THE STUDY OF OPEN SOURCE SOFTWARE COLLABORATIVE USER MODEL BASED ON SOCIAL NETWORK AND TAG SIMILARITY

Xiang Chen
School of Management and Economics
Beijing Institute of Technology
5South Zhongguancun Street, Haidian District, Beijing 100081, China
chenxiang@bit.edu.cn

Yao-hui Pan
School of Management and Economics
Beijing Institute of Technology
5South Zhongguancun Street, Haidian District, Beijing 100081, China
panyaohui2018@google.com

ABSTRACT

Open source software (OSS) has become a mainstream in software development, utilizing a global information infrastructure. OSS is a complicated social process to understand. OSS is a multi-faceted phenomenon including code, a licensing structure, a community, development best practices, a method of diffusion. However, the current OSS collaborative researches place too much emphasis on collaborative behaviors, but ignore the study on collaborative process. By using the social network theory to abstract collaborative network topology, this paper proposes a method for constructing social network model, which considers both the contact relationship and level of collaboration between collaborators. Based on the definition of three types of contact behavior, this paper presents an approach to measuring the contact relationship intensity. Based on introducing and improving TF-IDF (term frequency–inverse document frequency), this paper presents the methods for calculating tag weights and work similarity between collaborators. Finally, by evaluating the model using data from the OSS website www.Codeplex .com, we verify that our model outperforms conventional models in both describing and forecasting collaborative behavior.

Keywords: Social network; Open source software; Collaborative user model; Tag similarity

1. Introduction

With the increasing development of E-commerce and the Internet, data exchange and business collaboration among enterprises in open network environment become more and more frequent, which promotes the significant development of collaborative virtual community. Nowadays, a wide range of commercial collaborations, especially high-technology development collaborations such as OSS, are performed via virtual community due to its easy access to information and resources. However, participants in OSS community might have never cooperated with each other. Here comes the problem when participants are unfamiliar with each other. Thus, how to perform collaboration based on virtual community has become a hot issue, and people are eager to study its theory. First-class international journals, such as Nature, Science, and etc., have published more and more related research achievements. [Ohtsuki et al. 2006; Borgatt et al. 2009].

Compared with traditional collaboration, collaboration in OSS lacks rigorous organizational forms. Implementation of collaboration in OOS is always propelled by the social relationship between collaborators[Yang et al. 2013]. In recent years, some scholars in this field have introduced Social Network Analysis (SNA) as an alternative methodology to analyze and model network collaboration. Matthew [2010] analyzed some projects from the OSS website SourceForge.net and testified that previously existing developer-developer ties significantly impact on the past and future projects. Giunchiglia et al. [2012] made an endeavor to conglomerate the socio-psycho-technical aspect of so-called social networks which could be more realistic, logically inferable and convincible towards people to claim its analogousness with real society. Van Der Aalst et al. [2005] mined the interaction mode and organization structure between collaborators and evaluated an individual's role in business process by combining workflow management and SNA. Many scholars proposed approaches to build social network based on relationships between collaborators and testified that such approaches are effective in analyzing individual behaviors

and work roles [France 1999; Priscilla 2007; Raja & Tretter 2011]. Bryant & Colledge [2002] analyzed trust relationships in electronic commerce and pointed out that trust relationships can promote collaboration and co-operation. Von Krogh et al. [2012] proposed "motivation-practice" framework and derived six concrete propositions and suggested a new research agenda on motivation in OSS collaboration. Fu and Peng [2007] analyzed knowledge sharing relationship among Agents and studied the cooperation willingness of sharing knowledge among Agents. To sum up, a large body of collaboration studies focus on studying the developers' participation motivations, but one important motivation is a collaborator's desire to gain good community reputation which is largely based on collaborative process. Recent researches also reveal that a positive collaboration decision does not depend on his evaluatee's level of experience, but rather based on his past reputation and their shared affiliations such as mutual acquaintances etc [Hua et al. 2012].

Aiming at problems mentioned above, numerous models based on social network are built to extract behaviors in virtual community and to perform information recommendations. The mainstream of these models is Collaborative Filtering (CF) recommendation algorithm based on social network. Its theoretical basis is user model which is typically constructed based on original user-item rating matrix. Recently, the constructing of rating matrix emphasizes more on users' social network. By adding members considering social network to the nearest neighbors and assigning bigger weight values to 'overlapping' members, Liu et al. [2010] applied social network to CF recommendation and testified that the proposed model outperforms a traditional CF. Meo et al. [2011] proposed an approach to constructing user profile which considers not only explicit relationships among users but also implicit ones underlying users' shared interests and behaviors, thus making the model more flexible and reliable. Chandrashekhar & Bhasker [2011] proposed a neighborhood-based, memory-based CF approach using complex relationships between users and entropy, and testified that the proposed approach outperforms traditional CF approaches in dealing with the ratings sparsity. However, most of previous user models are built based on interactions or relationships among users, without revealing collaborations in social networks.

