Link Prediction in Microblog Network Using Supervised Learning with Multiple Features

Siyao Han*, Yan Xu
The Information Science Department, Beijing Language and Culture University, Beijing, China.

* Corresponding author. Tel.: +86-18311049766; email: hsy_blcu@163.com
Manuscript submitted January 20, 2015; accepted August 11, 2015.
doi: 10.17706/jcp.11.1.72-82

Abstract: Link prediction (LP) is a fundamental network analysis task. It aims to analyze the existing links and predict the missing or potential relations between users in a social network. It can help users in finding new friends, enhance their loyalties to the web sites and build a healthy social environment. In previous researches, much attention was focused on structure information or node attributes, in order to analyze the global or local properties. Considering the nature of Microblog social network, we proposed a link prediction system combining multiple features from different perspectives, and learn a classifier from these feature subsets to predict the potential links. We train classifiers using SVM, Naïve Bayes, and Random Forest and Logistic Regression algorithms and evaluate them using the microblog network dataset. The results show that our features perform better than the traditional features, and the combination of multiple features can achieve highest accuracy.

Key words: Link prediction, microblog, feature extraction, supervised learning.

1. Introduction

Recently, Social network has been flourishing with increasing influences. Online social network services such like Facebook, Twitter and Microblog, have become more and more popular around the world, offering individuals with similar interests and personalities the possibility of building relationships.

Along with the booming of social network, a considerable amount of attention has been devoted to link mining [1]. In this context, link prediction is one of the most fundamental tasks. Through analyzing the existing links, it can predict missing or potential relationship links between users in a social network, which can help users to find new friends, enhance their loyalties to the web sites and promotes the growth of the social networks [2].

Link problem can be solved under supervised learning framework, by extracting features from the network and building models to identify the missing or potential existing links [3]. In this context, feature extraction is the key problem. In social network (e.g. Microblog network), users' properties and social interaction information contain tremendous clues. So recent years, much work have been focused on solve LP problems using heterogeneous features of social network, such as node features, topological features and some specific features. M. Fire, *et al*. [4] propose some structural (topological) features to identify missing links. J. Valverde-Rebaza, *et al*. [5] try to exploit the behaviors information of Twitter users for link prediction. D. Liben, *et al*. [6] extract aggregated features, proximity features and topological features to solve the LP problem of coauthorship network. D. Wang, *et al.* [7] also take into account the mobility measures. Y. Jia, *et al.* [8] define a retweet similarity to measure the interaction based matrix factorization model for following link prediction. Recently, J. Narasimhan and L. Holder [9] incorporate time features into

LP problem. J. Kim, *et al*. [10] make use of cluster information. And F. Tan, *et al*. [11] reexamine the role of topology feature from the perspective of information theory. These work just consider features from one or two sources. In this paper, we collect information from multiple sources and extract four informative feature subsets (namely, Node feature subset, Topology feature subset, Social feature subset, Voting feature subset) to get more precise and comprehensive results.

Topological features are always some structure similarity indices, which can be categorized into 3 types: local indices, global indices, and quasi-local indices. Much work has been done to find a proper topology similarity index [12], [13]. However, there does not appear to be one best similarity index that is superior in all settings. LP techniques often focus on these structure similarity indices respectively, but rarely combine these signals. In this paper, we combine these standard measures together as the topology feature subset. The result of the combination is better than that of each single measure.

In addition to structure similarity indices, social relation information can also be used to enhance link prediction, in the case of social networks [2], [14], [15]. In micro-blogging application, "following" and "followed" relationships have two opposite directions. Compared with undirected social networks, directed social networks are more informative. So we define some measures for directed microblog network, and also introduce users' category, behavior and content similarity into the final social feature subset.

Due to the sociability property of Microblog network, the neighbors' impacts are also helpful for determination. Z. Huang, *et al*. [16] use these similarity indices in link prediction to enhance collaborative filtering (CF) and solve the sparsity problem. J. Kim, *et al.* [10] make use of cluster information. In our paper, we try to use the CF algorithm to collect the voting feature of neighbors to enhance the link prediction result.

Using these features, link problem can be converted into classification problem. There are several works related to our approach. D. Liben, *et al*. [6] and M. A. Hasan, *et al*. [3] try to solve LP problem of coauthorship network using supervised learning methods. S. Soundarajan, *et al.* [17] also use supervised learning methods to predict directionality of links in directed network. H. Rodrigues Sa' and R. Prudencio [18] use supervised learning methods in weighted networks. In our method, after extracting multiple feature subsets, we use some advanced machine learning algorithms to build models predicting the hidden links.

