Towards Social Recommendation System Based on The Data From Microblogs

Pei-Shan Chang  
Department of Information Management  
National University of Kaohsiung  
Kaohsiung City, Taiwan  
purplemio@gmail.com

I-Hsien Ting  
Department of Information Management  
National University of Kaohsiung  
Kaohsiung, Taiwan  
iting@nuk.edu.tw

Shyue-Liang Wang  
Department of Information Management  
National University of Kaohsiung  
Kaohsiung City, Taiwan  
slwang@nuk.edu.tw

Abstract—With the rapid growth of Internet and social networking websites, there are various services that provided in these platforms. For instance, Facebook focuses on social activities, Twitter and Plurk are both focus on the interaction of users through short messages (which are so-called microblogs). Therefore, there are more than millions of users registered in these websites and become places where full of marketing possibilities. Thus, it is an important issue to assist companies to understand the users in the social networking websites in order to enhance the accuracy and efficiency of target marketing.

In this paper, we have proposed the architecture of a social recommendation system based on the data from microblogs. The social recommendation system is conducted according to the messages and social structure of target users. The similarity of the discovered features of users and products will then be calculated as the essence of the recommendation engine. A case study will be included to present how the recommendation system works based on real data that collected from Plurk.

Keywords: Social Recommendation System, Social Networks Analysis, Microblogs, Target Marketing

I. INTRODUCTION

With the rapid development of Internet technology and the combination with the concept of Web2.0, the web has becoming a popular communication platform. Microblogging is one of the recent social phenomena of Web 2.0. Comparing with other web 2.0 services, it focuses on communication and immediately interaction, such as Twitter, Jaiku, Digu and Plurk. Microblogging allows users to post short message to describe, update and share their current status and opinion [10]. The users can post not only text message but also images, videos or links. They can also access the microblogs via many platforms, such as mobile phone, instant messenger , plug-in, and synchronize with Facebook.

Thus, there are more and more applications have been applied in microblogs, such as marketing. For example, Dell uses Twitter to sell off-season products and have gained about two million dollars in two years. GozCafe uses Karma value for product promotion, and it attracted more than 1500 customers.

Plurk is one of the most famous microblogs and provides faster communication between users and reduce their time in interaction, which is the biggest difference from traditional blogs. Besides, Plurk are close to users’ daily life as that topics are mostly related to their daily lives, current events, news, and other interests. Therefore, it increases the usage of users. In figure 1, it shows that Plurk has become a most popular platform with higher using frequency than other microblogs, such as blogspot.com and twitter.com.

![Figure 1. The comparison of using frequency of plurk, blogspot and twitter (From Alexa.com ,2011/03/12)](image)

In order to analyze the data in microblogs, social networks analysis (SNA) is an important research field due to it focuses on the analysis of social data and social relation. Social network is used to describe the relationship between people and also display what the role of an actor plays in a group. In particularly, there are more and more users and messages coming along with the communication which make large and complex social network. It also increases the difficulties to understand the structure of a social network and to analyze the social data. De Valck has pointed-out that it is important for marketers to face this challenge and try to extract useful knowledge for the users in those online social communities [4].

Therefore, we can use SNA to extract the features of users in online social communities’ user features and combine with marketing theories for meaningful to mapping the user feature patterns discovering. This can be used to conduct a recommendation system, not only can help marketers to catch the target customers more detail, but also can be used the mapping patterns to find target customer and develop marketing strategies that can increase the marketing efficiency and to develop appropriate marketing strategies.

In this paper, we will apply the methodologies of SNA understand more about the features of microblogs. The content and social structure of microblogs will be used for discovering the patterns. The discovered patterns will then be used to
conduct a recommendation system for business product recommendation.

The structure of this paper is organized as below: in sections 2, some related literatures of microblogs, social network analysis and recommendation system will be reviewed. In section 3, the system architecture of the proposed recommendation system will be introduced. In section 4, we will show a case study to demonstrate the proposed methodology. The paper will be concluded as well as the discussion will be provided in section 5.

II. LITERATURE REVIEW AND RELATED WORKS

In this section, related literatures will be reviewed, including microblog and marketing field, social networks analysis and recommendation system based on SNA.

A. Microblog and Marketing

Microblog is currently one of the most popular blog platforms. It allows only 140-character text messages input and users can post their daily life. It is also a very good platform for the users to maintain the friendship between users [26]. Another difference between microblog and traditional blog is the updating frequency [10]. The users usually update their status more than once a day in microblogs but usually update once in few days in traditional blogs. In Taiwan, Plurk is one of the most popular microblog. There are two different kinds of role, one is sharing friendship with other users and another is “follower” which means the user want to pay special attention to the users in the list.

