INITS model is based on the contextual information of the users.

Table 2. The overall experimental results on the complementary information sources based on normalized discounted cumulative gain

	nDCG[5]	Ratio
IM	0.6261	0.0%
WU	0.5828	7.43%
WS	0.5957	5.10%
WT	0.5607	11.66%
CF	0.5484	14.17%

Table 2 illustrates the experimental results of different methods based on nDCG[5], i.e. we measure the system performance based on the top five recommended shops. The third column denotes the improvement ratio of the proposed method. The integrating method that uses all the information sources achieves the best result. Similarly, the second best result is achieved when ignoring the shop-shop similarity. An interesting difference between the experimental results of MAE and nDCG[5] is that the shop-topic information is most important for the ranking quality among all the auxiliary data. The result indicates that the users will visit the shops more frequently if the topics of the shops can match the users' needs.

Table 3. The influence of the INITS model on different shops

	no description	description
HITS	0.01297	0.01336
INITS	0.01240	0.01304
Ratio	4.40%	2.40%

Table 3 shows the experimental results in terms of MAE on the shops which have a description and the shops without a description. The first model (HITS) is the integration model which uses the HITS algorithm to update the user-topic information, and the second model uses the INITS algorithm. The ratio is the improvement ratio. We can observe that on both datasets, the INITS model outperforms the baseline. Another observation is that, the INITS model has a higher ratio of improvement on the shops without descriptions than the shops with a description, which proves that the INITS model is useful for inferring the hidden topics and is effective for assigning the score to the sub-topics of an authority. Notice that the shops without descriptions have a lower error rate on average, for they are followed by less users and the relationships are more sparse, which lead to less prediction errors but a low recall.

Figure 4 illustrates the performances of our approach varying the parameters, where λ_1 controls the influence of user-user similarity, λ_2 controls the contribution of the shop-shop similarity to the objective function, λ_3 controls the information source of shop-topic, λ_4 is used to avoid over-fitting. Mean Absolute Error is adopted to measure the error rate. The four parameters are tested individually. When testing one of the parameters, the other three are fixed to be 0.1. The results show that the error rate increases when the parameters are either too large or small.

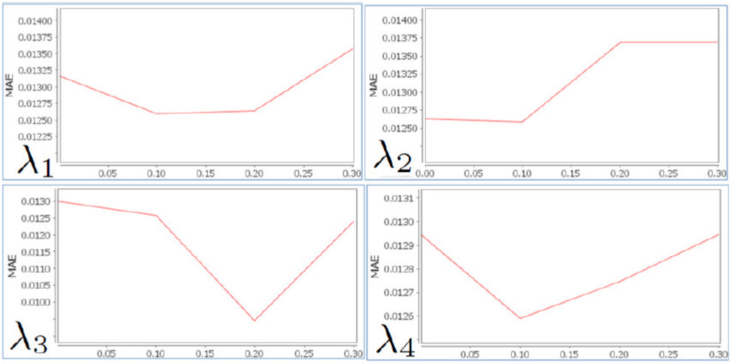

Fig. 4. The impact on Mean Absolute Error of different parameters

8 Conclusion and Future Work

In this paper, we propose a novel approach to predict the link relations between users and shops based on a real world application. The contributions of our work are the following. By surveying the application and analyzing the user logs, we put forward a new problem of discovering the potential followers of a shop. In order to better model the characteristics of the shops, LDA is adopted to process the descriptions offered by the merchants to recover the latent topics underlying the texts. We propose to use several useful auxiliary data sources to tackle the sparsity problem. A novel approach, namely INtent Induced Topic Search (INITS) is introduced to revise the coordinates of the merchants in the latent semantic space. In the future, we will validate our method with other datasets to see and improve the effectiveness of our approach.

Acknowledgments. We are very grateful to the WuXianGouXiang project team, Hao Yang, Minggang Wang, Shenghua Wang, Ke Zhang and Olivia Li. They designed and developed the service from scratch and allowed us to get access to the data and conduct this research work. Alvin Chin and Yuanchun Zhou are the corresponding authors. The work is supported by the Natural Science Foundation of China(NSFC) under Grant No. 61003138.

References

1. Backstrom, L., Leskovec, J.: Supervised random walks: predicting and recommending links in social networks. In: ACM International Conference on Web Search and Data Mining, pp. 635–644 (2011)
2. Baeza-yates, R.A., Ribeiro-neto, B.A.: Modern Information Retrieval (1999)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
4. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46, 604–632 (1999)
5. Leskovec, J., Huttenlocher, D.P., Kleinberg, J.M.: Predicting positive and negative links in online social networks. Computing Research Repository abs/1003.2:641–650 (2010)
6. Li, Y., Hu, J., Zhai, C., Chen, Y.: Improving one-class collaborative filtering by incorporating rich user information. In: International Conference on Information and Knowledge Management, pp. 959–968 (2010)
7. Liben-Nowell, D., Kleinberg, J.M.: The link prediction problem for social networks. In: International Conference on Information and Knowledge Management, pp. 556–559 (2003)
8. Lichtenwalter, R.N., Lussier, J.T., Chawla, N.V.: New perspectives and methods in link prediction. In: Knowledge Discovery and Data Mining (2010)
9. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to information retrieval (2008)
10. Newman, M.E.J.: Clustering and preferential attachment in growing networks. Physical Review E 64 (2001)
11. Singh, A.P., Gordon, G.J.: Relational learning via collective matrix factorization. In: Knowledge Discovery and Data Mining, pp. 650–658 (2008)
12. Srebro, N., Jaakkola, T.: Weighted Low-Rank Approximations. In: International Conference on Machine Learning, pp. 720–727 (2003)
13. Wang, C., Satuluri, V., Parthasarathy, S.: Local probabilistic models for link prediction. In: International Conference on Data Mining, pp. 322–331 (2007)
14. Xu, Z., Kersting, K., Tresp, V.: Multi-Relational Learning with Gaussian Processes. In: International Joint Conference on Artificial Intelligence, pp. 1309–1314 (2009)
15. Zheng, V.W., Cao, B., Zheng, Y., Xie, X., Yang, Q.: Collaborative Filtering Meets Mobile Recommendation: A User-Centered Approach. In: National Conference on Artificial Intelligence (2010)
16. Yoon, H., Zheng, Y., Xie, X., Woo, W.: Smart Itinerary Recommendation Based on User-Generated GPS Trajectories (2010)
17. Zheng, Y., Xie, X.: Learning Travel Recommendations from User-Generated GPS Traces 2, 1–29 (2011)
18. Zheng, V.W., Zheng, Y., Xie, X., Yang, Q.: Collaborative location and activity recommendations with GPS history data. In: World Wide Web (2010)
19. Park, M.-H., Hong, J.-H., Cho, S.-B.: Location-Based Recommendation System Using Bayesian User's Preference Model in Mobile Devices. In: Indulska, J., Ma, J., Yang, L.T., Ungerer, T., Cao, J. (eds.) UIC 2007. LNCS, vol. 4611, pp. 1130–1139. Springer, Heidelberg (2007)