JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 35, 429-445 (2019)
DOI: 10.6688/JISE.201903_35(2).0011

Guest Recommendation for Holding Social Activities Among Friends Through Social Platforms*

CHIH-HUA TAI
Department of Computer Science and Information Engineering
National Taipei University
New Taipei City, 23741 Taiwan
E-mail: hanatai@mail.ntpu.edu.tw

As more and more people rely on social platforms such as Facebook for holding social activities among friends, this paper addressed the need of guest recommendation to ease the task of selecting proper guests from a large number of friends for invitation to kinds of specific types of activities. For the problem, this paper proposed the Guest Invitation to Hosts (GIH) algorithm to learn the dominant factors behind the guest invitations from historical data. Concerning that there are usually multiple factors dominating the guest invitation and the factors differ from the types of social activities, GIH consists of two learning mechanisms: (1) the learning of discrimination between activity types, by the topic words and descriptions of activities; and (2) the learning of latent dominant factors for guest invitation (to specific types of activities), by the Non-negative Matrix Factorization and Top-*k* Frequent Pattern mining techniques. Evaluated on the Facebook data, GIH can reach an accuracy of at least 80% in the guest recommendation, and outperform naïve approaches in terms of precision, recall, F1 score and accuracy. The case studies showed that the dominant factors (rules) identified by GIH comply with the human intuition. The latent dominant factors (rules) identified by GIH for guest recommendation can further be used as references for advanced studies in social science.

Keywords: social activity, guest recommendation, matrix factorization, frequent pattern, clustering

1. INTRODUCTION

Social activity is a critical behavior in human life. Recently, it has been shown that the ways of people's social activities have gradually changed due to the popularity and convenience of social platforms such as Facebook, Twitter and Sina Weibo. For example, before the rise of social platforms, people cared friends, made new friends, held activities, and communicated with each other by visiting, calling, or writing letters. Nowadays, through the social platforms, people care friends by following pieces of online posts including texts, photos and videos, make new friends through online communities, hold activities by spreading online messages, and communicate with each other by twittering and re-twittering. The cost of social activities is spectacularly reduced, which further attracts more people to join social platforms and to rely on its convenience.

According to the announcement of Facebook, the number of monthly active users reached 1 billion in October 2012 and exceeded 1.86 billion in December 2016. When much more people have relied on social platforms, various applications for social activities have been expected. For example, Facebook used user profiles and social connec-

Received September 26, 2017; revised October 31, 2017; accepted November 29, 2017.
Communicated by Wen-Chih Peng.
* This work is supported in part by MOST though grants MOST 106-2221-E-305-014 and MOST 105-2221-E-305-012.

tions to help users get back their old friends; Facebook online games recommended new friends of common interests; automatic information pushing mechanism (such as post notification and birthday reminder) facilitated users to maintain friendships well; convenient social-platform embedded tools were developed for online organizing social activities and calling for participations. In addition, there were also many works [1-9] researching into novel social functions. All these achievements aimed at making all kinds of social activities more convenient. This paper addressed the need of guest recommendation for holding social activities among friends through social platforms. Note that due to the characteristics, *i.e.*, friend list, activity page, real-time event notification and calendar reminder, of social platforms, the task of holding social activities among friends is easier than before. However, this task is still time-consuming and challenging for activity hosts because of the large amount of friends and the multiple factors concerning the invitation to a specific social activity. For example, consider that a teenager is going to hold a personal concert. He/she needs to concern about his/her closeness to the invited friends, the interest and location of the invited friends, and so on. That is, he/she needs to think of 1-3 factors before determining whether or not to invite one friend to the personal concert. According to the survey from the Pew Internet & American Life Project, together with Harvard's Berkman Center, 94% of teenagers used Facebook and had 425.4 friends on average [10]. The teenager then needs to repeat the thinking hundreds of times, which is a heavy task for human brain. Therefore, guest recommendation is a practical problem and need.

Specifically, this paper studied a guest recommendation problem, whose purpose is to recommend activity hosts a ranked list of friends regarding a specific activity. For the problem, a novel method referred to as the Guest Invitation to Hosts (GIH) algorithm is proposed as an effective solution. Concerning that usually there are multiple factors dominating the guest invitation and the factors differ from the types of social activities, GIH consists of learning mechanisms of discriminating activity types and extracting the factors dominating the guest invitation to specific types of activities. To learn to discriminate activity types, GIH utilizes the words in the topics (*e.g.*, known as the event name in Facebook) and descriptions in the activity pages, as the topics and descriptions are expected to provide meaningful semantics to all participants. For learning the dominant factors, GIH constructs the host-guest relational matrices and uses Non-negative Matrix Factorization (NMF) and Top-k Frequent Pattern mining (TFP) techniques to extract the important/frequent relations from the matrices. Evaluated on the Facebook data, which consists of 9 social activities of 3 different types and totally 1,322 friends of 9 hosts, the GIH algorithm can reach an accuracy of at least 80% in the guest recommendation and outperform naïve approaches in terms of precision, recall, F1 score and accuracy. The case studies further showed that the dominant factors (rules) identified by GIH comply with the human intuition.

2. RELATED WORK

At the present stage, research on guest recommendation in social networks is still under development, and there are only a few of researches related to social activities. Table 1 compared the works related to guests of social activities and their problem defi-

nitions. Table 1 also compared social network platforms that enable users to organize activities online, such as Facebook and Meetup.

Table 1. Comparison between related works and systems.

