Followee Recommendation in Asymmetrical Location-Based Social Networks

Josh Jia-Ching Ying, Eric Hsueh-Chan Lu, and Vincent S. Tseng*
Institute of Computer Science and Information Engineering
National Cheng Kung University
No.1, University Road, Tainan City 701, Taiwan (R.O.C.)
{jashying, ericlu416 }@gmail.com, *Correspondence: tsengsm@mail.ncku.edu.tw

ABSTRACT
Researches on recommending followees in social networks have attracted a lot of attentions in recent years. Existing studies on this topic mostly treat this kind of recommendation as just a type of friend recommendation. However, apart from making friends, the reason of a user to follow someone in social networks is inherently to satisfy his/her information needs in asymmetrical manner. In this paper, we propose a novel mining-based recommendation approach named Geographic-Textual-Social Based Followee Recommendation (GTS-FR), which takes into account the user movements, online texting and social properties to discover the relationship between users' information needs and provided information for followee recommendation. The core idea of our proposal is to discover users' similarity in terms of all the three properties of information which are provided by the users in a Location-Based Social Network (LBSN). To achieve this goal, we define three kinds of features to capture the key properties of users' interestingness from their provided information. In GTS-FR approach, we propose a series of novel similarity measurements to calculate similarity of each pair of users based on various properties. Based on the similarity, we make on-line recommendation for the followee a user might be interested in following. To our best knowledge, this is the first work on followee recommendation in LBSNs by exploring the geographic, textual and social properties simultaneously. Through a comprehensive evaluation using a real LBSN dataset, we show that the proposed GTS-FR approach delivers excellent performance and outperforms existing stat-of-the-art friend recommendation methods significantly.

Author Keywords
Location-Based Social Network (LBSN), Followee Recommendation, Semantic Similarity, Data mining.

ACM Classification Keywords
H.2.8 [Database Management]: Database Applications – *Data Mining, Spatial Databases and GIS*

General Terms
Performance, Design, Experimentation.

UbiComp'12, September 5-8, 2012, Pittsburgh, USA.
ACM 978-1-4503-1224-0/12/09.

INTRODUCTION

With the rapid growth and fierce competition in the market of social networking services, many service providers have deployed various recommendation services, such as friend recommender, to promote users to understand each other in order to grow the underlying social networks. For example, several well known social networking systems, such as Facebook, Twitter, and FriendFeed, they have provided various services on friend search and recommendation. These services are very useful for users to find people who have similar interests, learn and share information/experiences with others, and make friends. Based on our observations, we could categorize these social networks into two classes:

- *Symmetrical Social Networks* (*SSNs*) that correspond to the general social relationship of users. This kind of social network is always represented as undirected graph, such as Facebook, Gowalla and Foursquare.
- *Asymmetrical Social Networks* (*ASNs*) that are likely represented as directed graph, such as Tweeter and Everytrail. In this kind of social network, users can follow other users whom they are interested in. If users follow somebody, they will receive notifications when their followees upload new trips or do something special on the social network website.

As contrasted with *SSNs*, the concept of social activity on *ASNs* is more complicated. Because people may not only want to make friend when they follow someone, they probably are more interested in the information which is provided by someone [12]. In other words, if some people have information needs, they will try to search and follow the persons who have the information. Here, we call this kind of asymmetric relationship "information-need relationship". As shown in Figure 1, user *A* and user *B* are friends each other if they have link in a symmetrical social network (see Figure 1(a)), but the reason user *A* follows user *B* may be user *A* and user *B* have the information-need relationship (i.e., user *B* provides some interesting

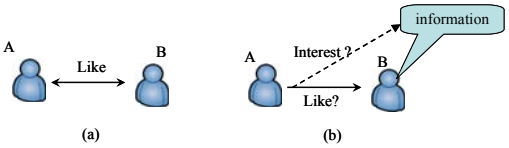


Figure 1. Two Types of Social Networks.

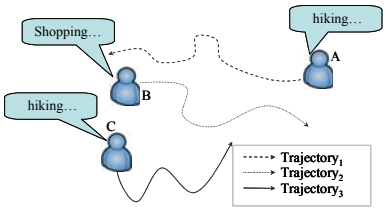


Figure 2. A scenario of Information Needs.

information for user *A*). As the result, *ASNs* always contain these two kinds of relationship, i.e., social relationship and information-need relationship. Thus we argue that these two kinds of relationship must be considered for followee recommendation.

However, most of the followee recommendation engines (called followee recommenders) just directly adopt friend recommendation techniques for recommending followees on *ASNs*. In other words, they only use the concept of social relationship to make recommendations (e.g., some systems often recommend followees' followees to their users) instead of capturing the information-need relationship. As the result, the existing works focus only on analysis of social properties, like followee of followee link, common followee, etc., to make recommendation. We argue that this recommending strategy could not work well on information-need relationship. The reason is that the social properties could not illustrate complete information-need relationship. For example, suppose that two users follow a lot of hikers, but the reasons of that the two users follow these hikers may be totally different. One of the two users may just like hiking and another may likes the pictures which are provided by these hikers. Accordingly, it is necessary to involve more information to make followee recommendation.