In this paper, we solve the link prediction problem under supervised learning framework. We extract individual features as the node feature and improve traditional structure similarities measures as topology feature. Based on the nature of microblog network, take the relation circle, user category, behaviors and posted messages into account. Furthermore, we consider the neighbors' impacts and calculate it as voting feature. Supervised learning methods (i.e., Naïve Bayes, SVM, Logistic Regressions and Random Forest algorithms) are applied on combination of these feature subsets (namely, node feature, topology feature, social feature and the voting feature) to find the coefficients, which optimize the proportion of correctly predicted links.

The contributions of this paper are as follows:

1) Link prediction techniques often focus on structure properties (such as Common Neighbors, Jaccard index and many other measures), but rarely combine these signals. We combine basic local structure measures together as a feature subset, rather than use them for make decision independently;
2) In social feature subset, except for the relationship and interaction information, we also refine some structure measures for directed social network and also take the users' category, behavior similarity and the message similarity into account;
3) Besides these common used features, we also collect the neighbor's opinions by calculating the rating score using traditional collaborative filtering method, and make the score as the vote decision of all neighbors;

4) This framework allows combination of multiple features and measurements, and can utilize some advanced machine learning methods to improve final performance.

The remainder of this paper is organized as follows. Section 2 will describe our method in detail. In Section 3, experimental evaluation is presented. And the paper is concluded in Section 4.

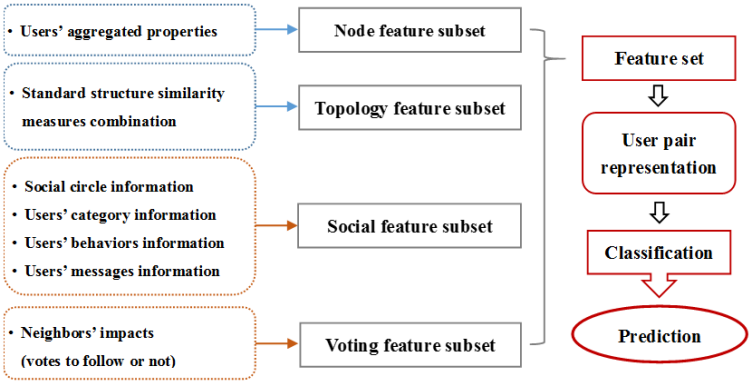

Fig. 1. The framework of our method.

2. Our Method

As illustrated in Fig. 1, given the real microblog network dataset, four informative features (namely, Node features (N), Topology features (T), Social features (S), and Voting features (V)) are extracted to represent each user pair (<user1, user2>). Then we convert link prediction problem to a binary classification problem, using four supervising learning methods (i.e., Naïve Bayes, SVM, Logistic Regressions and Random Forest algorithms) learn classifiers to identify whether the two users should have a following relationship (predict the likelihood of a link existence in the social graph).

2.1. Data Representation

In a normal social network, the link is undirected. Let $G\,(V,\ E)$ represent an undirected network where V is the set of nodes and E is the set of links. In microblog services, participants build own social circles by "following" (subscribing to) other users. Different from some online social networks such as Facebook, the followed user has the option but not the obligation to similarly follow back. In this case, the network is a direct graph, in which the direct edge stands for a following relationship. In this paper, we organize the whole social network as a list of following relationship $U = \{< u_1, u_2 >, < u_1, u_3 >, \ldots < u_i, u_j >, \ldots, < u_{n-1}, u_n >\}$,where $i,\ j = 1 \ldots n$, n is the number of user. Each term comprises a feature vector and a relationship label. The value of label is 1 when u_i following u_j, otherwise is 0. And the feature vector is composed by the four feature subsets.

2.2. Feature Extraction

Extracting an appropriate feature set is the most critical part of any machine learning algorithm. Traditional link prediction methods always consider the vertical, link information and the topological features [4], [6]. In the social network as we study, besides the topological features and the node attributions, the social circle information, the behaviors of users, the contexts users posting and the neighbors' impacts also give clues of whether a link is exist between the two users.

To describe a relationship between two users more comprehensively, we extract individual features as the node feature, and improve traditional structure similarities measures as topology feature. We take the relation circle, user category, behaviors and posted messages into account and propose social feature based on the nature of microblog network. Furthermore, the neighbors' impacts are considered and used as voting

feature. In conclusion, we arrange them as four types of feature subsets: Node features (*N*), Topology features (*T*), Social features (*S*), and Voting features (*V*).