	Problem definition	Using Interest Toward Activity?	Using Personal Interest Additional to Activity?	Using Social Tightness?	Incorporating Hosts?	Ranking?
WASO [7]	Find a set of guests that maximizes the group willingness to an activity.	Yes	No	Yes	No	No
PSGA [6]	Find a set of guests that maximizes the group willingness to an activity and optimally fits the activity cost.	Yes	No	Yes	No	No
MRGQ [4]	Find a set of guests that are more likely to make new friends in an impromptu activity.	No	No	Yes	No	No
HMGF [5]	Find a set of guests that are more likely to make new friends in an activity.	No	No	Yes	No	No
STGQ [11]	Find a set of guests that fits the activity time and social constraints.	No	No	Yes	Yes	No
SSGQ [9]	Find a set of guests that fits the activity spatial and social constraints.	No	No	Yes	Yes	No
Facebook Event Function	Recommend guests that contact frequently with the activity host.	No	No	Yes	Yes	Yes
Meetup Calendar	Create an activity for personal preference groups.	No	Yes	No	No	No
This work	Recommend a guest list that a host may want to invite for an activity.	Yes	Yes	Yes	Yes	Yes

Briefly speaking, Meetup and Facebook are social network platforms that enable users to organize activities online. Meetup forms groups according to personal interests and provides particular pages for activities. However, Meetup does not have the function of guest recommendation for forming new activities. Facebook provides a mechanism that recommends as the guests the friends who contact the most frequently with the hosts. But such an approach may miss the important friends who attend less online social activities with the hosts.

For guest recommendation, one group of studies aims at maximizing the participation willingness in activities. Reported by [12] and [13], there are usually two factors to affect the willingness of an attendee to participate in an activity: (1) having a certain degree of interest in the activity; (2) being familiar with a certain number of guests participating in the activity. Based on such theories, [7] measured the participation willingness

of each guest according to the guest's interest toward an activity and common friends with other guests, and formulated a problem called Willingness mAximization for Social grOup (WASO). Another work [6] took an additional consideration of activity cost and formulated the Participant Selection for Group Activity (PSGA) problem based on the same concept in WASO. The disadvantage of applying the above solutions to the problem proposed in this work is that the optimal group of recommended guests may not be desirable for the host since it lacks of the concept of hosts in the solutions. Consequently, in such cases, the activities are not going to be regarded successful from the hosts' point of view.

Another group of studies aims at recommending guests that are more likely to become friends with each other. The work [5] formulates a problem called Hop-bounded Maximum Group Friending (HMGF), and [4] formulates another one called Multiple Rally-Point Social Spatial Group Query (MRGQ). Based on the belief that people having common friends are more likely to become friends, the former utilizes the potential friendships happening along with the existing friendships to figure out the invitation group with the maximum likelihood of new friend making. Considering the spatial distance as the major concern for the guests to attend an activity, the latter identifies an invitation group whose spatial distance to the activity location is minimal within a specified range and whose intra-unfamiliarity is small enough to ensure the ease of making new friends. The major drawback of the above solutions to the problem of this work is that the interest toward the activity is not taken into consideration. As a result, the suggested guests may have little interest in the activity and reject the participation. Furthermore, these solutions also lack of the point of views from the activity hosts.

Indeed, still another group of works addresses the concept of activity hosts for guest recommendation. The work [11] addresses the need of concerning the guests' available time together with their social closeness to formulate a problem called Social-Temporal Group Query (STGQ). It thus also selects a proper activity period in addition to the invitation guests. The other work [9] focuses on organizing impromptu activities, and addresses the concerns of location constraint together with the social relationships for the so called Socio-Spatial Group Query (SSGQ) problem. There are two reasons making these solutions inappropriate to the problem of this work. First, they do not take the activity topic and the friends' interests into consideration, and thus may recommend a group of guests who are close to each other but have little interest in the activity. Second, these works rank all the host's direct friends equally for guest recommendation. It is therefore unable to reduce the task burden of guest selection for activity hosts.

3. PRELIMINARIES

3.1 Non-negative Matrix Factorization (NMF)

Non-negative Matrix Factorization (NMF) [14, 15] is a matrix factorization algorithm focusing on matrices with nonnegative values. Given a data matrix $A = [a_{ij}] \in \mathbb{R}^{m \times n}$, the objective of NMF is to find two nonnegative matrices $W \in \mathbb{R}^{m \times p}$, $H \in \mathbb{R}^{p \times n}$, $p \leq n$, such that the approximation of A by $A \simeq WH$ can be achieved by minimizing the following error function

$$\min_{p} \| A - WH \|_F^2,$$

where $\| \cdot \|_F$ represents the Frobenius norm of a matrix, *i.e.*, the squared root of the squared sum of all the elements in a matrix. From the row vectors of W, a maximum value located in a certain column vector represents that the corresponding document is classified in the corresponded group. Similarly, from the row vectors of H, several relatively large values located in some certain column vectors represents the relatively important words in a certain group. Due to the high interpretability of NMF, it is commonly preferred in document clustering [15] and various researches that do not need data with negative values. In this paper, GIH utilizes the advantages of NMF to find the important factors for guest invitation from the host-guest relational matrix.

3.2 Top-*k* Frequent Pattern Mining (TFP)

Top-*k* Frequent Pattern Mining (TFP) is a variation of Frequent Pattern Mining [16], which is to mine patterns with frequencies higher than a user-specified threshold min_{freq}. Differently, TFP prefers a user-specified number k and a minimal length min_{len} as the input parameters, and aims at mining the top-k frequent closed patterns whose length is at least min_{len}. As specified in the researches [17, 18], taking top-k and min_{len} rather than min_{freq} has the following advantages in practical applications. First, it is intuitionally easier for users to determine the values of k and min_{len} (such as top-10 coupled books sold). Second, the number of discovered patterns can be expected, *i.e.* by k, and the subsequent analysis or applications can thus stay focus. Third, the setting of min_{len} could be optional, which provides good flexibility for practical applications. In this paper, GIH therefore uses TFP to find the frequently co-occurrent factors as the rules for guest invitation.