Collaborative virtual community is an open social network. Tags, as a way of grouping content by category to make them easy to view by topic, are quite efficient in organizing a site and helping users find content they are interested in [Ricci et al. 2011]. Hung et al. [2008] calculated tag similarity based on coexisting information of user-tag matrix. Similarity between a user and an item was obtained by adding the biggest similarity between their tags. This model has been proved an excellent performance on social media recommendation. However, studies on tags are most limited to tag ratings, neglecting users' steady interests underlying tag ratings and unable to extract tags representing users' interests from multiple projects in collaborative environment.

Based on the researches above, this paper proposes a social network modeling approach in OSS collaboration, which comprehensively considers both contact relationship between collaborators and collaboration relationship generated during their work. By defining contact relationship intensity in collaborative work, this model presents a method of measuring contact relationship between collaborators. By defining work similarity between collaborators in virtual network, it proposes an approach to measuring level of collaboration. On the basis of the work above, we develop a collaborative user model for social network combining the contact relationship intensity and work similarity. We hope that it can deepen the understanding of collaboration in virtual network and provide a reference for OSS practice.

2. The Definition of The Collaborative User Model

A good user model is essential to perform recommendation efficiently. Hence, a collaborative user model is built considering both contact relationship and collaborative relationship. The collaborative user model in OSS community is mainly used to describe the relationship intensity between team members. Considering contact relationship intensity and level of collaboration, we define the OSS collaborative user model as follows.

Definition 1 Collaborative user model in OSS community can be defined as

$$VC = (C, DR, DS, DA), \text{ where}$$

(1) $C = \{c_1, c_2, ..., c_n\}$ denotes the set of collaborators;

(2) DR means the contact relationship intensity between collaborators, depending on frequency of contact, type of contact and intimacy of contact between collaborators.; DS means the work similarity between collaborators, reflecting the similarity between users' interests; DA means the total relationship intensity between collaborators.

Given a certain value set of relationship intensity, we can draft a network relationship graph and further study the collaboration relationship in virtual network based on social network theory. To transform the collaborative user model to social network graphs, we use nodes in social network graphs to denote the collaborators. Then we use edges connecting nodes to denote social relationship between collaborators. And we use the weights on these edges to denote the relationship intensity between collaborators. The aim of this model is to discover relationship between

collaborators, so directions in the network graph are not taken into consideration in this paper. While keeping all other properties constant, we remove directions of a network graph and get an undirected weighted graph. In the graph, edges between nodes denote social relationship between collaborators and weights on the edges are denoted by DA.

In order to make a further explanation, here we identify different kinds of relationship intensity. Relationship intensity is a "combination of the amount time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie" [Granoveter 1973]. Recent researches also reveal that a developer's positive evaluation decision is based on evaluatee's past reputation, their shared affiliations such as mutual acquaintances and their homophily in location, nationality etc [Hua et al. 2012].There are various ways of measuring relationship intensity. Many researchers obtain relationship intensity through a questionnaire, which always leads to a subjective result. For OSS community, we can extract interactions between users from massive open information and derive relationship intensity by analyzing these interactions.

Traditional methods for measuring relationship intensity based on interactions between collaborators only consider contact relationship between collaborators. But for network collaboration, interactions are not just routine contacts. Collaborative behaviors occurring at work also play an important role in the formation of social relationship between collaborators. Namely, users build their social relationship with others based on their routine contacts and collaborations at work. Accordingly, relationship intensity consists of contact relationship intensity and collaboration relationship intensity. Common used functions for the aggregation of a set of significance values include Max function (returning the maximum value), Min function (returning the minimum value) and Ave function (returning the average value). Here, we adopt Max function because of the mutual reinforcement between contact and collaboration relationship intensity. Thus, relationship intensity between collaborators is defined as follows.

Definition 2 Relationship intensity between i and j is defined as

$$DA_{ij} = Max - R(DR_{ij}, DS_{ij}), \text{ where}$$

$DA_{i,j}$ means the relationship intensity between i and j; $DR_{i,j}$ means the normalized contact relationship intensity; $DS_{i,j}$ means the normalized work similarity between i and j. $Max - R$ means fuzzy max operator.

3. Measurement of Contact Relationship Intensity

Intensity of contact between collaborators in real world will influence their collaborations in work to a great extent. Likewise, intensity of contact between collaborators in virtual environment will influence their trust for each other and affect their collaboration further. A common method of measuring contact relationship intensity is extracting collaborative behaviors between collaborators from contact information. Massa et al. [2006] propose a Trust-Aware solution. Assuming trust is propagated to the maximum propagation distance d, a user at distance q from source user will have a predicted trust value of $(d-q+1)/d$, where trust value is the weight on each edge. Zhang [2007] proposes a semantic-based reasoning, reputation-based, content-based and context-based combined trust model and a heuristic information-based targeted discovery approach. However, both of these two methods suffer some weaknesses. The Trust-Aware approach cannot identify different contact patterns, and the combined trust model is too complicated to understand and realize. In this part, considering different contact patterns, we propose a relatively simple approach to calculating contact relationship intensity between collaborators.