2.3. Node Features

There exist individual attributes that can provide helpful clues for link prediction. But with the limitation of privacy, some attributes only pertain to one vertex (user) in the social network; we cannot get the private profile of a user. Thus some aggregation functions need to be used to describe the corresponding vertexes (users) in a user-pair.

We define a user by aggregating the number of his fans, the number of his followed users, the number of his original messages, the number of his forward messages and the number of his reply. These features can reflect the activity of the user, and it provides a depiction of the user's character.

2.3.1. Topology features

We extract topology features for measuring the similarity between vertices (users) based on the assumption that similar vertices (users) are more likely to share same relations (links) [7]. The similarity indices can be classified into three categories: local indices, global indices, and quasi-local indices. Global indices may lead to higher accuracy, but their computation is very time-consuming and usually infeasible for large-scale networks. Local measures are generally faster but provide lower accuracy. Much work has been done to find a good index. However, there does not appear to be one best similarity index that is superior in all settings [19]. Different measures would have various performances depending on the network under analysis. In our paper, we incorporate these basic structure measures together as a feature set, rather than use them for make decision independently.

The basic structural definition for a vertex $x \in V$ is its neighborhood, which denotes the friends of x. In microblog networks, we treat the following persons of user *x* as the friends of x. Using these neighborhood definitions, we calculate following 9 basic stand measures based on local information [12], and combine them together as the topology feature subset for training the best weight vector. Table 1 shows the definitions of these measures [20].

Table 1. The Standard Similarity Measures

Common Neighbors (CN)	$\boldsymbol{CN(x,y)} = \lvert\boldsymbol{\Gamma(x)} \cap \boldsymbol{\Gamma(y)}\rvert$	*Salton Coefficient (Sal)*	$\boldsymbol{Sal(x,y)} = \frac{\lvert\boldsymbol{\Gamma(x)} \cap \boldsymbol{\Gamma(y)}\rvert}{\sqrt{\lvert\boldsymbol{\Gamma(x)}\rvert \times \lvert\boldsymbol{\Gamma(y)}\rvert}}$
Adamic Adar (AA)	$AA(x,y) = \sum_{z \in \Lambda_{x,y}} \frac{1}{\log \lvert\Gamma(z)\rvert}$	*Sorenson Coefficient (Sor)*	$Sor(x,y) = \frac{2\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\lvert\Gamma(x)\rvert + \lvert\Gamma(y)\rvert}$
Resource Allocation (RA)	$RA(x,y) = \sum_{z \in \Lambda_{x,y}} \frac{1}{\lvert\Gamma(z)\rvert}$	*Hub Depressed Index (HDI)*	$HDI(x,y) = \frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{max\{\lvert\Gamma(x)\rvert, \lvert\Gamma(y)\rvert\}}$
Preferential Attachment (PA)	$PA(x,y) = \lvert\Gamma(x) \times \Gamma(y)\rvert$	*Leicht-Holme-Newman Index(LHN-I)*	$\mathrm{LHN-I}(x,y) = \frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\lvert\Gamma(x)\rvert \times \lvert\Gamma(x)\rvert}$

2.3.2. Social features

In addition to topological similarity indices, social relation information can also enhance link prediction, in the case of social networks [8], [21]. For general social network (such as Facebook), it is undirected. Compared with undirected social networks, directed social networks are more informative. In micro-blogging application, "following" and "followed" relationships have two opposite directions, they

represent the most important behaviors of users. Let $\Gamma out(x) = \{y \mid (x,y) \in E\}$ be outgoing neighbors (followees) and $\Gamma in(x) = \{y \mid (y,x) \in E\}$ be incoming neighbors (fans). Based on the directed microblog network, we define some measures for directed social network.

M. Bilgic etc. [22] propose that object classification can improve the performance of link prediction, so we use the category information, which is obtain from our previous work [23], to get the category similarity. We also take advance of all kinds of messages (original, reply, topic) and calculate the similarities between them as the content features. Furthermore, users' behaviors are also helpful to identity the user pair (<user1, user2>) relationship.

In this feature set, six feature subsets can be considered to describe a user pair from social relation perspective: neighbor overlap, indirect neighbor overlap rates, same category rate, indirect same category rates, behavior overlap rates and content similarity.

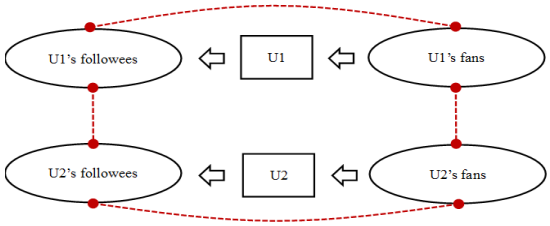

Fig. 2. Direct neighbor relationship.