4. GUEST INVITATION TO HOSTS (GIH)

4.1 Overview of GIH Mechanisms

The purpose of the GIH algorithm is to recommend activity hosts a ranked list of friends regarding the types of social activities. In order to learn the concerns (factors) behind the actions of inviting specific guests to specific activities, GIH is composed of two mechanisms as shown in Fig. 1. One is to learn to discriminate activity types, since the factors dominating the invitation differ from different types of social activities. GIH implements the mechanism by utilizing activity topic words and descriptions, because the activity topics and descriptions are expected to provide meaningful semantics to all participants. The other mechanism is to extract and learn the dominant factors of guest invitation regarding the types of social activities. For this mechanism, GIH learns from historical data by constructing the host-guest relational matrices and applying Non-negative Matrix Factorization (NMF) and Top-*k* Frequent Pattern mining (TFP) techniques to extract the important/frequent relations from the matrices. After the learnings, GIH can perform the guest recommendation given the topic and description of an activity and the friend list of the activity host.

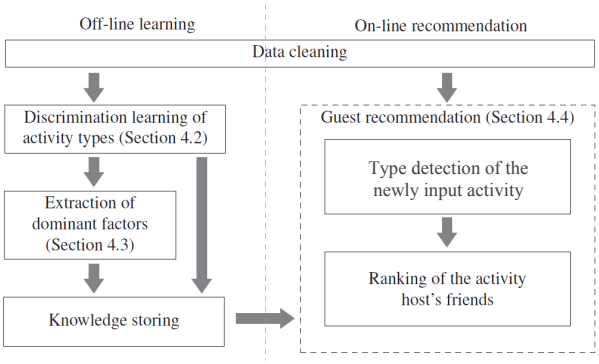


Fig. 1. The working flow of the GIH algorithm.

Specifically, GIH runs offline learning process and online recommendation process, respectively. In the offline learning process, GIH takes a set of historical social activity data including the friend lists of activity hosts, activity topics, descriptions and invited guests as the inputs. Given the inputs, GIH first runs a data cleaning process (1) to remove redundant words (*e.g.*, 'is', 'a' and so on) from activity topics and descriptions for the subsequent learning of discrimination of activity types, and (2) to construct a host-guest relational matrix of each activity for the subsequent learning of extraction of dominant factors. The host-guest relational matrix is a Boolean matrix representing the common characteristics (*e.g.*, education, job, interest, logs of participated activities, and so on) between the invited guests and the activity host. After then, GIH learns the discrimination of activity types (in subsection 4.2) according to the activity topics and descriptions by a clustering technique, and extracts the dominant factors regarding each type of activities (in subsection 4.3) by the NMF and TFP techniques. Finally, all the knowledge learned in the offline process is stored in a database.

An online recommendation process is launched, whenever the topic, description, and friend list of the host of a specific activity are given. After running the data cleaning process similar to the offline process, GIH finds the type of the input activity and calculates the invitation ranking score for each friend of the host (in subsection 4.4) according to the dominant factors offline-learned to be associated with that type of activities.

4.2 Learns the Discrimination of Activity Types

GIH learns the discrimination of activity types by clustering techniques due to the challenges listed in the following. First, the diversity of activities is large. This means that it is impractical to manually specify the differences between all types of activities. Second, the types of activities (such as a personal graduation concert) sometimes could be unclear. Therefore, instead of manually specifying the discrimination rules or learning by classification, GIH learns the discrimination of activity types from historical data by clustering on historical activities according to their topic words and descriptions. Clustering activities into groups of similar types also has the advantages of cleaning less-related data and increasing the volume of related data, which makes the dominant factors extracted in the subsequent step more reliable.

By clustering, GIH's intuition behind the learning is that activities of the same types will use similar keywords in the topics and descriptions in order to provide correct semantics to all participants. That is, the words appearing in the topics and descriptions are usually derived forms or synonyms for activities of the same types, although different activity hosts may use different words. Therefore, GIH considers the synonyms in the definition of similarity function of two activities. In addition, GIH weights the words differently. It is generally believed that the topic words are the most important, and the other words in the descriptions could be also important if they are derived forms or synonyms of the topics words. Consequently, the similarity between two activities d_i and d_j is defined through the semantics and importance of the words appearing in the topics and descriptions, *i.e.*,

$$Sim(d_i, d_j) = \frac{\sum_{v_x \in d_i, v_y \in d_j} Relevance(v_x, v_y) \times Weight(v_x, d_i) \times Weight(v_y, d_j)}{\sum_{v_x \in d_i, v_y \in d_j} Relevance(v_x, v_y)}, \quad (1)$$

where

$$Relevance(v_x, v_y) = \begin{cases} 1 & \text{if } v_x \text{ and } v_y \text{ are derived forms or synonyms} \\ 0 & \text{otherwise} \end{cases},$$

$$Weight(v, d) = \begin{cases} w_h & \text{if } v \in Topic(d) \\ w_m & \text{if } v \notin Topic(d) \text{ and } \exists u \in Topic(d), Relevance(v,u) = 1, \\ w_l & \text{otherwise} \end{cases}$$

and $Topic(d)$ represents the set of words appearing in the topic of an activity d. The denominator term in Eq. (1) is used for the normalization to prevent from leading to higher similarities for activities with longer topics and descriptions.