3.1. Definition of Information Issue Sequence (IIS) and Behavioral Rules

Social relationship between collaborators in virtual network is related to their attention degree. At the same time, collaboration as a behavior still needs a further analysis of its interactions. Considering the information exchanging frequency and time, this paper presents a definition of IIS. Based on the definition, it defines three behavioral rules for information exchanging between collaborators: reply rule, quote rule and co-occurring rule.

Definition 3 IIS in network is defined as

$$IIS = (C, I, f), \text{ where}$$

(1) $C = \{c_1, c_2, ..., c_n\}$ is the set of collaborators;

(2) $I = \{(i_1, t_1), (i_2, t_2), ..., (i_m, t_m)\}$ is the information sequence; t_m is the issue time of information i_m;

(3) f is information issue behavior; $f: c_k \to I$ means that collaborator c_k issues information sequence; $f: c_n \xrightarrow{t_m} i_m$ means that collaborator c_n issues information i_m at time t_m.

Rules 1 Reply Behavioral Rule: for an IIS, there is a reply behavior between c_k and c_l, if

(1) $\exists c_k, c_l \in C, \exists (i_q, t_q) \in I, c_k \to I, c_l \xrightarrow{t_q} i_q$;

(2) $\exists c_k, c_l \in C, \exists (i_p, t_p), (i_q, t_q) \in I, c_k \xrightarrow{t_p} i_p, c_l \xrightarrow{t_q} i_q, t_p < t_q$.

In a reply behavior, a replier has direct communication with the issuer instead of a strong response to the issued information. Thus, contact relationship intensity between them is relatively weak.

Rule 2 Quote Behavioral Rules: for an IIS, there is a quote behavior between c_k and c_l, if $\exists c_k, c_l \in C, \exists (i_p, t_p), (i_q, t_q) \in I, c_k \xrightarrow{t_p} i_p$, $c_l \xrightarrow{t_q} i_q$, where $t_p < t_q$ and $i_p \infty i_q$ (i_p is similar to i_q).

In a quote behavior, a replier acknowledges the information issued by the issuer, which conveys more information about the interaction and represents stronger contact relationship intensity than a reply behavior does.

Rule 3 Co-occurring Behavior Rule: for an IIS, there is a co-occurring behavior between c_k and c_l, if $\exists (i_p, t_p), (i_q, t_q), (i_r, t_r), (i_s, t_s) \in I$, $\exists c_k, c_l \in C$, $c_k \xrightarrow{t_p} i_p$, $c_l \xrightarrow{t_q} i_q$, $c_k \xrightarrow{t_r} i_r$, $c_k \xrightarrow{t_s} i_s$, where $t_p < t_q < t_r$ or $t_q < t_r < t_s$.

In a co-occurring behavior, users show interest in the same information within a short period of time. They have direct interaction with each other through reading and replying to the information issued by each other, which represents relatively strong contact relationship intensity.

3.2. Measurement of Contact Relationship Intensity

Many researches revealed that relationship intensity between collaborators in network is mainly determined by the following factors: frequency of contact, types of the tie, intimacy of the tie, similarity of interest [Joinson 2001; Haythornthwaite 2002]. Combining the characteristics of information interactions on the Internet and the definitions of IIS and behavioral rules, we assess contact relationship intensity by using a combination of frequency of contact, type of contact and intimacy of contact.

Frequency of contact is derived from the count of interactions between collaborators per unit of time. To avoid the impact of difference of activity level between users, we apply the incidence rate of interactions instead of count of interactions to measure contact relationship intensity between collaborators. For $c_k, c_l \in C$:

$$N_{ij,t} = T_{ij,t} / (T_{i,t} + T_{j,t}), \text{ where}$$

$T_{i,t}$ denotes the total count of interactions between c_i and others at time t; $T_{j,t}$ denotes the total count of interactions between c_j and others at time t; $T_{ij,t}$ denotes the count of interactions between c_i and c_j.

Types of contact include reply behavior, quote behavior or co-occurring behavior occurring in an IIS. For an IIS, we denote the set of interactions by $V = \{v_r, v_i, v_c\}$, where v_r, v_i, v_c means reply behavior, quote behavior and co-occurring behavior respectively.