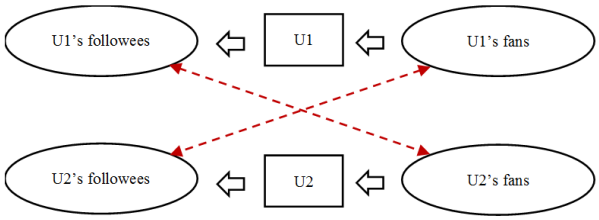

Fig. 3. Indirect neighbor relationship.

2.3.2.1. Neighbor overlap rate

Neighbor overlap rates consist of four parts: followee overlap rate, fans overlap rate, and mutual follow rate of user1 and mutual follow rate of user2. Fig. 2 shows the direct neighbor relationships.

- The overlap rate of u1's followees and u2's followees: SameFolloweeRate = $\frac{2*|\Gamma out(u1) \cap \Gamma out(u2)|}{|\Gamma out(u1)|+|\Gamma out(u2)|}$
- The overlap rate of u1's fans and u2's fans: SameFanRate = $\frac{2*|\Gamma in(u1) \cap \Gamma in(u2)|}{|\Gamma in(u1)|+|\Gamma in(u2)|}$
- The overlap rate of u1's followees and u1's fans: MutualFollowRate (u1) = $\frac{2*|\Gamma in(u1) \cap \Gamma out(u1)|}{|\Gamma in(u1)|+|\Gamma out(u1)|}$
- The overlap rate of u2's followees and u2's fans: MutualFollowRate (u2) = $\frac{2*|\Gamma in(u2) \cap \Gamma out(u2)|}{|\Gamma in(u2)|+|\Gamma out(u2)|}$

2.3.2.2. Indirect neighbor overlap rate

Indirect neighbor overlap rate is used to describe the indirect connection relationship between two users. Different from neighbor overlap rates, indirect user overlap rate is about bidirectional relationship. Fig. 3 shows the direct neighbor relationships.

- How many u1's followees in u2's fans: $\frac{|\Gamma out(u1) \cap \Gamma in(u2)|}{|\Gamma in(u2)|}$
- How many u1's fans in u2's followees: $\frac{|\Gamma in(u1) \cap \Gamma out(u2)|}{|\Gamma out(u2)|}$
- How many u2's followees in u1's fans: $\frac{|\Gamma out(u2) \cap \Gamma in(u2)|}{|\Gamma in(u1)|}$
- How many u2's fans in u1's followees: $\frac{|\Gamma in(u2) \cap \Gamma out(u1)|}{|\Gamma out(u1)|}$

2.3.2.3. Same category rate

Some research has been focused on combing object classification and link prediction[22]. Object classification is not provided with all the links relevant to correct classification and link prediction is not provided all the labels needed for accurate link prediction. Thus, M. Bilgic and his coworkers interleave object classification and link prediction in a collective algorithm.

Users in the same category are likely having a relationship. This feature describes whether two users in a

user pair is the same occupation class (such as IT, Sports, entertainment, etc.), which is obtained from[23]. If the two users belong to the same category, this feature value is set to "1", otherwise, it is "0".

2.3.2.4. Indirectly same category rate

Two users following the users with same category always have similar interests. Indirect same category rate is used to describe the indirect category relationship between two users. Table 2 shows how to get indirectly same category rate in four aspects. Let $C = \{c_0, c_1, \dots, c_i, \dots, c_n\}$, denoting all the categories of users, $C(u)$ be the category of u, $N(c_i, \Gamma\text{out}(u))$ be the number of u's followees of c_i category.

2.3.2.5. Behavior overlap rate

Micro-blogging platform also employs other social-networking models called "replying", "forwarding" and "mentioned" (@), they also represent microblog user behaviors. Accordingly, this feature set includes three behavior overlap rate: the "relaying" overlap rate, "forwarding" overlap rate and "@" overlap rate. Let $R(u)$ be the users who u replying, $F(u)$ be the users who u forwarding, $M(u)$ be the users who u mentioning Table 3 shows the behavior overlap rates.

