Journal of Universal Computer Science, vol. 18, no. 4 (2012), 554-576
submitted: 28/9/11, accepted: 14/12/11, appeared: 28/2/12 © J.UCS

Understanding Microblog Users for Social Recommendation Based on Social Networks Analysis

I-Hsing Ting, Pei Shan Chang, Shyue-Liang Wang
(Department of Information Management
National University of Kaohsiung, Taiwan
iting@nuk.edu.tw, M0993306@mail.nuk.edu.tw, slwang@nuk.edu.tw)

Abstract: With the rapid growth of Internet and social networking websites, various services are provided in these platforms. For instance, Facebook focuses on social activities, Twitter and Plurk (which are called microblogs) are both focusing on the interaction of users through short messages. Millions of users enjoy services from these websites which are full of marketing possibilities. Understanding the users can assist companies to enhance the accuracy and efficiency of the target market. In this paper, a social recommendation system based on the data from microblogs is proposed. This social recommendation system is built 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 included in the paper presents how the recommendation system works based on real data from Plurk.

Keywords: Social Networks Analysis, Social Recommendation System, Microblogs, Target Marketing
Category: H.3.5

1 Introduction

With the rapid development of Internet technology and its combination with the concept of Web 2.0, the web has become a popular communication platform. Microblogging is one of the recent social phenomena of Web 2.0. Comparing with the other web 2.0 services, Microblogging, such as Twitter, Jaiku, Digu and Plurk, focuses on the communication and immediate interaction. It allows users to post short messages sharing their statuses and opinions [Java et al. 2007]. Users can post not only text messages text but also images, videos or links. Microblogging is accessible through many platforms and devices, such as mobile phone, instant messenger, and many other plug-ins which can help synchronizing messages to Facebook. Thus, more and more applications have been applied by marketers in microblogs. For example, Dell used Twitter to sell off-season products ended up selling about two million dollars in two years. GozCafe used Karma for product promotion to attract more than 1500 customers.

Plurk is one of the most famous microblogs which provides a faster way to communicate between users and reduces the interaction time in comparison with traditional blogs. In addition, topics in Plurk are close to users' daily life. Therefore, the usage of users is increased. Figure 1 shows that Plurk becomes the most popular platform with the highest usage than the other microblogs, such as blogspot.com and twitter.com.

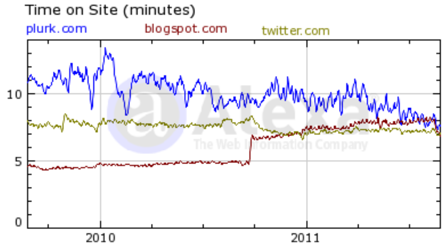


Figure 1: The average time spent on different microblogs (From: Alexa.com, 2011/08/28)

Social networks analysis (SNA) is an important research field to analyze the data in microblogs due to its focus on the analysis of social data and social relations. In particularly, there are more and more users and messages coming along with the communication which make large and complex social network. This increases the complexities of the social network which has made itself difficult to understand its structure and analyze the social data. For example, the huge number of interaction generates complicated social relationships [Yamaguchi et al. 2011]. A huge dataset can cost a great deal of time to process, (such as removing and replacing stop words). In addition, although a semantic method is indispensible to recognize the meaning in the context. [Hannon et al. 2010] suggests that the tweets of user's followees are not a good predictor. De Valck also points 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 [De Valck et al. 2009].

Therefore, we incline to use SNA to extract the users' features in online social communities. Social networks describe the relationships or interactions between actors (individuals) or events among the groups. It is helpful for researchers to find the unknown structures or characteristics of users and groups. It can also help to discover heterogenic structures that exist in different product communities [Wang et al. 2011] as well as different roles of actors that represent different abilities, for example, the diffusion ability [Ho et al. 2011].

A recommendation system is developed by combining users' features through SNA and marketing theories. It helps marketers to catch the target customers; moreover, the mapping patterns can be used to find target customers and to develop marketing strategies

In this paper, the methodologies of SNA are applied to understand more features of microblogs. The content and social structure in microblogs will be used to discover the patterns. Discovered patterns will then be used to develop a recommendation system for business product recommendation.

The structure of this paper is organized as below: in section 2, we review related literatures of microblogs, social network analysis and recommendation system. In section 3, we describe the system architecture of the proposed recommendation system. In section 4, a case study will be applied to demonstrate the proposed methodology. Conclusion and discussion will be provided in section 5.

2 Literature Review and Related Works

2.1 Microblogs and Marketing

Microblogs is currently one of the most popular social networking platform~~s~~. It allows users to type in no more than 140characters per text message. It is also a very good platform for the users to maintain their friendship [Zhao and Rosson 2009]. Another difference between microblogs and traditional blog is the updating frequency [Java et al. 2007]. In microblogs, users usually update their status more than once a day but only update once in a few days in traditional blogs.