Intimacy of contact involves in the time factor of occurrence of interactions. There is a strong tie between users who interact with each other in a short period of time. Otherwise, the tie is weak. In the measurement of relationship, we denote the time factor in interactions by time coefficient λ_t. To ensure that λ_t descends with the increasing of time intervals, a descending function to measure λ_t was given [Gu & Zhang 2010]. Gao et al. [2009] analyzed the time intervals between posts in online communities and verified that the time interval between posts conforms to exponential distribution. To test the distribution of time interval between collaborators' interactions in virtual community, we applied 5375 time intervals captured from www.Codeplex .com to picture the time-interval P-P graph (Figure 1), which shows that the time interval (/h) has an asymptotic exponential distribution despite the existence of inaccurate deviation. To further verify its accuracy, we test its goodness of fit in the whole dataset using K-V test. The R-square of K-V test is 0.973, which approximately approaches 1 and demonstrates that exponential distribution can predict distribution of time interval well with a high goodness of fit. Therefore, here we assume that time intervals conform to exponential distribution. Then we perform a Kolmogorov-Smirnov test and the Z-value is 6.378 with a most extreme difference of 0.087 (Table 1), which doesn't reach the 0.05 significant level. However, we deem that time intervals between interactions have an exponential distribution given the following facts: (1) Z-value varies from the critical value slightly. (2) The dataset is obtained with crawler. It makes a closest-time inference for partial missing value and loses some reply data when processing time data.

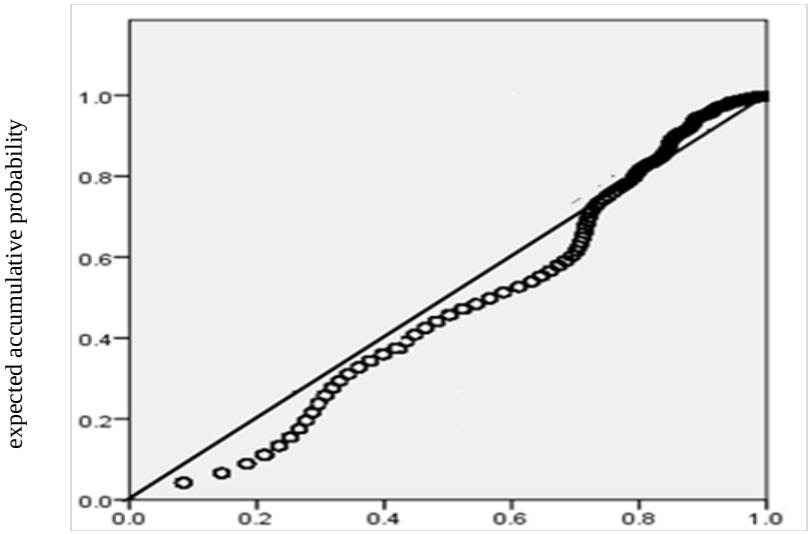

Figure 1: Time-interval P-P graph

Table 1: Kolmogorov-Smirnov test for time intervals between posts

Statistical Variables		f
N		5375
Exponential Parameters[a,b]	Mean	39.7148
Most Extreme Differences	Absolute	.087
	Positive	. 087
	Negative	-.054
Kolmogorov-Smirnov Z		6.378
Asymp . Sig. (2-tailed)		.000

Combining the assumption described above and descending function proposed by Gu et al., this paper defines a descending function as

$$\lambda_t = a^{(T-t)/\theta}, \text{ where}$$

$a(0 < a < 1)$ denotes time sensitivity factor and θ denotes the expectation of time intervals.

Based on the above analysis, we define a measurement for contact relationship intensity as follows.

Definition 5 Contact relationship intensity per unit of time between collaborators is determined by frequency of contact, type of contact and intimacy of contact. Let $C = \{c_1, c_2, ..., c_n\}$, for user c_i, c_j, contact relationship intensity is defined as

$$DR_{ij} = \sum_{t=0}^{T} (\sum_{v \in V} \omega_v * T_{v,ij,t} / (T_{v,i,t} + T_{v,j,t}) \times \lambda_t), \text{ where}$$

(1) ω_v denotes normalized weight of influence of reply behavior on contact relationship.

(2) $T_{v,i,t}$ is the total count of interactions of type v that user c_i has with other users at unit time t; $T_{v,j,t}$ is the total count of interactions of type v that user c_j has with other users at unit time t; $T_{v,ij,t}$ is the count of interactions of type v between c_i and c_j at unit time t.

(3) λ_t denotes the time descending function.

4. Measurement of Similarity between Collaborators

Contact relationship intensity describes intensity of contact between collaborators. But collaboration behaviors

in virtual environment are more arbitrary. Collaborators in OOS collaborations may participate in a project only out of interests. Therefore, finding content collaborators are interested in is beneficial to mining OOS collaborative relationship. Tags are main tools describing collaborators' interests and specialties. Compared with tags in social network websites, tags in collaborative websites are usually administrated by website administrators and users can only employ tags defined in the website, which have less arbitrariness and stronger pertinence. In a collaborative website, projects are characterized by different tags. Because collaborators have distinct degrees of attention on different projects, we can compare collaborators by comparing tags of their projects. Besides, there exists an affiliating relationship between collaborators and projects, projects and their tags. We can derive similarity between collaborators from such affiliating relationship. Here we define tag set for each collaborator firstly.