Table 2. Indirectly Same Category Rate

How many u1's followees are same category with u2	$\frac{N(C(\text{u2}), \Gamma\text{out}(\text{u1}))}{\lvert\Gamma\text{out}(\text{u1})\rvert}$
How many u1's fans are same category with u2	$\frac{N(C(\text{u2}), \Gamma\text{in}(\text{u1}))}{\lvert\Gamma\text{in}(\text{u1})\rvert}$
How many u2's followees are same category with u1	$\frac{N(C(\text{u1}), \Gamma\text{out}(\text{u2}))}{\lvert\Gamma\text{out}(\text{u2})\rvert}$
How many u2's fans same category with u1	$\frac{N(C(\text{u1}), \Gamma\text{in}(\text{u2}))}{\lvert\Gamma\text{in}(\text{u2})\rvert}$

Table 3. Behavior Overlap Rate

The rate u1 and u2 replying the same user	$\frac{2 * \lvert R(\text{u1}) \cap R(\text{u2})\rvert}{\lvert R(\text{u1})\rvert + \lvert R(\text{u2})\rvert}$
The rate u1 and u2 forwarding t same user	$\frac{2 * \lvert F(\text{u1}) \cap F(\text{u2})\rvert}{\lvert F(\text{u1})\rvert + \lvert F(\text{u2})\rvert}$
The rate u1 and u2 mentioning(the same user	$\frac{2 * \lvert M(\text{u1}) \cap M(\text{u2})\rvert}{\lvert M(\text{u1})\rvert + \lvert M(\text{u2})\rvert}$

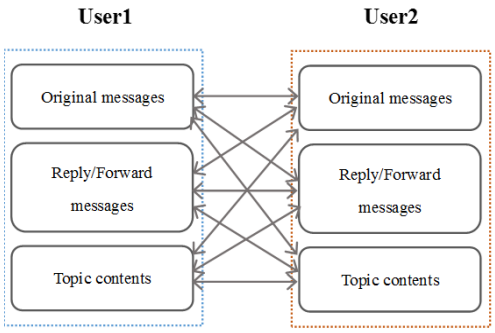

Fig. 4. The content similarities.

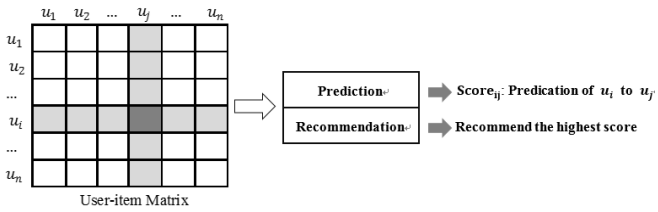

Fig. 5. The process of collaborative filtering.

2.3.2.6. Content similarity

Content features represent the similarities of the interested contents of two users. We obtain them by calculating the similarities between the original messages, reply/forward contents and topic contents of the two users.

Fig. 4 shows how to get user-user pair social content features. The left column shows content features of u1 and right column is those of u2. Lines in the middle connecting texts between u1 and u2 represent the similarity relationships between the two texts. We use text similarity algorithm to calculate the similarities between each pair of contents to get the nine social content features.

2.3.2.7. Voting features

In a social network (e.g., Microblog), besides some individual interests and attributes, users are always impacted by other users, especially their neighbors (the ones have some relations with them). In the link

prediction problem, we can make user of this character, extracting the neighbors' decisions to get better performances in the final result. Collaborative filtering (CF) algorithm in recommendation filed is a well-studied and efficient algorithm. We use the user-based CF algorithm to collect the votes of the neighbors and predict the final score (the probability of user1 following user2).

Fig. 5 shows two processes of collaborative filtering algorithm: rating score predicting and recommendation. The system records users' rating scores, keeps them in the user-item matrix. Traditional user-item matrix records users' rating scores on items, while in this paper user-item matrix is actually a user-user matrix, where every entry represents the relation of the two users. If user_i and user_j has following relationship, Matrix $[i][j]$ is "1". Conversely, the value is "0". CF algorithm predicts the un-rated scores using the rated scores, then recommends the ones with high score to users. In this paper, we only consider the rating score predicting process. We use the predicted rating score as the voting feature.

3. Experiments and Results

3.1. Data Set

All the data in this paper is crawled from Sina microblog, including 37660 users, 1959276 users' following relationship and 1,737,329 user messages.

Because our data is static, we do not know which links are the missing or future links. To evaluate our method, the observed/exiting links, E, is randomly divided into two parts: the training set E^{Tain}, which is treated as known links, while the test set E^{Test}, which is used for testing. Clearly, $E^{\text{Tain}} \cap E^{\text{Test}} = \emptyset$.