Although there are several previous studies [9, 10, 11]on Location-Based Social Networks (LBSNs) involve the information of user movements for potential friend recommendation, these existing techniques mostly focus on analyzing the similarity of moving sequences (i.e., geographical or semantic trajectories). Due to the experience binding of users' movements, these recommendation techniques only recommend people for the users who have geographical or semantic common movements. Take Figure 2 as an example, there are three trajectories provided by the three different users. There is no user who is similar to user *C* since there is no trajectory which is similar Trajectory_3. Thus, the traditional recommendation techniques would suffer for the problem of experience-limitation (i.e., only recommending the users who have similar movements). However, the textual information, such as travelogues and comments of trips, did not be involved in the existing work. Actually, the textual information could represent the intension of users' preference or fancy. Take Figure 2 as an example again, we can see that the user *A* and user *C* often talk about "hiking" in their provided textual information. Thus, we may recommend them to each other as their followee.

To address the above-mentioned problem, we propose a novel approach named *Geographic-Textual-Social Based Followee Recommendation (GTS-FR)* for recommending users the followees based on not only social factors but also users generated data. As shown in (1), given a set of users *U*, the problem of followee recommendation can be formulated as classifying the relation of a given ordered user pair, *u* and *v,* into the binary class, *1* and *0*. Here, class *1* means that user *u* follows user *v*, and class *0* means that user *u* does NOT follow user *v*.

$$f(u \mid v) \rightarrow \{0,1\}, \text{where } u \in U \text{ and } v \in U \quad (1)$$

Note that $f(u|v) \neq f(v|u)$ because the "follow" is asymmetric relation. Hence, followee recommendation in LBSN can be addressed as the problem of binary-class classification for each individual user (i.e., to classify all other users into "followee" class and "non- followee" class). While binary-class classification techniques have been developed for many applications, such as protein function classification [4], music categorization [6] and semantic scene classification [2], the problem has not been explored previously under the context of asymmetrical LBSN. Furthermore, the geographical and textual information changes quickly especially in LBSNs. How to extract appropriate features to support the recommendation from such heterogeneous data is also a critical and challenge issue. To support followee recommendation based on user-generated data and social properties, we address this problem by learning a SVM classifier for each individual user. To do so, a fundamental issue is to identify and extract a number of descriptive features for each user in the system. Selecting the right features is important because those features have a directed impact on the effectiveness of the prediction task. As mentioned earlier, only considering the common movements and social properties did not work well. Therefore, we explore the users' textual information and seek unique features of users captured in their own information and information need for followee classification.

By dealing with the observations prompted in the above examples, we extract features of user pair in three different but complementary aspects: 1) *Social Property* (*SP*), 2) *Geographical Property* (*GP*), and 3) *Textual Property* (*TP*). The features extracted from *Social Property*, corresponding to a given user pair, can be derived from the intersection among their followees and followers based on statistical analysis. To consider the factor of users' provided information, we extract the features from *Geographical Property* to capture the relevance between users' provided trips by a HITS-Based random walk model [3]2. To involve the factor of users' information need, we extract the features from *Textual Property* to capture the relevance between users' provided textual information and information needs by exploiting the representative keywords of their travelogues and comments of trips. To facilitate feature extraction from *Textual Property*, we propose a family of graph representations that capture the

user-keyword and location-keyword relationships from the users' textual information. We develop an algorithm to build a captures the information needs of each user by exploring above-mentioned two graphs. Accordingly, for each ordered user pair (*u, v*), we derive the probability to evaluate the closeness between the information provided by *v* and *u*'s information needs. This following probability is thus treated as a feature of *Textual Property*, along with the features derived from *Social Property* and *Geographical Property*, to feed the binary SVM in our *GTS-FR* model.

This research work has made a number of significant contributions, as summarized below:

- We propose to tackle the problem of user textual information mining in users' relations, which is a crucial prerequisite for effective followee recommendation in an asymmetrical LBSN.
- We propose *Geographic-Textual-Social Based Followee Recommendation* (*GTS-FR*), a new approach for users' similarity mining and followee recommendation on an asymmetrical LBSN. The problems and ideas in *GTS-FR* have not been explored previously in the research community.
- We formulate the problem of followee recommendation in an asymmetrical LBSN as the problem of binary class classification and propose *GTS-FR* to learn a SVM for each user to estimate possibilities of other users. In the proposed *GTS-FR*, we explore 1) *Social Property* (*SP*), 2) *Geographical Property* (*GP*), and 3) *Textual Property* (*TP*) by exploiting the LBSN data to extract descriptive features.
- We use a real dataset, which was crawled from *EveryTrail* [1], to evaluate the effectiveness of our proposed *GTS-FR* in a series of experiments. The results show *GTS-FR* delivers superior effectiveness over other recommendation strategies in terms of the popular measures precision, recall and F-measure.