Definition 6 For collaborator j, tag set is defined as

$$Tc_j = \bigcup_{w_i \in W_j} Tw_i \text{, where}$$

$C = \{c_1, c_2, ..., c_n\}$ denotes the set of collaborators; $W = \{w_1, w_2, ..., w_m\}$ denotes the set of projects; $Tw_i = \{t_1, t_2, ..., t_k\}$ denotes the tag set of project i; w_i denotes the set of projects that colaborator j has participated in.

Based on this definition, we calculate weight for each tag in Tc_j. Because tag is a type of text-information, we apply term frequency-inverse document frequency (TF-IDF) to calculate its weight. TF-IDF is a weighted algorithm to evaluate how important a word is to a document in a collection or corpus, which is commonly used as a weighting factor in information retrieval and text mining. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus [Aizawa 2003].

Term frequency (TF) is a measure of how often a term is found in a document, which increases proportionally to the number of times a word appears in a document. In collaborative website, we can build a document based on the tag set of a collaborator. Considering the most common structure for calculating TF and characteristics of tags, we define the calculating formula of TF as

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}},$$

where $n_{i,j}$ denotes the total number of times tag t_j appears in the project set of collaborator i; $\sum_k n_{k,j}$ denotes the total number of times tag t_j appearing in all projects.

Inverse document frequency (IDF) is a measure of whether the term is common or rare across all documents, which increases disproportionally to the number of times a word appears in the corpus. It reflects that less common words have a better ability to distinguish different types of documents. Combining the most common structure for calculating IDF and characteristics of tags, we define the formula for calculating IDF as

$$idf_j = \log \frac{|D|}{|\{k : t_j \in d_k\}|},$$

where $|D|$ denotes the total count of projects; $|\{k : t_j \in d_k\}|$ denotes the count of projects including tag t_j ($n_{i,j} \neq 0$). Because tags are set following the existence of projects in a collaborative website, it avoids encountering the situation that dividend is zero. When a tag is not attached to any project, we adjust divisor to $1 + |\{k : t_j \in d_k\}|$.

Then, TF-IDF is calculated as

$$tfidf_{i,j} = tf_{i,j} \times idf_j$$

A high weight in TF-IDF is reached by a high term frequency (in the given project set of collaborator i) and a low document frequency of the tag in the whole collection of projects. Hence the weights are consistent with the practical assumption.

Based on the formula of TF-IDF, we define tag vector for each collaborator.

Definition 7 For collaborator j and $t_l \in Tc_j$, tag vector is defined as $Uc_j = \{(t_1, r_1), (t_2, r_2)....., (t_l, r_l))\}$, where r_l is calculated by applying the formula of $tfidf_{i,j}$.

According to the definition of social tagging, when describing a collaborator's interest, we usually describe the content of projects using tag vector. The vector dimension denotes the total count of tags in all projects. Each value in the vector is the respective tag's weight in the collaborator's tag set. Then we can calculate normalized TF-IDF vector which represents the collaborator's tag set with the tag vector. Similarity between two tag libraries hence is denoted by the reciprocal of distance between two TF-IDF vectors, where the distance is obtained using cosine similarity. Thus, the calculation of similarity between tag sets is simplified as the calculation of similarity between TF-IDF vectors of tag sets. Based on the above work, we can define similarity between collaborators as follows.

Definition 8 For collaborator i and j, the similarity between them is defined as

$$DS_{ij} = \frac{\sum_{k=1}^{n} r_k^i r_k^j}{\sqrt{[\sum_{k=1}^{n} (r_k^i)^2][\sum_{k=1}^{n} (r_k^j)^2]}}),$$

where Uc_{i} and Uc_j denote tag vectors of collaborator i and j respectively; n is the total count of tags in Uc_{i} and Uc_j; r_k^i denotes the weight of the k th tag in tag set of collaborator i.

5. Experiments and Results

To testify the efficiency of the proposed model, this paper chooses classic indicates and construct representative indicates as baseline. Based on actual data, this paper makes comparisons between our proposed model and other models and evaluates the efficiency of our proposed model. The process is as follows.

5.1. Collection of actual data

Using Nutch crawling code, this paper captures webpage data from the OSS website *www.Codeplex .com* and obtains experimental data based on collaboration information by formatting webpage data with programming. Then we test the model using the experimental data mentioned above. Codeplex is Microsoft's open source project hosting website with a total number of 19449 projects and a rigorous project framework. Here, we applied open source projects with over six collaborators and with Alpha published to test the model. These projects involve 3815 members and 19707 posts in thread discussing.