Due to there are more and more users on microblogs, many companies run different promotion activities on it. Such as in the period of US presidential election, Barack Obama used Twitter as a platform to attract the focuses of young people. In addition, Dell also uses Twitter to be their marketing place for promotion. In Taiwan, GozCafe is the first case that use microblog for marketing. They use Plurk as a virtual channel to implement Click-and-Mortar strategy.

About microblog marketing, viral marketing was the most adopted strategy. However, target marketing is another good option if the companies can understand the users’ characteristics.

Normally, there are three steps in target marketing which are Segmentation, Targeting, and Position. In the stage of segmentation, a market can be separated to many groups according to different users’ behaviors, or distinguishable features. McCarthy pointed out that segmenting is dividing the market into several tiny homogeneous markets, in order to suit all possible needs and also provide benefits for firms to develop marketing strategies [15]. Following this, the microblog message and the structure of social networks will be take place in this paper to support target marketing.

B. Social Network Analysis

SNA is an important domain to analysis social network’s structures and relationships. We can understand the characteristics of a social network through SNA [22]. In SNA, there are some common methods such as density, closeness, centrality, and betweenness. SNA can also be used to discover a role or an actor, as detailed in figure 2, such as the role of star, social, and bridge. Besides, there are some advance methods, for example, cluster, it can be used to cluster similar groups. The analysis result can then be used in other fields meaningfully, such as marketing.

The research methodology of SNA has been developed wildly in many domains such as sociology, management, commerce, biology and computer science [11][25]. In these research fields, not only SNA can be used to deal with large amount of social data more efficiency [5][7][16] but also other technologies including HITS (Hypertext Induced Topics Selection), Semantic Web, PageRank…etc. The most important measurements of SNA include diameter, centrality, degree centrality, closeness centrality and betweenness centrality.

The diameter is used to measure the amount of nodes between two nodes in a network. Centrality includes three degree measurement, Degree Centrality, Closeness Centrality and Betweenness Centrality. Degree centrality is to measure which actors have most relation or link with others in a network. Closeness centrality is used to measure how close an actor is to all the other actors in a network, in other words, the actor can quickly interact with all others. Betweenness centrality is that an actor may have some control over some of the paths. These measurements are common used in many social network related researches and will be used in this paper as well.

In this paper, we will propose a method to analyze the microblog data and to extract social network. There are some researches about how to exact social network from unstructured web-page data, such as Matsuo et al. made a POLYPHONE system by using searching techniques to conducting social networks with correlation and co-occurrence [14]. Currently, in the field of computer science, related researches mostly use Data Mining and Web Mining as the technique for social network extraction. The web mining techniques can be divided into three types, which are Web Content Mining, Web Usage Mining and Web Structure Mining [24]. In this research, we are focusing on online social networks analysis. Text mining and web content mining are the main technologies which used in this area. These techniques are also can support to understand the preferences of users. Web usage mining can be used to translate the web usage into social networks relation data [13] Web structure mining is a technique that can be used to analyze the network to find the paths, reachability or structure hole.
C. Recommendation System

There are typically two different types of recommendation systems - collaborative filtering and content-based methods. The first method is according to someone who have the same taste or opinion of the recommended items with the target user (is a recommendation method of using similar taste with target user) [2][3][19], the second one is a recommendation method base on the item attributes to measure the preference of target user [17][18][20].

However, there are still many issues should be discussed about recommender systems. First, content-based methods have to rely on explicit item descriptions to measure item similarity, therefore, it is difficult to extract such descriptions from user’s ideas or opinions [9]. Collaborative filtering has cold-start and data sparsity problem [1]. The cold-start problem is that when a user initially joins, this user may has only a few reviews (or none) on this system. Therefore, it’s hard to obtain user’s preference to make recommendations from past reviews. The data sparsity problem usually happens in recommender systems with great number of items. Sometimes, user usually only rates or purchases few number of items. Consequently, it is difficult to measure the similarity of users for recommendation based on limited number of reviews.

Previously, traditional techniques that used in recommendation system only utilize the characteristics of users for recommendations but ignore the influence of relationships in social networks. For example, the purchase activities of users will affected by their friends [8]. On the other hand, Golbeck pointed out that a recommendation can be made if there is just one path exist in a social network [6]. In particularly, the efficiency and accuracy can be increased and combined not only utilize the preference of users, but also combine the structure of social networks.

Recently, there are some researchers are related to social network-based recommendation systems, such as Gupta et al. pointed that friend’s in the network will influence user’s decision [8]. Golbeck using the weight of trust to make recommendation for users [6]. Seth and Zhang proposed a method by using strength-of-weak-ties to bring diverse recommendations to target user [23]. Scott et al. focus on people who are in the second-degree network of the target user to find who are sharing interests or have a relation of reciprocity [21]. However, most those researches focus on the data from Facebook.com, the user profiles in Orkut. There are few researches focus on the message content, and structure in microblog’s. Therefore, in this paper, Plurk (which is the microblog with highest usage in Taiwan) will be utilized as the platform to implement the proposed recommendation system.