The rest of this paper is organized as follows. We briefly review the related work in 2nd Section and provide our followee recommendation approach GTS-FR in 3rd Section. Finally, we present the evaluation result of our empirical performance study in 4th Section and discuss our conclusions and future work in 5th Section.

RELATED WORK

Actually, apart from friend-of-friend strategy, most existing friend recommendations on LBSN focus on dealing with users' similarity measurement for making recommendations. Many studies [5, 6, 8, 11] have proposed to discuss the problem of similarity measurement in the field of data mining. Trajectory similarity measurement [5] and user similarity measurement [6, 8, 11] are two hot topics in this problem. In [5], Lee *et al.* proposed a Partition-and-Group method to calculate the similarity between two trajectories. For all trajectories, they first find the characteristic points to form line segments and then apply three kinds of distance measures, i.e., perpendicular distance, parallel distance and angle distance, on these segments to group the trajectories. However, these distance measures are only applicable to geographic information and thus can not be used to measure user similarity based on semantic trajectories.

The main idea of trajectory-based user similarity measurement is to derive the user similarity by analyzing the movement behaviors of mobile users. In [11], Zheng *et al.* proposed a personalized friend and location recommendation system which is called *HGSM-based recommender*. To explore users' similarities, the system considers users' movement behaviors in various location granularities. Based on the definition of stay point which is the geographic region where mobile users usually stay for over a time threshold, the system discovers all of the stay points in trajectories and then employ a density-based clustering algorithm to organize these stay points as a hierarchical framework. Such cluster is named stay region (or stay location). As such, a personal hierarchical graph is formed for each user. For each level of hierarchical graph, a user's trajectory can be transformed as a sequence of stay regions. To measure the similarity of two users, some common sequences, named similar sequence, are discovered by matching their stay region sequences in each level of hierarchical graph. Then, for each stay region, the TFIDF value for a similar sequence is calculated, where TF value represents the minimum frequency of the two users accessed this stay region within the similar sequence, while the IDF value indicates the number of users who have visited this stay region. Finally, the similarity between two users is derived by the summation of the TFIDF values of all stay regions within the similar sequences. However, this approach treats every stay region in the similar sequence independently, i.e., without considering the sequential property of stay regions in the similar sequence. In [8], the *LBS-Alignment* method was proposed to calculate the similarity of two mobile users. The LBS-Alignment method calculates the similarity of two users by using the longest common sequence within their Mobile Sequential Patterns. By analyzing such longest common sequences, the ratio of common part in the Mobile Sequential Patterns are taken as the similarity. Although all these approaches have considered temporal information and location hierarchy, they do not take into account the semantics of locations.

GTS BASED FOLLOWEE RECOMMENDATION

The proposed *GTS-FR* approach is designed a two-phase algorithm, as shown in Figure 3, to address the problem of users' similarity mining for followee recommendation. The first phase deals with the feature extraction (lines 1 to 5), while the second phase explains the followee recommendation (lines 7 to 11). The task of feature extraction explores three aspects that are discussed in Introduction. For a user pair, we explore the *Social Property* (*SP*) as population features which abstract the aggregated number of followee-of-followee links of two users. On the other hand, we explore the *Geographical*

Input: Social Links Set L
Users' Trips T
Users' Textual Information I
Users U
Output: relation between each pair of users
1 *Phase 1. Feature Extraction*
2 Feature Set $F \leftarrow \varnothing$
3 $F \leftarrow F \cup SP(L, I)$
4 $F \leftarrow F \cup GP(T, I)$
5 $F \leftarrow F \cup TP(I)$
6
7 *Phase 2. Feature Extraction*
8 Training Set $T \leftarrow F \cup L$
9 Classifier $C \leftarrow$ SVM(T)
10 Classification Result $R \leftarrow C(U \times U)$
11 Return R

Figure 3. GTS-FR algorithm.

Property (*GP*) between two users to formulate descriptive features of a specific user pair. Moreover, to overcome the experience-limitation problem, *Textual Property* (*TP*) is considered as a feature to represent information needs of users in our recommendation model. The features derived from *Social Property*, *Geographical Property* and *Textual Property* are used to learn a SVM model for each user to classify whether other users could be followed in the phase of followee recommendation. For a user, other users are classified into followee and non-followee classes by the individual SVM model of the user. After checking all users, we obtain all qualified potential followees for the user under examination.

Features from Social Property

As discussed earlier, the traditional social-based friend recommendations could not work well. The reason is that the traditional social-based friend recommendations always make recommendation by friend-of-friend links. The concept of reformation by using friend-of-friend links is that if an user B is a friend of user A's friends, B is likely to be a friend of A. If we directly adopt such recommendation concept, we should modify it by using followee-of-followee link. In other words, such followee-of-followee recommendation strategy is based on the concept that if user X is followed by user Y's followees, X may be followed by Y. Take Figure 4 as an example, we may recommend user *b* to user *k* because user *k* follows user *j* and user *j* follows user *b*. However, recommending followee's followee can not reflect the relation of information need and offered information. In other words, the reason of user *k* following user *j* is different with the reason of user *j* following user *b*. We argue that the "transitivity" of followee-of-followee link should be considered.