Many companies are starting to run different promotional activities on microblogs due to its increased users. Besides, the speed of information exchange~~s~~ in microblogs is very fast. During the US presidential election, Barack Obama used Twitter as a platform to draw young people's attentions and to converge the political views in his constituencies. In addition, Dell also used Twitter as a marketing place to announce its promotions, discount and new arrival messages. It continues to provide great price of goods and first-hand e-coupons. [Hsu et al., 2010]. In Taiwan, GozCafe is the first case that uses microblogs for marketing. They hold an activity called "What is karma? Is it edible?" This activity allows Plurk users to pay for their meals once a day by Karma points which can be increased if posting meaningful messages or acquiring responses. For example, if your karma value is 50.03, then you can get a price discount of 50 dollars. At the same time, you also have to post a message on Plurk that you had used karma to get a discount in GozCafe. As a result, this promotion message spread out through Plurk. GozCafe uses Plurk as a virtual channel to implement Click-and-Mortar strategy and also conducted a great success.

As for microblogs marketing, viral marketing is one of the most adopted strategy. However, target marketing is another good option if the companies understand the users' characteristics. Normally, there are three steps in target marketing, they are segmentation, targeting, and position. In the stage of segmentation, a market can be separated into many groups according to users' different behaviours, or distinguishable features. McCarthy and Perreault point 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 [McCarthy and Perreault 1990]. With this in mind, the microblog messages and the structure of social networks will be taken place in this paper to support target marketing.

2.2 Social Networks Analysis

SNA is an important methodology which can be used to analyze the structures and relationships in social networks [Scott 2000]. In SNA, there are some commonly used measurements such as density, closeness, centrality, and betweenness. SNA can also be used to discover a role or an actor, (such as the role of star, social), and bridge (see figure 2).These roles can be identified by different SNA measurements. For example, the role of *social* may have higher out-degree (outward arrows). In contrast, *star* have higher in-degree (inward arrows). And *bridge* can connect two different groups in a social network. In other words, if there is no bridge in a network, these two groups will never be connected. Besides, there are some advanced analyzing methods in SNA, for example, cluster can be used to discover groups according to their similarities.

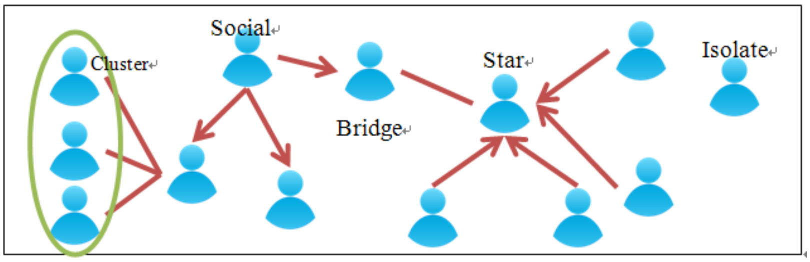


Figure 2: A simple social networks

SNA has been developed widely in many domains such as sociology, management, commerce, biology and computer science [Wilson 1989, Jun et al 2006]. In these research fields, not only SNA can be used to efficiently deal with not only large amount of social data [Mika 2005, Godbole et al. 2007, Goodreau 2007] but also other information technologies including HITS (Hypertext Induced Topics Selection), Semantic Web, PageRank…,etc. The most important measurements of SNA include *diameter and centrality*. 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 a measurement that can be used to measure which actor has most relationships or links with others. In other words, this actor is the most active one and can access the most resources in a social network. In a graph *G(V,E)*, there are a set of vertexes V and a set of edges E, the maximum numbers of actors is *g*. In graph *G*, the *degree centrality* $C_D(n_i)$ is an actor n_i's centrality index, defined in the following formulation, and the suffix *D* is representing as degree [Scott 2000]:

$$C_D(n_i) = d(n_i) = \sum_{\mathrm{j}} x_{ij} = \sum_{\mathrm{i}} x_{ij} \tag{1}$$

Closeness centrality is used to measure how close an actor is to all others in a network. In other words, the actor can have a quick interact with all others. Therefore, he may spread information and influence other actors in a quicker and easier way. The closeness centrality $C_c(n_i)$ is defined as the following formula and the suffix *C* represents the closeness. $d(n_i,n_j)$ is the number of lines in geodesic that links actors *i* and *j*. General speaking, the index is the inverse of the sum of the distance from actor *i* to all the other actors [Scott 2000]:

$$C_c(n_i) = \left[\sum_{j=1}^{g} d(n_i, n_j)\right]^{-1} \tag{2}$$

Betweenness centrality reflects the extent to which an actor lies between two nonadjacent ones. In other words, this actor is in a critical position that can control the information diffusion in the networks. The *betweenness centrality* $C_B(n_i)$ is defined as the following formula and the suffix *B* represents betweenness. The g_{jk} is the number of the shortest path of two actors, and $g_{jk}(n_i)$ is the number of the shortest path of two actors contain actor *I* [Scott 2000]:

$$C_B(n_i) = \sum_{\mathrm{j}<k} \left(\frac{\mathrm{g_{jk(n_i)}}}{\mathrm{g_{jk}}}\right) \tag{3}$$

These measurements are commonly used in many social network related research fields and will be used in this paper as well.