After data preprocessing, we randomly select 500 active users in sport, entertainment, IT, real estate categories respectively. A user pair $< u_i, u_j >$ is regarded as a positive sample when u_i follows u_j, otherwise it is a negative sample. In our comparative study, we generate the data set as following steps:

- Generating all the positive user pairs of each selected user.
- Randomly selecting the same number of negative user pairs of each user.
- Combining positive and negative samples.
- Randomly selecting 100,000 samples.

3.2. Experiment Measurements

Precision, recall and f-measure, are the most popular criteria used in classification. In this paper, we solve link prediction problem under supervised learning framework, using several machine learning algorisms train classifiers to classify linked and un-linked user pairs. So in our study, we just use precision, recall, f-measure of class 1(linked user pair) as the measurements of classifiers' performance.

3.3. Experiment Platform

We choose Weka as our experiment platform, which is a widely used data mining and machine-learning tool [24]. As a public platform for data mining, Weka aggregates a mass of machine learning algorithms including data classification, regression, clustering, etc.

We train classifiers using SVM, Naïve Bayes, Logistic Regression and Random Forest algorithms, with 10-cross validation. All the algorithm parameters are the default ones in Weka platform.

3.4. Compare Results of Single Measures and Their Combination

The traditional local structure measures do not appear to be one best similarity index that is superior in all settings [6]. To select a good index for particular network is redundant and dull. In our method, we combine these measures together as a topology feature. Table 4 shows the results of each measure and their combination (topology feature).

Table 4. The Results of Single Measures and the Combination

	Naïve Bayes			SVM			Logistic			Random Forest		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
AA	0.558	0.92	0.695	0.544	0.926	0.685	0.583	0.905	0.709	**0.917**	**0.879**	**0.898**
CN	**0.8**	**0.928**	**0.86**	0.607	0.865	0.714	0.622	0.862	0.723	0.806	0.909	0.854
HDI	0.546	0.134	0.22	0.557	0.879	0.682	0.599	0.867	0.709	0.813	0.905	0.856
Jaccard	0.575	0.901	0.702	0.532	0.927	0.676	0.572	0.899	0.699	0.733	0.898	0.807
LHN-I	0.546	0.076	0.134	0.557	0.879	0.682	0.599	0.867	0.709	0.813	0.905	0.856
PA	0.797	0.917	0.853	0.641	0.859	0.734	**0.656**	**0.858**	**0.743**	0.797	0.917	0.853
RA	0.797	0.924	0.856	0.513	0.921	0.659	0.548	0.889	0.678	0.823	0.877	0.849
Sahon	0.649	0.862	0.74	**0.663**	**0.86**	**0.749**	0.671	0.859	0.753	0.811	0.892	0.849
Sorenson	0.806	0.906	0.853	0.56	0.875	0.683	0.616	0.863	0.719	0.806	0.909	0.854
Topology	**0.815**	**0.916**	**0.862**	**0.737**	**0.871**	**0.799**	**0.801**	**0.894**	**0.845**	**0.957**	**0.914**	**0.935**

Based on microblog network, these measures have distinct performances in these four classifiers. Such as LHN-I, it has a bad result (0.134 of F-measure) using Naïve Bayes (NB), but a good performance (0.856 of F-measure) using Random Forest (RF). Conversely, the combination of them—topology feature (T) always keeps good and steady performances and can far exceed these single measures. Thus, the combination we proposed has indeed improved the performance.

3.5. Compare Results of Each Feature Subset

In addition to the typical similarity measures based on the topology structure and the vertex attribution, we propose two feature subsets for microblog real data, which are called social feature subset and voting feature subset. We apply them on link prediction problem of microblog real data respectively and then combine them together.

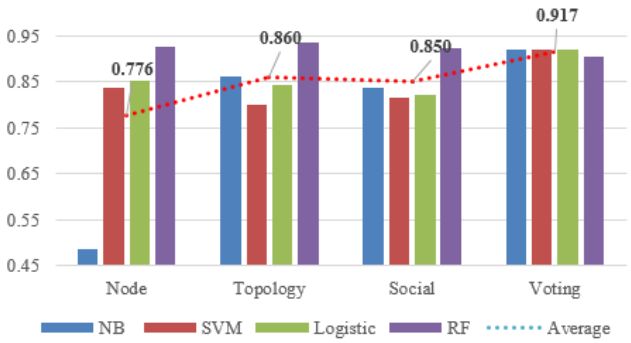

Fig. 6. F-measure values of four feature subsets on different classifiers.