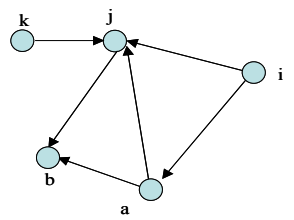


Figure 4. transitivity and followee-of-followee link

Definition 1. Transition-Setter. Given a followee-of-followee link, denoted ($u \rightarrow v$, w), is a directed path in an asymmetrical social network from user u to user v via user w. The middle user w is called *Transition-Setter*.

Accordingly, given a user-user order pair (u, v), here (u, v) $\neq$(v, u), the features extracted from *Social Property* could be generally formulated as (2).

$$SP(u,v) = \sum_{t \in T(u,v)} Transitivity(t) \quad (2)$$

where $T(u, v)$ indicates the set of Transition-Setters of all followee-of-followee links from u to v.

As mentioned above, we can significantly observe that measuring transitivity of two users' Transition-Setters is the key of *Social Property* features. Intuitively, population of common followees of an user and his followers could be utilized for measuring the transitivity of the user because their information needs are satisfied with the information offered by their common followees. As a result, different transitivity, naturally formed in aggregated relations of followers to followees, is embedded in the followers' following behaviors. In an asymmetrical social network data, the most important information is user's following behaviors among users for user transitivity measurement. In the following, we propose to extract two population features to depict users' Transition-Setter as below.

- **Transitivity by Links between Followees and Followers (LinkTran)** - As discussed above, some people will follow another people by cooperating of followee-of-followee link and Transition-Setters' transitivity. Based on this idea, the design idea of LinkTran focuses on the proportion of pairs of follower and followee are linked from follower to followee. Accordingly, we formulate the LinkTran of a Transition-Setter as (3).

$$LinkTran(i) = \frac{1}{|P(i)| \times |S(i)|} \times \sum_{\forall j \in P(i)} \sum_{\forall k \in S(i)} I(j,k) \quad (3)$$

  where $P(i)$ indicates the set of followers of user i, $S(i)$ indicates the set of followees of user i, $I(j, k)$ is an indicator function which indicates whether user j follows user k. Take Figure 4 as an example. The followers of user j are user a user i, and user k. The followee of user j is user b. Thus, the LinkTran of user i and is (1+0+0)/(3×1) ≒ 0.33

- **Transitivity by Communications between Followees and Followers (CTran)** - We employ the χ^2 test for testing relation of texting behaviors of EveryTrail users and their followee. If the test shows significant, it means that the user always comments his followees' trips. Based on the

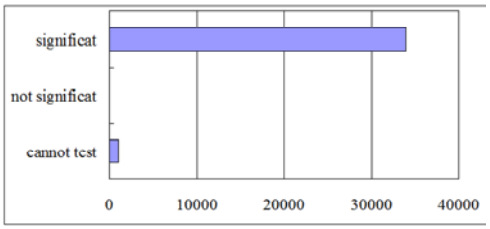


Figure 5. result of χ^2 test.

observations from the EveryTrail dataset, shown in Figure 5, we find most of users will comment their followees' trips. Hence the number of comments is a good index for measuring users' transitivity. Based on the observations, we replace the indicator function, i.e., $I(j, k)$, of formula (3) by following (4).

$$CTran(j,k)=\begin{cases}1-\dfrac{Comment(j,k)}{\max\limits_{f\in S(j)}\{Comment(j,f)\}} & \text{, if user } j \text{ follows user } k\\ 0 & \text{, otherwise}\end{cases} \quad (4)$$

where $Comment(j, k)$ indicates the number of comments offered by user j to user k and $S(j)$ indicates the set of followees of user j.

Features from Geographical Property

As mentioned above, the reason of a user follows other users in the asymmetrical LBSN is either the information need or making friends. To make a complete recommendation, users' interest should be considered because people always make friend who have similar interest. In EveryTrail website, there are two kinds of user-generated data could reflect their interest, i.e., trips and tags of trips, as shown in Figure 6. The trip is also called trajectory typically consists of a sequence of geographic points (represented as <latitude, longitude>). The trajectory could reflect the detail of user's activity. On the other hand, the tag of trajectory could reflect the high level concept of user's activity. Each trajectory just have only one tag, e.g., hiking, biking, etc.

Accordingly, given an ordered user pair (u, v), the features extracted form Geographical Property could be generally formulated as (7).

$$GP(u,v)=\frac{1}{|Tr(u)|\times|Tr(v)|}\times\sum_{\forall p\in Tr(u)}\sum_{\forall q\in Tr(v)}Similarity(p,q) \quad (7)$$

where $Tr(u)$ indicates the set of trajectories of user u.