Table 2. Examples of the topics and descriptions of two activities.

Activity	Topic	Description
1	Music concert	Celebrate the musical talents of the students
2	Choir concert	An evening of celebration

Example 1: Consider the two activities listed in Table 2. We have

$d_1 = \{music, concert, celebrate, musical, talents, students\}$,
$d_2 = \{choir, concert, evening, celebration\}$.

Accordingly, the pairs of words from d_1 and d_2 with a relevance value 1 include

$Relevance(concert, concert) = 1$,
$Relevance(celebrate, celebration) = 1$.

Note that $Weight(concert, d_1) = Weight(concert, d_2) = w_h$ since both words belong to the topics of the corresponding activities. $Weight(celebrate, d_1) = Weight(celabration, d_2) = w_l$ because they neither belong to the topics nor the derived forms or synonyms of the topic words. Given $w_h = 1$, $w_m = 0.5$, and $w_l = 0.1$, the similarity is then calculated as

$$Sim(d_1, d_2) = \frac{(1 \times 1 \times 1) + (1 \times 0.1 \times 0.1)}{2} = 0.5. \quad \square$$

Given the similarities between all pairs of activities, GIH then applied the complete link algorithm [19] for clustering, because the complete-link ensures that any pair of activities in the same clustering group is sufficiently similar to each other. Finally, GIH stores the clustering results together with the topic words and descriptions of the corresponding activities.

Later in the online recommendation process (detailed in subsection 4.4), the type of the newly input activity can be determined according to the stored knowledge.

4.3 Learns the Dominant Factors

GIH learns the dominant factors behind the guest invitation to each type of activities by the NMF and TFP techniques. The NMF is used to extract the relatively important factors (*i.e.* the characteristics that the guests and host have in common) from each activity, while the TFP is applied to identify the dominant rules from the aspect of a specific type of activities.

Specifically, as illustrated in the overview of GIH and the preliminary, a host-guest relational matrix of each activity is a Boolean matrix representing the common characteristics between the invited guests and activity host. That is, $a_{ij} = 1$ if the ith guest has the jth character in common with the activity host, and $a_{ij} = 0$, otherwise. Then, similar to the clustering on general documents [19], by performing the non-negative matrix factorization $A \simeq WH$, the largest value in each row of the resulting matrix W indicates the cluster each guest belonging to, while the relatively large values in each row of the other resulting matrix H reveal the important factors of each cluster. Formally, given the number of clusters set as p on the non-negative matrix factorization of a host-guest relational matrix A, the set of important factors for each cluster $q (q \in [1, p])$ is defined as follows.

$$F_q^p(A) = \{f_j\}$$

where f_j is the jth character corresponding to the relatively large value in the jth column of the qth row in H. Here for each cluster q, GIH distinguishes the characters with relatively large values from those with relatively small values by the complete-link algorithm with the number of clusters set as 2. As a result, the collections of the sets of important factors corresponding to non-empty clusters can be regarded as the important factors for guest invitation to a specific activity.

Example 2: Consider an activity with its host-guest relational matrix A consisting of 5 guests and 7 factors as follows.

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}_{5 \times 7}$$

When the number of group $p = 3$, the resulting W and H of NMF are

$$W = \begin{bmatrix} 1.8 & 0 & 0.3 \\ 1.8 & 0 & 0.3 \\ 2 & 0.4 & 0.2 \\ 0 & 1.6 & 1.1 \\ 0 & 1.7 & 0 \end{bmatrix}_{5\times 3},$$

$$H = \begin{bmatrix} 0.3 & 0.5 & 0.5 & 0.5 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.6 & 0.5 & 0.5 \\ 0.9 & 0 & 0 & 0 & 0 & 0.1 & 0 \end{bmatrix}_{3\times 7}.$$

According to the values of each row in W, the first three guests belong to the 1st group since they have the largest values in the first column of their rows. Likewise, the last two guests are the 2nd group, while no one in the third group. On the other hand, according to the values of each row in H, the sets of relatively important factors for the 1st and 2nd groups are $F_1^3(A) = \{f_2, f_3, f_4\}$ and $F_2^3(A) = \{f_5, f_6, f_7\}$, respectively. Note that the factor set of the 3rd group is neglected since it contains no guests. Hence, $F_1^3(A) = \{f_2, f_3, f_4\}$ and $F_2^3(A) = \{f_5, f_6, f_7\}$ are learned as the important factors to invite different groups of guests to the specified activity. ❑

Nevertheless, note that the relatively important factors learned from one pair of resulting matrices W and H of NMF can have two problems: (1) sensitive to the parameter of cluster number p, which generally varies as the types of activities and the hosts' personal concerns, and (2) bound to the host-guest relational matrix A associated to one particular activity. Such factors will thus incur parameter effects and personal concerns as noises, and reduce the generality for guest invitation to a type of activities.[1]

In order to get rid of such noises for reliable guest recommendation, GIH performs a second-stage extraction to learn the dominant rules from the aspect of a group of activities of the same type. Specifically, a factor-set collection T is first formed by the factor sets learned with respect to various p given A associated to different activities of the same type, *i.e.*,

$$T = \bigcup_{A,p,q\in[1,p]} F_q^p(A),$$

where every considered $F_q^p(A)$ should correspond to a non-empty cluster. Then, with the specified number k and minimum length min_{len} of expected rules, GIH applies the TFP technique on T to mine the top-k frequent patterns as the dominant rules for guest invitation to activities of the corresponding type.