From Fig. 6, voting feature has the best and steadiest performance, with the highest average f-measure value (0.917), followed by topology feature subset (0.86 of f-measure value). Node feature subset has the relative lower f-measure value, but the highest precision in all the four classifiers. The detailed results can be seen in Table 5.

Topology feature subset and social feature subset are composed by heterogeneous features, they have better results under Random Forest classifier, but a relatively poor performances under other three classifiers. Because random forest is good at the heterogeneous feature issues [25]. The strengths of the random forest method include feature selection and consideration of many feature subsets (instead of focusing on just a few features that best separate the training data).

Table 5. The performances of Each Feature Subset in 4 Classifiers

	Naïve Bayes			SVM			Logistic			Random Forest		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Node	**0.958**	0.327	0.487	**0.945**	0.752	0.837	**0.951**	0.773	0.853	**0.971**	0.886	0.927
Topology	0.815	0.916	0.862	0.737	0.871	0.799	0.801	0.894	0.845	**0.957**	**0.914**	**0.935**
Social	0.878	0.802	0.838	0.77	0.866	0.815	0.782	0.864	0.821	0.945	0.906	0.925
Voting	**0.933**	**0.908**	**0.921**	**0.934**	**0.908**	**0.921**	**0.936**	**0.907**	**0.921**	0.916	0.894	0.905

3.6. Compare Results of Combined Feature

We apply topology feature subset and node feature subset respectively on these four classifiers, and then combine social feature subset and voting feature subset sequentially. We can see from Table 6, social feature subset and voting feature subset can enhance both the performances of topology and node feature in all the four classifiers, especially voting feature, which raise the node feature result from 0.487 to 0.927, using Naïve Bayes algorithm. And the combination of four subsets can have the best results. So we can conclude that the features we proposed can improve the performance of link prediction.

Table 6. The Effects of Adding Feature Subsets

	Naïve Bayes			SVM			Logistic			Random forest		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
T	0.815	0.916	0.862	0.737	0.871	0.799	0.801	0.894	0.845	0.957	0.914	**0**.935
T+S	0.881	0.902	0.891	0.838	0.873	0.855	0.869	0.893	0.881	0.969	0.922	0.945
T+V	0.945	0.921	0.933	0.94	0.921	0.93	0.945	0.924	0.934	0.972	0.921	0.946
T+S+V	0.962	0.911	0.935	0.963	0.911	0.936	0.952	0.92	0.936	0.978	0.924	0.95
N	0.958	**0.327**	**0.487**	0.945	0.752	0.837	0.951	0.773	0.853	0.971	0.886	0.927
N+S	0.973	**0.588**	**0.733**	0.837	0.882	0.859	0.873	0.875	0.874	0.979	0.919	0.948
N+V	0.98	**0.869**	**0.921**	0.953	0.892	0.922	0.951	0.915	0.933	0.978	0.92	0.948
N+S+V	0.982	**0.878**	**0.927**	0.969	0.909	0.938	0.955	0.92	0.937	0.981	0.919	0.949
T+N+S+V	**0.977**	**0.895**	**0.934**	**0.968**	**0.912**	**0.939**	**0.957**	**0.921**	**0.938**	**0.979**	**0.924**	**0.951**

4. Conclusion

In this paper, we extract several informative and effect feature subsets (i.e., Node feature subset, Topology feature subset, Social feature subset and Voting feature subset) from microblog network, then solve the link prediction problem using supervised learning methods (i.e., Naïve Bayes, SVM, Logistic Regressions, Random Forest). Results show that the topology feature subset is superior to all single similarity indexes. And the social feature subset and voting feature subset we proposed could largely enhance the results of topology and node feature subsets. Finally, the combination of these feature subsets reaches the best and most steady results in all the four classifiers.

Acknowledgment

This article is supported by National Key Technology Research and Development Program of China (2012BAH39B02) and National Natural Science Foundation of Beijing. No: 4122076. Thanks for providing valuable comments and advices from the reviewers.

References

[1] Gettor, L., & Diehl, C. P. (2005). Link mining: A survey. *SIGKDD Explorations Newsletter, 7(2).*

[2] Latha, R. H., & Kumari, K. S. (2012, August). Survey on link prediction in Facebook and Twitter.

International Journal of Engineering Research and Application(IJERA), 2(5), pp. 1631-1637.

[3] Hasan, M. A., Chaoji, V., Salem, S., & Zaki, M. (2006). Link prediction using supervised learning. *Proceedings of SDM 06 Workshop on Link Analysis, Counterterrorism and Security.*

[4] Fire, M., Tenenboim, L., Lesser, O., Puzis, R., Rokach, L., & Elovici, Y. (2011). Link prediction in social networks using computationally efficient topological features. *Proceeding of 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)* (pp. 73-80). Boston, MA: IEEE.