III. The System Architecture

According to the research background and motivation, we have designed the system architecture to address the related issues. The proposed system architecture is presented in figure 3. In this paper, target marketing is the marketing theory that used in the system, instead of other marketing theories, such as Vital Marketing for unspecific target, Word-of-Mouth Marketing based on trust.

The purpose of the system is to understand more about the characteristic of the data in microblogs and to study how to extract social networks. The source message and social data will be pre-processed and keep the meaningful keywords. In another part, the target marketing theory will be used to pre-define the categories of products and target users, in order to conducting the relationship between product features and social network structure of users.

In figure 3, the system can be divided into three parts, the first part is the introduction of plurk message and data pre-processing. The second part is to conduct the characteristic vector of users and products, and the third part is the recommendation mechanism that used in the recommendation system. The process of the three parts will be introduced in detail below.

A. Introduction to the Plurk message and SNA

1) Data Collection

The message of microblog is very short, simple, update quickly, and conversational, totally different from traditional blogs. For the features of short and conversational message, the data could be more close to user’s real preference.

2) Database
After data collection and extraction, the output will then be stored in a database. The database is designed according to the structure of the message in microblog. There are two fields in the table which are PlurkContent and PlurkUser. The fields are used to record user’s profiles and messages in Plurk. Furthermore, the responses of the messages and the list of responders will be stored in the field of the table ResponseContent and ResponseUser.

3) Keyword Extraction

Due to the message data from microblog are un-structured data, thus the keywords in the message must be extracted first by the process of Naturally Language Process (NLP). The stop-words in the message must be removed and then the keywords can be extracted by measuring the keyword frequency. In NLP, the measurement of TF and TF-IDF are basic methods which can be used to measure the keyword frequency. TF means the term frequency in a document. TF-IDF represents the importance of words among all document. Those measurements are defined as below:

\[
\begin{align*}
    TF &= \frac{freq(i,j)}{\maxfreq(l,j)} \quad \text{(1)} \\
    IDF &= \log\left(\frac{N}{ni}\right) \quad \text{(2)} \\
    TF-IDF &= TF \times IDF \quad \text{(3)}
\end{align*}
\]

4) SNA Modules

In this paper, the adoption of SNA methods is according to the features and network structure of Plurk. In order to extract the important characteristic from Plurk, HITS(Hypertext induced topic selection) can be used to define the role of Hub or Authority for a node. Hub can be the information provider and Authority can be the node with highest prestige in a network. \(A(u)\) represents the times of other users \(v\) make responses to target user \(u\) and \(H(u)\) is the times of target user \(u\) makes responses to other users \(f\) [12].

\[
\begin{align*}
    Authority(u) &= \sum_{v \in \mathcal{U}, u \rightarrow v} H(v) \quad \text{(4)} \\
    Hub(u) &= \sum_{u \in \mathcal{U}, u \rightarrow f} Authority(f) \quad \text{(5)}
\end{align*}
\]

B. The Definition of Target Marketing

In this paper, the main purpose is to conduct the user characteristic vector of different products which is based on target marketing theory. As shown in figure 4, it can be divided into two different parts. In the figure, user and product will be treated as different target in marketing.

For example, different users may have different interests. Whose favor is music and his message will full of many related words, such as linkin park or Mariah Carey, or have some links such as the links to YouTube. Therefore, for this example, the user will be categorized as the type of “MUSIC”. In particular, the definition of product characteristics would be conducted from the social data of users. Preliminary, a questionnaire will be used to collect the interests of users. The collected data will then be used as the training data set in our experiment.

According to user’s historical message, we can acquire the social network features to define the relationship within the network structure and content of products. Finally, we will get the relation mapping table of users and products.

![Figure 4 The relation mapping of users and products](image)

A case study will be included in next section to present how the recommendation system works based on real data that collected from Plurk.

IV. A CASE STUDY

A. The pre-processed Data

The pre-processed raw data will then be stored in the database, and the structure of the database is shown in table 1 and table 2. Table 1 is the database structure and table 2 is the plurk message database structure.