To examine the model's accuracy, we divide the whole dataset into two pieces by time point, where data before Jan 1. 2010 is used as the training set and data from Jan 1, 2010 to June 1, 2011 is used as the test set.

5.2. Indicates of model evaluation

To evaluate the model's accuracy in predicating collaborative characteristics, we propose the following indicates to measure the model's effectiveness.

Definition 9 Predicted relationship intensity D_m is defined as the average of relationship intensity between users who have interactions with each other. The formula for calculating D_m is given by:

$$D_m = (\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij} * D_{ij}) / (\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}),$$

where C_{ij} is the count of interactions occurring between c_i and c_j in test set; D_{ij} is the relationship intensity between c_i and c_j, which is calculated by using DR, DS, DA model with training set. Predicted relationship intensity reflects the effectiveness of the relationship intensity model in measuring level of interaction between users. A bigger value indicates a better performance.

Definition 10 Precise of the model P_m is defined as the average of the model's precise in detecting collaborative teams. The precise function is given by:

$$P_m = (\sum_{i=1}^{n} C_i / N_i) / n,$$

where N_i denotes the total count of interactions occurring between c_i and others in test set; C_i denotes the count of interactions occurring between c_i and the others in his team detected by the model. The precise reflects the accuracy of the model in detecting collaborative relationship, which indicates a better performance with a bigger value.

Besides, this paper focuses on recommending accurate collaborators. Statistical accuracy and decision support accuracy are two main approaches to evaluating performance of recommendation. And decision support accuracy is usually measured by Recall and Precision.

Precision is the value of hits in recommended collaborators list compared with test set divided by the number of recommended collaborators.

$$Precision = Hits/TopN ,$$

where *Hits* denotes accurate collaborators in recommended collaborators list (excluding previous collaborators of projects); *TopN* denotes total number of collaborators in recommended collaborators list.

Recall is the value of hits in recommended collaborators list compared with test set divided by the number of collaborators who participate in projects in test set.

$$Recall = Hits/N ,$$

where *Hits* denotes accurate collaborators in recommended collaborators list (excluding previous collaborators of projects); N denotes total number of collaborators in real projects.

Precision and Recall are two contradictory terms. Recall increases with the increasing of $TopN$, while Precision decreases with the increasing of $TopN$. Considering both Recall and Precision play important roles in evaluating the efficiency of recommender system, $F1$ is constructed to find the optimal balance point of Recall and Precision.

$$F1 = 2 * Precision * \frac{Recall}{Precision + Recall}$$

This paper applies Recall, Precision and $F1$ to evaluate efficiency of recommender system. In every round of cross-validation, recommender system predicts and ranks possible collaborators participating in the required projects. Then we calculate Recall, Precision and $F1$ by comparing recommended collaborators list obtained using the proposed model with the collaborators list in test set.

The efficiency of recommender system should be evaluated from the whole dataset. Hence, we calculate the statistics by adding values of all the required projects instead a single project when counting *Hits*, *TopN* and N. *Hits* denotes total number of collaborators recommended successfully in all the required projects. *TopN* denotes total number of collaborators recommended in all the required projects. N denotes total number of collaborators participating in all the required projects actually.

5.3. Selecting of compared models and experimental results

The experiments are conducted by programming and comparing the running results. All the experiment programs are developed with the same java development platform based on the same data set.

With regard to the excellent performance of Trust-aware model proposed by Massa et al. in predicting trust relationship and describing social network characteristics in OOS websites, we adopt the model to evaluate collaborative characteristics. When evaluating collaborative characteristics, we initially construct three user models namely DA, DS and DR models, and then draw social network graph based on these models. Finally, we apply Girvan-Newman algorithm to detect collaborative teams for the obtained social network graph.

Based on the above definition, we calculate the predicted relationship intensity and precise. The results are shown in Table 2.

Table 2: Experimental Results of Different Models

	N of isolated nodes	*N of teams*	*Dm*	*Pm*
DA	103	173	0.0332	81.48%
DS	189	181	0.0301	68.19%
DR	385	189	0.0299	52.06%
Trust-1	601	282	0.0218	40.11%
Trust-2	543	245	0.0231	44.15%
Trust-3	486	231	0.0242	46.30%

As shown in Table 2, the DR model outperforms Trust-Aware approach in situation where both the predicted relationship intensity and observed precise are sufficiently larger than that of Trust-Aware approach. But compared with increasing in D_m value, the DR model contributes little to improve its precise, which is mainly because some participants have never issued posts in the projects. Thus, both the Trust-Aware model and DR model do not provide a good solution to isolated node problem. For this problem, the DS model provides a comparatively good solution, which employs similarity between tag sets that contains all tags in collaborators' projects to describe similarity between collaborators. This model enables us to detect work characteristics of collaborators well even without posts and hence has a great improvement in its precise P_m. Compared with improvement in value, the DR model has no significant improvement in D_m, which is mainly because work characteristics cannot represent intimacy or

intention to cooperation between collaborators comprehensively. Compared with the DR and DS models, DA model demonstrates a better performance in both relationship intensity and observed precise, which provides a better solution to predict users' future interactions and to describe work characteristics and competence.