Table 3. Example of a factor-set collection *T*.

$\{f_2, f_3, f_4\}$, $\{f_5, f_6, f_7\}$, $\{\{f_1, f_2, f_4\}$, $\{f_2, f_4, f_7\}$
$\{f_2, f_4, f_5\}$, $\{f_3, f_5, f_6\}$, $\{\{f_5, f_6\}$, $\{f_1, f_5, f_6\}$

Example 3: Consider a factor-set collection T in Table 3. Given $k = 2$ and $min_{len} = 2$, by

[1] For the same reasons, the factors that incorporate more details from $\boldsymbol{W}$ and $\boldsymbol{H}$ (such as the values of $\boldsymbol{H}$ as the factor weights) will suffer and reduce their generality.

TFP, *GIH* then identifies the two most frequent rules of lengths at least 2:

$R = \{\{f_2, f_4\}, \{f_5, f_6\}\}$. ❑

Given historical data of different activity types, GIH can thus learn the corresponding dominant rules for recommending guests to activities of different types.

4.4 Guest Recommendation

In an online recommendation process, GIH consists of two major steps: (1) detecting the type of the newly input activity, and (2) ranking the friends of the activity host for guest recommendation.

In particular, GIH detects the type of the newly input activity by evaluating its similarity to all the clustered activities in each type. That is, according to the topic and description, GIH calculates the similarities between the newly input activity and every already-clustered activity by Eq. (1). The newly input activity d is then judged to belong to a type that has in average the highest similarities between its clustered activities and the newly input activity, *i.e.*,

$$Type(d) = \arg\max_{t} Average(Sim(d, d_t)),$$

where the type of activity d_t is t.[2]

After the detection of activity type, the friends of the activity host are ranked for guest recommendation according to the previously off-line extracted rules corresponding to $Type(d)$. Specifically, note that there usually exist multiple rules (such as 'a relative' or 'a college friend') for inviting different groups of guests to a specific activity. However, for a guest, s/he just needs one rule as the reasoning of being invited. That is, as long as there exists one rule recommending a friend to be the guest, the friend is worth considering for guest invitation. Furthermore, also note that a guest-invitation rule may consist of multiple factors (such as 'a college friend interested in classical music'). The matching degree of one's characteristics to a rule's factors indicates his/her recommended degree by the rule. Due to the concerns above, GIH hence ranks the host's friends for guest recommendation by finding everyone's best matching rule.

Example 4: Suppose that there are two guest-invitation rules $R = \{\{f_2, f_4\}, \{f_5, f_6\}\}$ for a specific type of activities. Consider a friend whose common characteristics with the host is $\{f_1, f_2, f_4, f_6\}$. GIH calculates the matching degrees of the friend's characteristics to the two rules, respectively, as

$$\frac{|\{f_1, f_2, f_4, f_6\} \cap \{f_2, f_4\}|}{|\{f_2, f_4\}|} = \frac{2}{2} = 1,$$
$$\frac{|\{f_1, f_2, f_4, f_6\} \cap \{f_5, f_6\}|}{|\{f_5, f_6\}|} = \frac{1}{2} = 0.5.$$

This friend is thus recommended for guest invitation by a ranking score 1. ❑

[2] A soft clustering of a newly input activity to multiple types is an alternative choice. However, such a method not necessarily leads to a better performance, but very probably worsens the recommendation instead due to the noises from less-relevant types of activities. For example, for a host who wants to invite friends to the performance of "The Marriage of Figaro", which type is musical hobbies rather than wedding. Using a soft clustering will then take the misleading rules from the type of wedding into account so that inappropriate guests will be invited for the opera performance.

5. PERFORMANCE EVALUATION

5.1 Data Sets and Experimental Setting

In this section, the experiments were conducted on the Facebook dataset to demonstrate the performance of the GIH algorithm. From these experimental results, a case study on rules used among different types of social activities was further presented to show the knowledge of the GIH algorithm. All programs are implemented in C++ and Matlab. The experiments are performed on an Intel Core i5 PC with 16GB RAM using Windows 8.

Data Sets The event data from the Facebook were crawled as the social activities for the evaluations. Table 4 lists the details of the collected 9 activities. For each activity, the characteristics of the host and all his/her friends were obtained from the user profiles, including 'about', 'likes', and 'events', of which the redundant words were cleaned using the Chinese Word Segmentation System provided by Academia Sinica. Here in order to ensure the diversity and generality of the data and for the interests of case studies, we artificially collected the 9 activities from three types of events: campus activities, wedding/funeral invitation, and hobbies. In applications, however, there is no need to specify the activity types in advances since the proposed GIH algorithm applies the clustering approaches to group activities based only on the topic words and descriptions.

Table 4. Summary of the event data from Facebook.

Activity	# of friends	# of invited guests	Topic
1	132	35	Wrap-up presentation of NTPU Pop Dance club
2	139	35	NTPU Christmas party
3	133	36	NTPU CS party
4	129	40	Sherry's memorial gathering
5	123	20	Welcome to Henry and Vivian's wedding
6	138	59	I'm getting married! Welcome to my wedding!
7	185	81	TCSB: an evening of animation music
8	195	93	Chess tournament of NTUST Cup
9	158	43	Brain Theatre: Riverside Drive

Table 5. Precision, recall, accuracy, and F1-measure among all approaches.