[5] Valverde-Rebaza, J., & Lopes, A. A. (2013, Dec.). Exploiting behaviors of communities of twitter users for link prediction. *Social Network Analysis and Mining*, *3(4)*, 1063-1074.

[6] Liben-Nowell, D., & Kleinberg, J. (2007). The link prediction problem for social networks. *Proceedings of the 20th International Conference on Information and Knowledge Management* (pp. 556-559).

[7] Wang, D., Pedreschi, D., Song, C., Giannotti, F., & Barabasi, A. (2011). Human mobility, social ties, and link prediction. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1100–1108). New York.

[8] Jia, J., Wang, Y., Li, J., Feng, K., Cheng, X., & Li, J. (2013). Structural-interaction link prediction in Microblogs. *Proceedings of the 22th International Conference on World Wide Web companion* (pp. 93-194).

[9] Narasimhan, J., & Holder, L. (2014). Feature engineering for supervised link prediction on dynamic social networks. *Proceedings of the 10th International Conference on Data Mining*.

[10] Kim, J., Choy, M., Kim, D., & Kang, U. (2014). Link prediction based on generalized cluster information. *Proceedings of the 23rd international conference on World Wide Web Companion, WWW Companion '14* (pp. 317-318).

[11] Tan, F., Xia, Y., & Zhu, B. (2014, Oct.). Link prediction in complex networks: a mutual information perspective. *PLOS ONE, 9(9).*

[12] Pan, Y., Li, D., Liu, J., & Liang, J. (2010, July). Detecting community structure in complex networks via node similarity. *Physica A: Statistical Mechanics and its Applications , 389(14),* 2849–2857.

[13] Leicht, E. A., Holme, P., & Newman, M. E. J. (2006, February). Vertex similarity in networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)* , pp. 1-10.

[14] Lü, L., Jin, C., & Zhou, T. (2009, Oct). Similarity index based on local paths for link prediction of complex networks. *Physical Review E*.

[15] Yin, D., Hong, L., & Davison, B. D. (2011). Structural link analysis and prediction in microblogs. *Proceedings of the 20th ACM Conference on Information and Knowledge Management* (pp. 1163-1168).

[16] Huang, Z., Li, X., & Chen, H. (2005). Link prediction approach to collaborative filtering. *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries.* New York: ACM.

[17] Soundarajan, S., & Hopcroft, J. E. (2012). Use of supervised learning to predict directionality of links in a network. *Proceeding of 8th International Conference, ADMA 2012* (pp. 395-406).

[18] Rodrigues, S. H., & Prudencio, R. B. C. (2011). Supervised Learning for Link Prediction in Weighted Networks. *Proceedings of The 2011 International Joint Conference on Neural Network* (pp. 2281-2288). San Jose, CA.

[19] Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and Its Applications, 390(6)*, 1150-1170.

[20] Adamic, L. A., & Adar, E. (2003, July). Friends and neighbors on the web. *Social Networks, 25(3),* 211–230.

[21] Lu, Z., Savas, B., Tang, W., & Dhillon, I. (2010). Supervised link prediction using multiple sources. *Proceedings of 2010 IEEE 10th International Conference on Data Mining (ICDM)*, (pp. 923–928).

[22] Bilgic, M., Namata, G. M., & Getoor, L. (2007). Combining collective classification and link prediction.

Proceedings of 7th IEEE International Conference on Data Mining Workshops (pp. 381-386). IEEE.

[23] Zhou, M. L., Xu, Y., & Zhao, X. D. (2012). Study of feature extract on microblog user occupation classification. *Information Science and Engineering*, 20-23.

[24] The University of Waikato. (n.d.). *Weka 3: Data Mining Software in Java*. From http://www.cs.waikato.ac.nz/~ml/weka/

[25] Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32.

Han Siyao is a graduate student studying computer science at Beijing Language and Culture University, China. Her research interests include data mining, machine learning, social network and text classification. Siyao received her bachelor's degree in computer science from Beijing Language and Culture University.

Xu Yan is an associate professor in the College of Information Science, Beijing Language and Culture University and visiting professor in Institute of Computing Technology, Chinese Academy of Sciences. She received the M.S. degree in computer science in 1996 and the Ph.D. degree in computer science in 2004 from Beijing University of Aeronautics & Astronautics. Her research interests include data mining, text classification and information retrieval.