To sum up, the DR model provides a better solution to predict collaborative behavior and intention, while the DS model performs well in predicting work characteristics and collaborative competence. The results suggest that the DA model, a combination of the DA and DS model, provides a better solution to detect project teams and predict future interactions between collaborators in social network accurately, which better reflects social relationship and collaborative relationship in real world.

To testify the efficiency of the proposed model in recommending collaborators, we compare DS model with recommendation model proposed by Hung et al. (TBUPR). User-tag matrix corresponds to user-interest matrix in TBUPR model when TBUPR model is applied to recommend collaborators in virtual collaborative community. The results of comparison are listed in Figure 2.

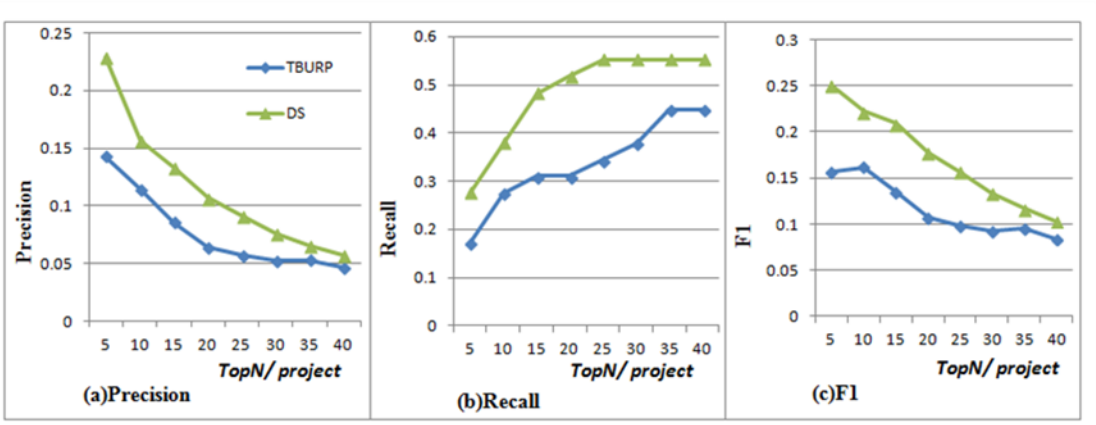

Figure 2: Compared results of DS model with TBUPR model

As shown in Figure 2, DS model outperforms TBUPR model in recommending accuracy. Recall suggests that hits of collaborators list of DS model are higher than that of TBUPR. Besides, DS model offers an efficient solution to matrix sparseness by measuring similarity between collaborators and matching degree between collaborators and projects, thus resulting in a better recommending performance. In sum, based on collaborators' tags, our proposed model can express users' preferences and work capabilities well and improve the efficiency of recommending by decreasing the influence of matrix sparseness on recommendation with introducing tag similarity.

6. Conclusions and Future Research

In this paper we propose a social collaborative user model based on social network. The model detects contact relationship and collaborative relationship between collaborators from their contact and work information. It also constructs a social network model to reflect social relationship between collaborators. And we conducted an empirical validation on Codeplex dataset to evaluate the feasibility and effectiveness of the proposed model. The experimental results show that our proposed model is effective in improving the measuring accuracy of relationship between collaborators, detecting a collaborator's role in a work team and detecting work teams in social network. Thus it provides a reference and foundation for the future studies on Internet-based collaboration problem.

However, our study also has some limitations. Though defining rules for three types of behavior in this paper, we did not research further on their relationship and evolvement. Thus, future study should focus on this field combining more collaboration data from the Internet.

Acknowledgment

The research has been supported by the Natural Science Foundation of China (No. 71102111).

REFERENCES

Aizawa, A., "An Information-Theoretic Perspective of TF-IDF Measures," *Information Processing & Management*,Vol.39, No.1:45–65, 2003.

Borgatt, S.P., A. Mehra, D.J. Brass and G. Labianca, "Network Analysis in the Social Sciences," *Science*, Vol.323:892 – 895, 2009.

Bryant, A. and B. Colledge, "Trust in Electronic Commerce Business Relationships," *Journal of Electronic Commerce Research*, Vol. 3, No. 2:32-39, 2002.

Chandrashekhar, H. and B. Bhasker, "Personalized Recommender System Using Entropy Based Collaborative Filtering Technique," *Journal of Electronic Commerce Research*, Vol. 12, No. 3: 214-237, 2011.