	Precision (SD)	Recall (SD)	Accuracy (SD)	F1-measure (SD)
Naïve	0.54 (±0.20)	0.54 (±0.16)	0.71 (±0.08)	0.53 (±0.18)
NMF	0.65 (±0.17)	0.67 (±0.17)	0.79 (±0.06)	0.65 (±0.15)
TFP	0.81 (±0.19)	0.89 (±0.09)	**0.92 (±0.07)**	0.86 (±0.12)
NMF+TFP	**0.87 (±0.17)**	**0.90 (±0.08)**	**0.92 (±0.07)**	**0.88 (±0.11)**

Parameter Settings For reliable evaluation, the leave-one-out cross validation is performed. The training activities are clustered using the complete-link algorithm with the number of activity types set according to the fact (*i.e.*, 3), and the weight setting in Eq. (1)

is $w_h = 2w_m = 4w_l$. The parameters k and min_{len} are both set as 2 when the TFP technique is applied. For testing, the number of guests recommended by GIH is decided by the number of host's friends in the testing activity and the average ratio of invited guests to host's friends in the training activities that belong to the same type as the testing activity.

Competitive Approaches For comparison, NMF and TFP were used alone in the GIH algorithm to utilize the host-guest relational matrix as two competitive approaches. Besides, a naïve technique that extracts the top 50% common characteristics as the dominant factors from the host-guest relational matrix in GIH was also implemented as another competitive approach.

5.2 Performance Comparison

5.2.1 Effectiveness

First, the precision, recall, accuracy, and F1-measure were used as the measurements, where the precision is the fraction of recommended friends that are invited, the recall is the fraction of invited friends that are recommended, the accuracy is the fraction of all friends that are predicted correctly, and the F1-measure is the harmonic mean of precision and recall. Table 5 lists the precision, recall, accuracy, and F1-measure among all approaches on all 9 social activities, where the best results are bolded.

From these results, it can be observed that among all approaches, using NMF+TFP in GIH is the best. This is because NMF first figures out the important relations for each cluster and TFP then extracts the frequent relations among them so that the dominant factors found by NMF+TFP are reasonable. Using TFP alone is better than using NMF alone, but both of them are worse than using NMF+TFP together, which shows these two techniques complement each other. At last, Naïve is the worst since it only considers the most common characteristics and they may not be the concerns behind the hosts to invite their guests.

Then, the ROC curve was used to evaluate the effectiveness of each approach, where the ROC curve plots the true positive rate against the false positive rate. Fig. 2 shows the ROC curves of all approaches, where the plots on each curve highlight the results when the number of recommended friends is 10, 20, …, and so on. First, as the larger area under the curve is the better, it is obvious that NMF+TFP is the best, followed by TFP and NMF, while Naïve is the worst. Second, the false positive rate of NMF+TFP raises quickly when the number of recommended friends exceeds 40 due to

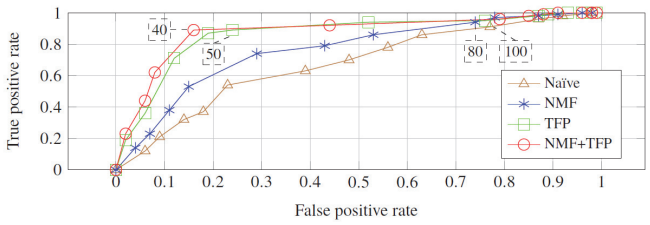


Fig. 2. The ROC curve.

the average invited guests of our Facebook data is only 41, which means using the NMF+TFP techniques is capable to recommend the suitable guests with a shorter recommendation list than the other approaches. For example, NMF+TFP has true positive rate 0.89 as the number of recommended friends is 40. To achieve the similar true positive rate, TFP, NMF, and Naïve need 50, 80, and 100, respectively, recommended friends.

5.2.2 Effects of parameters

Next, the effect of the parameters in the GIH algorithm was explored. Specifically, first the cluster weights (when GIH discriminates the types of activities), then the profiles used to construct the relational matrix, and finally the length of frequent patterns extracted by the TFP technique were discussed.

In order to show the effects of different cluster weights, here are three kinds of settings: $w_h = w_m = w_l$, $w_h = w_m = 2w_l$, $w_h = 2w_m = 4w_l$.[3] Besides, the results without clustering are also shown as the baseline. Fig. 3 shows the precision, recall, accuracy, and F1-measure of different cluster weights. First, for all four measurements, $w_h = 2w_m = 4w_l$ is the best, followed by $w_h = w_m = 2w_l$, while $w_h = w_m = w_l$ is the worst in three weight settings. Then, compared with the results without clustering, all results of $w_h = 2w_m = 4w_l$ and $w_h = w_m = 2w_l$ are better than that of without clustering, while the results of $w_h = w_m = w_l$ are similar to that of without clustering. Since $w_h = 2w_m = 4w_l$ weights topic, topic-related words, and the others in a descending order and $w_h = w_m = 2w_l$ puts higher weighting on topic and topic-related words, this indicates that preferring the topic words for clustering is a good idea, which complies with the common sense that people usually put enough information on the title of an activity in order to attract the attention of their guests.

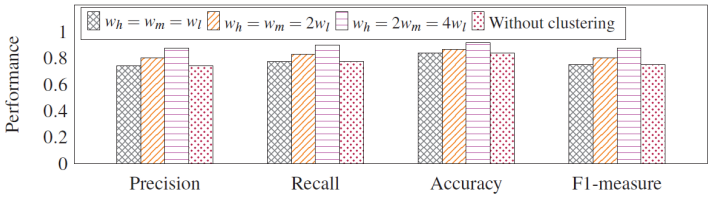


Fig. 3. The performance comparison of different cluster weight settings in terms of precision, recall, accuracy, and F1-measure, respectively.