We can significantly observe that measuring similarity of two trajectories is the key of Geographical Property features. Intuitively, the regions which user stays in could reflect the user's preference. As a result, for each user, we adopt the notion of *stay locations* [11] to represent the users' movement behavior as shown in Figure 7. To discover stay locations, we first detect the regions, called stay points, where a user stayed in, i.e., *s1* and *s2* in Figure 7. Then we cluster all detected stay points to form stay locations, i.e., *location2* and *location5* in Figure 7. As shown in Figure 7, the trajectory could be transformed as the sequence <*location2*, *location5*>. As the result, the similarity measurement could be modeled as the sequences matching problem.

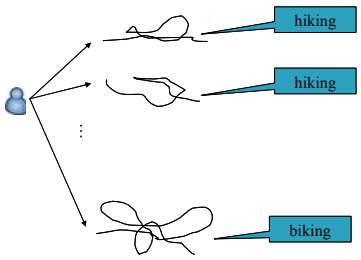


Figure 6. trips and tags of trips.

Given two sequences, we argue that they are more similar when they have more common parts. Thus, we use the *Longest Common Sequence* (*LCS*) of these each pair of sequences to represent their longest common part. For example, given a sequence $P = <A, B, C, D>$ and a pattern $Q = <A, D, C>$, their longest common sequence is $LCS(P, Q) = <A, C>$. Accordingly, we define the *participation ratio* of the common part to a pattern P as follows.

$$ratio(LCS(P,Q),P)=\frac{|LCS(P,Q)|}{|P|} \quad (8)$$

Intuitively, the tags of two sequences could reflect basic concepts of them. Therefore, the similarity of two sequences will be evaluated as 0 if their tags are different. Thus, we calculate the similarity of two sequences by averaging the participation ratios of their common part to them. Given sequences P and Q, a simple approach is to directly compute the average of the two ratios to P and Q, as shown in Equation (9). Thus, we call this approach *Equal Average (EA)*. On the other hand, as shown in Equation (10), we can compute the *Weighted Average (WA),* in proportion to the lengths of the two sequences. The argument is that a longer pattern provides more information about user behaviors than a shorter pattern. Therefore, the longer pattern gives more weight than the shorter one in measuring the similarity between two sequences.

$$Similarity_{EA}(P,Q)=I_T(P,Q)\times\frac{ratio(LCS(P,Q),P)+ratio(LCS(P,Q),Q)}{2} \quad (9)$$

$$Similarity_{WA}(P,Q)=I_T(P,Q)\times\frac{|P|\times ratio(LCS(P,Q),P)+|Q|\times ratio(LCS(P,Q),Q)}{|P|+|Q|} \quad (10)$$

where $I_T(P, Q)$ is an indicator function which indicates whether the tags of P and Q are the same. Note that we could extract two features from Geographical Property, namely *EA* and *WA*.

Features from Textual Property

As discussed earlier, we intend to exploit the users' information needs in LBSN for matching other users' provided information by a HITS-Based random walk model [3]. We believe that users comment other users' trip or write travelogue within their trips can represent their information needs. Therefore, we build a *User-Keyword* (*UK*) graph, which consists of users and keywords connected in accordance with the textual records. Let $t(u_i, w_j, l_s) \in TI$ denotes a textual record describing that user u_i has provided textual information which contains the keyword w_j and associate the location l_s, where *TI* indicates

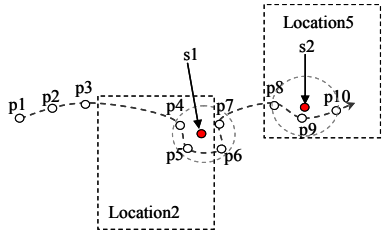


Figure 7. transitivity and followee-of-followee link

the collection of all textual records. Here, the keywords are extracted from all textual information with high TFIDF value. Definition 2 gives the formal definition of the *UK* graph.

Definition 2. User-Keyword (UK) Graph, denoted by $G_u(V_u,E_u)$, is an undirected bipartite graph (as illustrated in Figure 8(a)). Here $V_u = U \cup K$, where U and K are the sets of all users and keywords, respectively, and $E_u = \{e_{i,j} \mid t(u_i, w_j, \cdot) \in TI\}$, where $t(u_i, w_j, \cdot)$ denotes that user u_i has texted keyword w_j in some textual information. In this graph, each edge $e_{i,j} \in E_u$ is weighted by the number of keyword w_j has been texted by user u_i.