Table 1. The database structure

<table>
<thead>
<tr>
<th>RowName</th>
<th>Meaning</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>User unique id</td>
<td>3182573</td>
</tr>
<tr>
<td>nick_name</td>
<td>User account</td>
<td>mynameishi</td>
</tr>
<tr>
<td>karma</td>
<td>A kind value of usage and popular</td>
<td>93.23</td>
</tr>
<tr>
<td>gender</td>
<td>0=female , 1=male</td>
<td>0</td>
</tr>
<tr>
<td>location</td>
<td>User location</td>
<td>Kaohsiung, Taiwan</td>
</tr>
</tbody>
</table>
Table 2. The plurk message database structure

<table>
<thead>
<tr>
<th>RowName</th>
<th>Meaning</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>userID</td>
<td>Target user</td>
<td>3182573</td>
</tr>
<tr>
<td>ownerID</td>
<td>Message author</td>
<td>3182573</td>
</tr>
<tr>
<td>qualifier</td>
<td>The type of post (is, says, asks,...)</td>
<td>says</td>
</tr>
<tr>
<td>content_raw</td>
<td>Message content</td>
<td>Good night</td>
</tr>
<tr>
<td>plurkID</td>
<td>This message’s unique id</td>
<td>589257961</td>
</tr>
<tr>
<td>responsesCount</td>
<td>Response count</td>
<td>3</td>
</tr>
<tr>
<td>favorCount</td>
<td>The number like this message</td>
<td>1</td>
</tr>
<tr>
<td>favorers</td>
<td>The list who like this message</td>
<td>3408049</td>
</tr>
</tbody>
</table>

B. Keywords and The Social Structure

We choosing the keywords related to Food (such as the name of a restaurant), 3C (such as ipad), and a dataset has been extracted accordingly. Then, UNINET is used to generate the social graph and calculate the SNA measurements to help the understanding of unknown features within products and users' social network structure.

In figure 5, the clusters in the target network A and B can be defined very easy through the visualized social network graph. (a) in figure 5 is the network graph based on the message type of food of target A and (b) is the network graph of target B. (a) in Figure 6 is the network graph based on the message type of 3C of target A and (b) is the network graph of target B. As shown in the graphs, the degree measurement in 3C network is greater than the network of Food.

The detail SNA measurements are shown in table 3. In table 3, we can find that in type Food, target A and B’s SNA value are very close. As the same phenomenon, in type 3C the value also closes with each other. For type 3C, the measurement of density and closeness all pointed that actors are close to each other and more close than type Food. Also in Avg Dist, in type 3C, the average geodesic path length is shorter than type Food. The analysis results indicate that users in type 3C are more condense. In another hand, the higher density means the more homogeneity of users in this product type.

Figure 6. Target B in type 3C (a) 23 actors (b) 4 messages

From table 3, we also can found there is a big gap of those SNA measurement values between type Food and 3C. In another word, the SNA measurement can be used to define the characteristics vector of different products based on social network features. For the analysis results, it indicates that the proposed method is reasonable and can be discussed further in the future.

Table 3. The value of SNA measurements of target A and B with type Food and 3C

<table>
<thead>
<tr>
<th>Type \ User</th>
<th>Food Target A</th>
<th>Food Target B</th>
<th>3C Target B with 23 actors</th>
<th>3C Target B with 4 messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree</td>
<td>31.225</td>
<td>34.783</td>
<td>51.799</td>
<td>48.000</td>
</tr>
<tr>
<td>Closeness</td>
<td>60.260</td>
<td>61.475</td>
<td>68.547</td>
<td>66.937</td>
</tr>
<tr>
<td>Betweenness</td>
<td>3.275</td>
<td>3.106</td>
<td>2.296</td>
<td>2.167</td>
</tr>
<tr>
<td>Density</td>
<td>0.631</td>
<td>0.643</td>
<td>0.726</td>
<td>0.714</td>
</tr>
<tr>
<td>Avg Dist</td>
<td>1.866</td>
<td>1.854</td>
<td>1.769</td>
<td>1.781</td>
</tr>
<tr>
<td>Norm Dist</td>
<td>0.798</td>
<td>0.803</td>
<td>0.842</td>
<td>0.837</td>
</tr>
</tbody>
</table>
Avg Dist is the average geodesic path length in the bipartite graph within components.

Norm Dist is Avg Dist divided into minimum possible in bipartite graph of given node-set sizes.

V. CONCLUSION AND FUTURE RESEARCH

In this paper, we have proposed the architecture of a social recommendation system based on the data from microblogs. The social recommendation system is conducted according to the messages and social structure of target users. The similarity of the discovered features of users and products will then be calculated as the essence of the recommendation engine. A case study will be included to present how the recommendation system works based on real data that collected from Plurk. From the analysis results, we can find the difference of the SNA measurement between different products is significant. Therefore, it shows that the recommendation system is workable to recommend different products to target customers.

In the future, we will try to implement the proposed recommendation system based on our propose methods by using the characteristics of social networks.

REFERENCES