Next, the performance of the profiles used to construct the relational matrix was compared. Fig. 4 shows the precision, recall, accuracy, and F1-measure of using 'about', 'likes', and 'events', separately. For all four measurements, using 'likes' is the best, followed by 'about', while 'events' is the worst. This complies with the common sense that people express various interests by clicking the 'likes' button so that using the information in 'likes' to construct the relational matrix is meaningful. By contrast, the information in 'about' and 'events' only expresses part of people's interests so that the performance is thus poor.

[3] Generally, the words weighted higher bring more effects on the similarity value. However, a wide disparity in w_h, w_m, and w_l overemphasizes some words while little difference of disparity may lead to similar results of the equal-weighting setting. In our dataset, we obtain $w_h = 2w_m = 4w_l$ as a good solution to the weight setting.

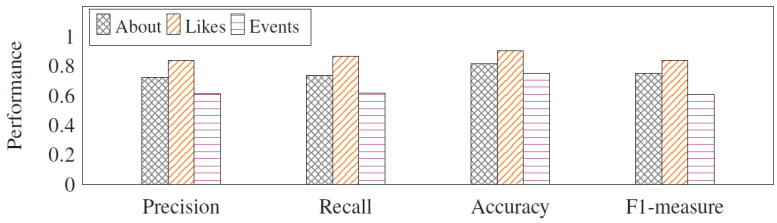


Fig. 4. The performance comparison of different profiles used in terms of precision, recall, accuracy, and F1-measure, respectively.

At last, the length of the frequent patterns identified in the TFP technique was discussed. Fig. 5 shows the precision, recall, accuracy, and F1-measure as the length is set as 1 to 5. From these results, it can be observed that setting the length of the frequent patterns as 2 is the best and using frequent patterns with longer length is meaningless. This makes sense since the concerns behind the hosts to invite their guests are not the more the better, and taking too much concern into account will result in noise.

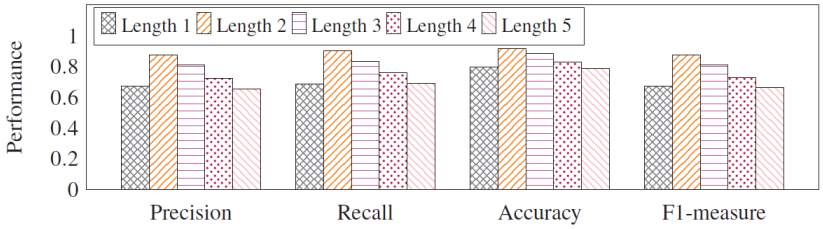
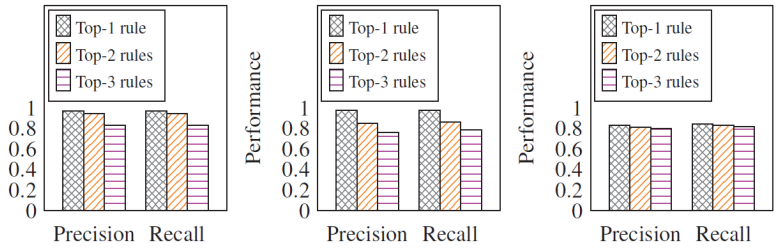


Fig. 5. The performance comparison of different length of the frequent patterns identified in terms of precision, recall, accuracy, and F1-measure, respectively.

5.3 Case Study and Discussion

In this subsection, it studied one activity in each type of social activities to show the knowledge of our GIH algorithm. Table 6 lists the top-5 rules of these three cases and Fig. 6 shows the precision and recall of the GIH algorithm when top-k rules are applied, for k from 1 to 3.



(a) Activity 3 (b) Activity 6 (c) Activity 7

Fig. 6. The performance comparison of using top-k rules in three cases in terms of precision, recall, accuracy, and F1-measure, respectively.

Table 6. Top-5 rules of the three cases.

Case	Clustered type	Top-5 rules of the cluster
Activity 3	Campus activities	1. University, National Taipei University 2. News, media 3. Public figure, public 4. News, artist 5. Actor, company
Activity 6	Wedding/funeral invitation	1. Pages, director 2. Actor, clothing 3. Musician, band 4. Travel, business 5. Service, Japan
Activity 7	Hobbies	1. Actor, director 2. Performance art, performance and event venue 3. Musician, band 4. TV show, performance 5. Orchestra, musician

The first case – Activity 3, "NTPU CS party", is a campus activity held by the computer science student in National Taipei University. By the human intuition, the concerns behind the invitation should be the same school and the same department. In Table 6, the first rule of campus activities exactly matches what the host thinks, which brings high precision and recall when only the top-1 rule is used, as shown in Fig. 6 (a). The second and the third rules help find the friends with similar interests. However, since not all interests are related to "NTPU CS party", the precision and recall slightly drop down when the second rule is applied, and drop down more when the third rule is further applied, as shown in Fig. 6 (a).

The second case – Activity 6, "I'm getting married! Welcome to my wedding!" is held by a bride who works in the entertainment circles. The first and second rules well represent the characteristics of her friends and colleagues so that the performance of using the top-2 rules is still great, as shown in Fig. 6 (b). The third rule "musician, band" is a little far away from the bride's interest. That is why using the top-3 rules results in obvious decline in performance, as shown in Fig. 6 (b).

The last case – Activity 7, "TCSB: an evening of animation music", is related to hobbies. In our thoughts, a host would prefer his/her guests with similar interests in the band when holding a concert. GIH retrieves the third rule that matches our thoughts. However, since having hobbies like performance and music is pervasive in our training data, the performance of using different top-k rules is thus comparable, as shown in Fig. 6 (c).