Given m users and n keywords, we build an $m \times n$ adjacency matrix M for *UK* Graph. Formally, $M = [c_{ij}]$, $0 \leqq i < m$; $0 \leqq j < n$, where c_{ij} represents how many times the i_{th} user has texted the j_{th} keyword. Formally, the random walk model applied to UK Graph can be described as follows:

$$\begin{aligned} x_{keyword}^{k+1} &= (\varepsilon M_{col}^{T} + (1-\varepsilon)\delta_1)x_{user}^{k} \\ x_{user}^{k+1} &= (\varepsilon M_{row} + (1-\varepsilon)\delta_2)x_{keyword}^{k+1} \end{aligned} \quad (11)$$

where k is the number of iterations, M_{col} is the column stochastic matrix of M (M_{col} is computed by normalizing each column in M), M_{row} is the row stochastic matrix of M (M_{row} is computed by normalizing each row in M), δ_1 is a matrix with all elements equal to $1/m$, δ_2 is a matrix with all elements equal to $1/n$, and ε is the "teleport probability," which represents the probability of a random surfer teleporting from a keyword node to a user node (respectively from a user node to a keyword node) instead of following the links in *UK* Graph.

As the above-mentioned random walk model, the users' relevance can be obtained. However, such random walk model do not consider the relationship among keywords. We argue that the similar keywords could represent similar information needs. Intuitively, similar keywords could be texted with the same locations. Therefore, we build a *Location-Keyword* (*LK*) graph, where the locations are the same as stay locations which is extracted in the *Geographical Property* feature extraction step. Definition 3 gives the formal definition of the *LK* graph.

Definition 3. Location-Keyword (LK) Graph, denoted by $G_l(V_l, E_l)$, is an undirected bipartite graph (as illustrated in Figure 7(b)). Here $V_l = L \cup K$, where L and K indicate the sets of all locations and keywords, respectively, and $E_l = \{e_{j,s} \mid t(\cdot, w_j, l_s) \in TI\}$, where $t(\cdot, w_j, l_s)$ denotes that location l_s has been texted with keyword w_j in comments or travelogues. In this graph, each edge $e_{j,s} \in E_l$ is weighted by the proportion of keyword w_j that has been texted in the comments or travelogues of user location l_s.

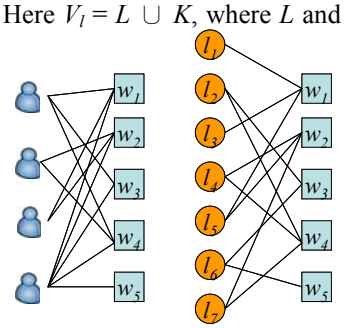


Figure 8. trips and tags of trips.

Given r locations and n keywords, we build an $n \times r$ adjacency matrix N for *LK* Graph. Formally, $N = [v_{ij}]$, $0 \leqq i < n$; $0 \leqq j < r$, where v_{ij} represents how many times the i_{th} keyword appears in the textual information associated with the j_{th} location. Formally, the random walk model applied to *LK* Graph can be described as follows:

$$\begin{aligned} x_{keyword}^{k+1} &= (\varepsilon M_{col}^{T} + (1-\varepsilon)\delta_1)x_{user}^{k} \\ y_{location}^{k+1} &= (\varepsilon N_{col}^{T} + (1-\varepsilon)\delta_2)x_{keyword}^{k+1} \\ y_{keyword}^{k+1} &= (\varepsilon N_{row} + (1-\varepsilon)\delta_3)y_{location}^{k+1} \\ x_{user}^{k+1} &= (\varepsilon M_{row} + (1-\varepsilon)\delta_4)y_{keyword}^{k+1} \end{aligned} \quad (12)$$

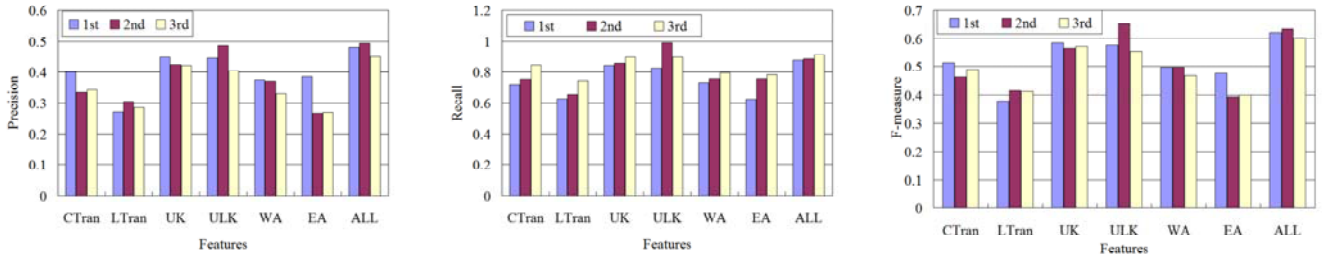
where k is the number of iterations, M_{col}, M_{row}, ε, δ_1 and δ_2 are the same as the random walk model applied to *UK* Graph, i.e., formula (11). Similarly, N_{col} is the column stochastic matrix of N (N_{col} is computed by normalizing each column in N), N_{row} is the row stochastic matrix of N (N_{row} is computed by normalizing each row in N), δ_3 is a matrix with all elements equal to $1/n$, δ_4 is a matrix with all elements equal to $1/r$. Note that we could extract two features from *Textual Property*, namely *UK* and *LK*.