France, B., "Communication Patterns in Distributed Work Groups: A Network Analysis," *IEEE Transactions on Professional Communication*, Vol.42, No.4: 261-275, 1999.

Fu, X.J. and Y.H. Peng, "Social Network Analysis Based Optimization for Knowledge Flows Planning," *Computer Integrated Manufacturing Systems*, Vol.13, No.11: 2168-2177, 2007. (in Chinese)

Gao W, Q.H. Li, B. Zhao and GH. Cao, "Multicasting in Delay Tolerant Networks: A Social Network Perspective," *Proceedings of the 10th International, Symposium on Mobile and Hoc Network & Computin*, New York, USA, 299-308, 2009.

Giunchiglia, F., S.H. Mukta, M.T. Nayeem and K.T Hasan, "Semantic Enabled Role Based Social Network," *International Journal of Intelligent Systems and Applications*, Vol.4, No.12: 1-11, 2012.

Granoveter, M.S., "The Strength of Weak Ties," *American Journal of Sociology*, Vol.78, No.6: 1360-1380, 1973.

Gu, C.J. and S.Y. Zhang, "A Novel Trust Management Model for P2P Network with Reputation and Risk Evaluation," *2010 International Conference on E-Business and E-Government*, Guangzhou, China, 3544-3547, 2010.

Haythornthwaite, C., "Strong, Weak, and Latent Ties and the Impact of the New Media," *The Information Society*, Vol.18: 385-401, 2002.

Hua, D., J.L. Zhao and J. Cheng, "Reputation Management in an Open Source Developer Social Network: An Empirical Study on Determinants of Positive Evaluations," *Decision Support Systems*, Vol.53, No.3: 526–533, 2012.

Hung, C.C., Y.C. Huang, J.Y. Hsu and D.K.C. Wu, "Tag-Based User Profiling for Social Media Recommendation," *Workshop on Intelligent Techniques for Web Personalization & Recommender Systems at AAAI2008*, Chicago, Illinois, 49-55, 2008.

Joinson, A.N., "Self Disclosure in CMC：The Role of Self Awareness and Visual Anonymity," *European Journal of Social Psychology*, Vol.31:177-192, 2001.

Liu, Z.B., W.Y. Qu, H.T. Li and C.S. Xie, "A Hybrid Collaborative Filtering Recommendation Mechanism for P2P Networks", *Future Generation Computer Systems*, No.26: 1409-1417, 2010.

Meo, P.D., A. Nocera, G. Terracina and D. Ursino, "Recommendation of Similar Users, Resources and Social Networks in a Social Internetworking Scenario," *Information Sciences*, Vol.181, No.7: 1285-1305, 2011.

Mattew, V.A., "The Importance of Social Network Structure in the Open Source Software Developer Community," *Proceedings of the 43rd Hawaii International Conference on System Sciences*, Hawaii, USA, 1-10, 2010

Massa, P. and P. Avesani, "Trust-Aware Collaborative Filtering for Recommender System," *Proceeding of Fourth International Conference on Trust Management*, Pisa, Italy, 492-508, 2006.

Ohtsuki, H., C. Hauert, E. Lieberman and M. A. Nowak, "A Simple Rule for the Evolution of Cooperation on Graphs and Social Networks," *Nature*, Vol.441: 502-505, 2006.

Priscilla, A., "Redefining and Measuring Virtual Work in Teams: An Application of Social Network Analysis," *40th Annual Hawaii International Conference on System Sciences*, Hawaii, USA, 1001 -1010, 2007.

Raja, U. and M.J. Tretter, "Predicting OSS Development Success: A Data Mining Approach," *International Journal of Information System Modeling and Design*, Vol.2, No.4: 27-48, 2011

Ricci, F., L. Rokach and B. Shapira, "Recommender Systems Handbook," *Springer:* 2011.

Van Der Aalst, W.M.P., H.A. Reijers and M. Song, "Discovering Social Networks from Event Logs," *Computer Supported Cooperative Work*, Vol.14, No.6: 549–593, 2005.

Von Krogh, G., S. Haefliger, S. Spaeth and M.W. Wallin, "Carrots and Rainbows: Motivation and Social Practice in Open Source Software Development," *MIS Quarterly*, Vol.36, No.2: 649-676, 2012.

Yang, M.H., J.C.H. Chen, C.L. Tsai and H.Y Chao, "Investigating Collaborative Commerce System from the Perspective of Collaborative Relationship," *Journal of Electronic Commerce Research*, Vol.14, No. 1:85-98, 2013.

Zhang, Y., "Research on Models and Algorithms about Trust Computing and Mining Analysis of Online Social Network," *Ph.D. Dissertation of Zhejiang University*, 2009. (in Chinese)