In addition, according to the results of these three cases, setting k as 1 is the best. This is because the information on the Facebook is not comprehensive so that GIH misses some key rules and takes irrelevant rules into account instead, resulting in poorer performance when k increases. For example, in the second case (Activity 6), by intuition, the guests usually include the relatives of the groom or bride. However, people seldom fill the information about their relatives on the Facebook. GIH thus cannot retrieve this kind of rules due to the limitation of missing data on the Facebook. If the information of

relatives is available, it is possible to achieve better performance with the top-2 rules in the second case. In other words, there could be chances for a larger k to take really meaningful rules into account when more comprehensive data are available, while in contrast, a small k would be better in order to avoid misleading rules.

In the experiments, the GIH algorithm with NMF+TFP has the best precision, recall, accuracy, and F1-measure, and can achieve a high true positive rate with a lower false positive rate, which means GIH outperforms other approaches with an even shorter recommendation list.

6. CONCLUSION

This paper addressed the need of guest recommendation to ease the task of selecting proper guests from a large number of friends for invitation to kinds of specific types of activities. For the problem, this paper proposed the Guest Invitation to Hosts (GIH) algorithm to learn the dominant factors behind the guest invitations from historical data. The GIH algorithm consists of two learning mechanisms: (1) the learning of discrimination between activity types, by clustering on the topic words and descriptions of activities, and (2) the learning of latent dominant factors for guest invitation (to specific types of activities), by the Non-negative Matrix Factorization and Top-k Frequent Pattern mining techniques. Evaluated on the Facebook data, GIH can reach an accuracy of at least 80% in the guest recommendation, and outperform naïve approaches in terms of precision, recall, F1 score and accuracy. The case studies further showed that the dominant factors (rules) identified by GIH comply with the human intuition.

REFERENCES

1. X. Cai, M. Bain, A. Krzywicki, W. Wobcke, Y. S. Kim, P. Compton, and A. Mahidadia, "Collaborative filtering for people to people recommendation in social networks," in *Proceedings of Australasian Conference on Artificial Intelligence*, 2010, pp. 476-485.
2. J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, "Make new friends, but keep the old: recommending people on social networking sites," in *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems*, 2009, pp. 201-210.
3. A. D'cunha and V. Patil, "Friend recommendation techniques in social network," in *Proceedings of IEEE International Conference on Communication*, *Information and Computing Technology*, 2015, pp. 1-4.
4. C.-Y. Shen, D.-N. Yang, L.-H. Huang, W.-C. Lee, and M.-S. Chen, "Socio-spatial group queries for impromptu activity planning," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, 2016, pp. 196-210.
5. C.-Y. Shen, D.-N. Yang, W.-C. Lee, and M.-S. Chen, "Maximizing friend-making likelihood for social activity organization," in *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2015, pp. 3-15.
6. H.-H. Shuai, D.-N. Yang, P. S. Yu, and M.-S. Chen, "Scale-adaptive group optimization for social activity planning," in *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2015, pp. 45-57.

7. H.-H. Shuai, D.-N. Yang, P. S. Yu, and M.-S. Chen, "Willingness optimization for social group activity," *Proceedings of the VLDB Endowment*, Vol. 7, 2013, pp. 253-264.
8. S. Wan, Y. Lan, J. Guo, C. Fan, and X. Cheng, "Informational friend recommendation in social media," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2013, pp. 1045-1048.
9. D.-N. Yang, C.-Y. Shen, W.-C. Lee, and M.-S. Chen, "On socio-spatial group query for location-based social networks," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 949-957.
10. M. Madden, A. Lenhart, S. Cortesi, U. Gasser, M. Duggan, A. Smith, and M. Beaton, "Teens, social media, and privacy," *Pew Research Center*, Vol. 21, 2013, pp. 2-86.
11. D.-N. Yang, Y.-L. Chen, W.-C. Lee, and M.-S. Chen, "On social-temporal group query with acquaintance constraint," in *Proceedings of the VLDB Endowment*, Vol. 4, 2011, pp. 397-408.
12. M. Deutsch and H. B. Gerard, "A study of normative and informational social influences upon individual judgment," *The Journal of Abnormal and Social Psychology*, Vol. 51, 1955, p. 629.
13. M. F. Kaplan and C. E. Miller, "Group decision making and normative versus informational influence: Effects of type of issue and assigned decision rule," *Journal of Personality and Social Psychology*, Vol. 53, 1987, p. 306.
14. D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, Vol. 401, 1999, p. 788.
15. W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 267-273.
16. B. Goethals, "Survey on frequent pattern mining," *University of Helsinki*, Vol. 19, 2003, pp. 840-852.
17. J. Han, J. Wang, Y. Lu, and P. Tzvetkov, "Mining top-*k* frequent closed patterns without minimum support," in *Proceedings of IEEE International Conference on Data Mining*, 2002, pp. 211-218.
18. J. Wang, J. Han, Y. Lu, and P. Tzvetkov, "Tfp: An efficient algorithm for mining top-k frequent closed itemsets," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, 2005, pp. 652-663.
19. J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier, UK, 2011.

Chih-Hua Tai (戴志華) received the Ph.D. degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan. She is currently an Assistant Professor in the Department of Computer Science and Information Engineering, National Taipei University, New Taipei, Taiwan. Her research interests include privacy-preserving data sharing and mining, healthcare data mining, and social computing and marketing.