Followee Recommendation

After the phase of feature extraction, features derived from all of social, geographical and textual properties are used as inputs for the followee recommendation phase to learn a classification model for each individual user. We choose SVM as the classifier because it has shown excellent performance in similar tasks [2, 4, 6]. The reason why we select SVM as our classifier is that SVM is hard to be effected by class-imbalanced problem. In our approach, for each user, all of other users are used for his SVM training, i.e., an instance followed by the user under examination is considered as a positive example, while users without being followe by the user serve as negative examples. For instance, users followed by *user 1* are positive examples for a classifier for *user 1*, but negative examples for a classifier for *user 2*.

EXPERIMENTS

In this section, we conduct a series of experiments to evaluate the performance for the proposed GTS-FR using EveryTrail dataset . All the experiments are implemented in Java JDK 1.6 on an Intel Core i7-2600 CPU 3.40 GHz machine with 7GB of memory running Microsoft Windows win7. We first describe the data preparation on the EveryTrail dataset and then introduce the evaluation methodology. Finally, we show our experimental results for following discussions.



(a) (b) (c)

Figure 9. Comparison of Various Features

EveryTrail Dataset

EveryTrail is a trip-sharing and social networking website on which users can upload, share and find trips. On EveryTrail, users can upload GPS logs and write travelogues and comments within a trip. Users also can label a tag on a trip. While the EveryTrail website provides the public API to let other applications integrate with their service, some functionality in the API is broken. For this reason, we mainly use the API and with crawling web pages as support to get all the data we need. We extract the data from 12/2011 to 3/2012, each month is a time period. We got 35,153 users and 4 snapshots. The data description of each snapshot is given in Table 1.

Snapshot	1st	2nd	3rd	4th
# of trips	116179	145,662	193,331	196,949
# of comments	337,519	293,453	315,585	379,020
# of links	700,103	777,738	1,056,077	1,139,832

Table 1. Data Description of Each Snapshot

All of the data is divided into the training data and the testing data. The first, second and third snapshots are formed as the training data, and the remaining snapshots are formed as the testing data. For example, if we select the first snapshot as training data, the testing data will be extract second snapshot. Since the problem we address is followee recommendation, we only care about the following links which are not created in training data. Thus, the testing data will be formed by ordered user pairs who do not link in training data.

Evaluation Methodology

The follows are the main measurements for the experimental evaluations. The Precision, Recall and F-measure are defined as Equations (13), (14) and (15), where p^+ and p^- indicate the number of correct recommendations and incorrect recommendations, respectively, and R indicates the total number of links in the testing data.

$$\text{Precision} = \frac{p^+}{p^+ + p^-} \tag{13}$$

$$\text{Recall} = \frac{p^+}{R} \tag{14}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

We divide the experiment into two parts: 1) Comparison of Various Factors or Features (i.e., Internal Experiments) and 2) Comparison of Existing Recommenders (i.e., External Experiments). For the comparison of various features, we first compare the performance of our proposed Social Property, Geographical Property and Textual Property. Then, we compare the effectiveness of all of features. For the comparison of existing recommenders, we compare the effectiveness of GTS-FR with HGSM-based recommender [11] and followee-of-followee strategy in terms of Precision, Recall, and F-measure.

Comparison of Various Features

This experiment evaluates effectiveness of each factor in the proposed GTS-FR in terms of Precision, Recall, and F-measure. Figure 9 shows that, on average, all of Precision, Recall, and F-measure value of GTS-FR, under different features, i.e., CTran, LTran, UK, ULK, WA and EA, respectively. We observe that all of Precision, Recall, and F-measure of UK and ULK are better than those of other four features. The result shows that the features extracted from Textual Property are more important than those extracted from other features. If we focus on comparison of UK and ULK, overall, the UK is more stable but ULK could achieve highest recall. This is because the locations we detect are always changed. Sometimes, the change might benefit effectiveness but not usually. Moreover, we also can observe that the values of recall are always greater than the values of Precision. The reason is that social link is created slowly because the Everytrail website is established a long time. There are many links we recommend are created in the future but not in the next snapshot. Accordingly, we analyze the incorrect recommendations by the first snapshot model, which is tested by the second snapshot, whether they will become correct in the further snapshot. As shown in Figure 11, we can find that most incorrect recommendations become correct in the further

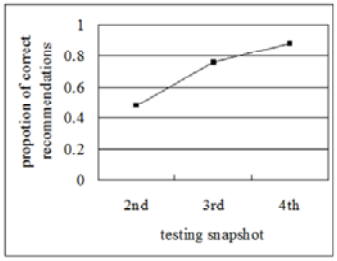
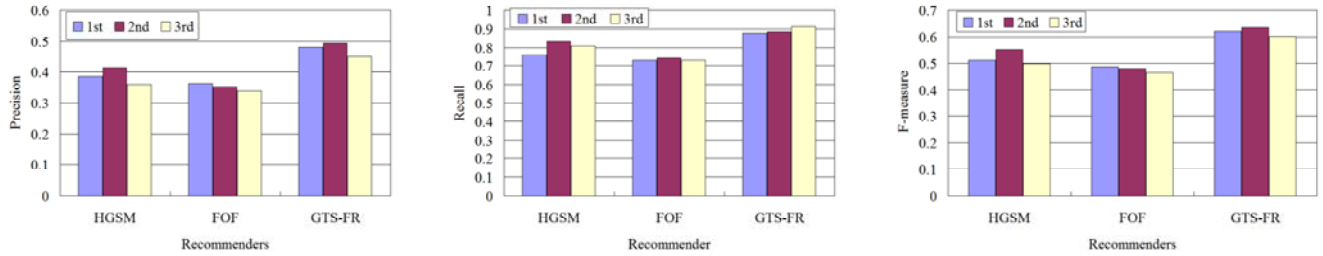


Figure 11. Analysis of incorrect recommendations.



(a) (b) (c)

Figure 12. Comparison of Various Recommenders

snapshot.

Comparison with Existing Recommenders

This experiment evaluates the effectiveness of our proposed GTS-FR comparing HGSM-based recommender and followee-of-followee strategy (FOF) in terms of Precision, Recall, and F-measure. HGSM-based recommender relays on the similarity of users' uploaded trajectory. It is similar to our proposed factor Geographical Property but more effective. followee-of-followee strategy (FOF) is widely used for followee recommendation in many existing LBSN websites. It is similar to our proposed factor Social Property. Figure 12 shows GTS-FR outperforms HGSM-based recommender and followee-of-followee in terms of Precision, Recall, and F-measure. The reason is that we consider users' relationship in the factor of users' information need, reflected by textual information, while other methods do not.

CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a novel approach named *Geographic-Textual-Social Based Followee Recommendation* (*GTS-FR*) for recommendation of interesting followees by mining users' information needs. Meanwhile, we have tackled the problem of user texting behaviors mining in information need discovering, which is a crucial prerequisite for effective recommendation of followees in a LBSN. The core task of followee recommendation in a LBSN can be transformed to the problem of the problem of binary classification. We evaluate the possibility of each ordered user pair by learning an SVM model. In the proposed *GTS-FR*, we have explored i) *Social Property* (*SF*), ii) *Geographical Property* (*GP*) and iii) *Textual Property* (*TP*) by exploiting the LBSN data to extract descriptive features. To our best knowledge, this is the first work on followee recommendation that consider social property, geographical property and textual property in LBSN data, simultaneously. Through a series of experiments by the real dataset obtained from EverTrail, we have validated our proposed *GTS-FR* and shown that *GTS-FR* has excellent effectiveness under various conditions. As for the future work, we plan to design more advanced link prediction strategies to further enhance the quality of followee recommendation for location-based social networks.

ACKNOWLEDGMENTS

This research was supported by National Science Council, Taiwan, R.O.C. under grant no. NSC101-2221-E-006-255-MY3 and NSC100-2218-E-006-017.

REFERENCES

1. EveryTrail: http://www.everytrail.com/.
2. Boutell, M. R., Luo, J., Shen, X. and Brown, C. M. Learning multi-label scene classification. Pattern Recognition, 37(9):1757–1771, 2004.
3. Cao, X., Cong, G. and Jensen, C. S. Mining significant semantic locations from GPS data, Proceedings of the VLDB Endowment, v.3 n.1-2, September 2010
4. Clare, A. and King, R. D. Knowledge Discovery in Multi-label Phenotype Data. In European Conference on PKDD, pages 42–53, 2001
5. Lee, J.-G., Han, J. and Whang, K.-Y. Trajectory Clustering: A Partition-and-Group Framework. In Proceedings of ACM SIGMOD, pp. 593-604, Jun. 2007.
6. Li, T. and Ogihara, M. Detecting emotion in music. In International Society for Music Information Retrieval Conference (ISMIR), 2003.
7. Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W. and Ma, W.-Y. Mining User Similarity Based on Location History. In Proceedings of ACM GIS, Irvine, CA, USA, Nov. 2008.
8. Lu, E. H.-C. and Tseng., V. S. Mining Cluster-Based Mobile Sequential Patterns in Location-Based Service Environments. In Proceedings of IEEE MDM, May. 2009.
9. Ye, M., Yin, P. and Lee, W.-C. Location Recommendation for location-based Social Network. In Proceedings of GIS, pages 458-461, 2010.
10. Ying, J. J.-C., Lu, E. H.-C., Lee, W.-C., Weng, T.-C. and Tseng, V. S. Mining User Similarity from Semantic Trajectories. In Proceedings of LBSN' 10, San Jose, California, USA, November 2, 2010.
11. Zheng, Y., Zhang, L. and Xie, X. Recommending friends and locations based on individual location history. ACM Transaction on the Web, 2011
12. Zheng, Y. Location-based social networks: Users. Computing with Spatial Trajectories, Yu Zheng and Xiaofang Zhou, Eds. Springer, 2011.