In this paper, we propose a method to analyze the data in microblog and to extract social networks. There are some researches about how to extract social network from unstructured web-page data. For example, Matsuo et al. made a POLYPHONET system by using searching techniques to conduct social networks with correlation and co-occurrence [Matsuo et al. 2006]. Up to date, Data Mining and Web Mining are frequently used as the techniques for social networks extraction in the field of computer science. The web mining techniques can be divided into three types: Web Content Mining, Web Usage Mining and Web Structure Mining [Ting 2008]. In this research, we focus on online social networks analysis. Text mining and web content mining are the main technologies used in this area. These techniques also can be used to understand the preferences of users. In SNA related researches, Web usage mining can be used to translate the web usage into social networks relation data [Lento et al. 2006]. Web structure mining is a technique to analyze the network to find the paths, reachability or structure hole.

2.3 SNA in Microblogs

Due to the huge amount of information on microblogs, the requirement of filtering-out useful information is growing [Yamaguchi et al. 2011]. Lee et al.

proposed several text analysis methods in order to detect emerging topics more effectively in Twitter. They use spatio-temporal data, such as latitude and longitude, time zone and content to enhance the data in microblogs. Making the data a reliable source [Lee et al. 2011, Jung 2009].

In addition, in microblogs, it is important not only to get useful information, but also to understand users' interests. Yamaguchi et al. use a name list which is tagged by users themselves. Users make appropriate tag names for other users in the list. That is to say the users in the list are most related to the characteristic of the list. For example, author A made tag "weather" in his list A1, it means the topics that posted by the users in list A1 are most related to weather. Therefore, if a user is described by the same tag from other users, the interests of the users can then be identified [Yamaguchi et al. 2011].

Some researchers study the issue of propagation in microblogs. Ho et al. studied how to measure and visualize the information propagation in Plurk. If companies can understand a user's influence, it can help companies to promote its products to the right person to promote. Their methodologies quantify a person's ability of propagation, measure and visualize the propagation by defining rigid and loose relationships in the messages. The rigid relationship is used to measure the reply and repost behaviour by the friends of the users. The loose relationship, on the other hand, only performs reply action. A system was implemented as a search platform for users to see the propagations and relevant information [Ho et al. 2011].

Furthermore, many researchers are using social relation to analyze the data in microblogs. It is a challenge to locate the most interesting and authoritative author, who can quickly and easily draw the users' attentions. Pal and Counts try to identify such author in Twitters. They define a lot of quantitative parameters, such as the number of retweets or number of used keyword hash tags [Pal and Counts 2011].

Ediger et al. proposed a graph characterization toolkit to analyze the massive relations in social networks. In particular, they use '@' tag and unique Twitter interactions to reduce the size of the networks. They also intend to identify influential sources by analyzing the interesting characteristics of users. The characteristics include degree distributions, connected components, betweenness centrality [Ediger et al. 2010].

According to the literatures above, we find out most researches were focusing on how to identify the most influential users in mircoblogs. However, there are few researches devoting on the area of recommendation system with the combination of SNA and the analysis of messages in microblogs. Recently, Hannon et al. proposed a recommendation method by combining content analysis with the social relation. They find out that social relation follower is more appropriate to be a recommendation resource. Moreover, they point out that micro-blogging service, such as Twitter, can be used as a useful recommendation resource [Hannon et al. 2010]. However, they are focusing on how to recommend follower to users. In contrast, our method is focusing how to recommend target users for companies.

2.4 A Review of Recommendation Systems

Typically, there are two different types of recommendation systems: collaborative filtering and content-based methods. The first method is according to someone who has the same taste or opinion of the recommended items with the target user [Resnick et al. 1994, Billsus and Pazzani 1998, Breese et al. 1998], the second method is a recommendation method based on the item attributes and measurement of the similarity between the attributes and the preference of target user [Pazzani and Billsus 1997, Mooney and Roy 1999, Sarwar et al. 2001].

However, there are still many issues needed to be discussed concerning recommendation systems. First, content-based methods have to be relied on explicit item descriptions to measure the similarity of items. That is to say, it is difficult to extract such descriptions from users' ideas or opinions [He and Chu 2010]. Collaborative filtering has *cold-start* and *data sparsity* problem [Adomavicius and Tuzhilin 2005]. The cold-start problem is that when a user initially joins, this user may have only a few reviews (or sometimes even none) on this system. Therefore, it's hard to obtain the user's preference to make recommendations from past reviews. The data sparsity problem usually exists in recommender systems with great number of items. Sometimes, users only rate or purchase a small 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 are sometimes affected by their friends [Gupta 2008]. On the other hand, Golbeck pointed out that a recommendation can be made if there is just one path existing in a social network. In particular, the efficiency and accuracy can be increased by not only utilizing the preference of users, but also combining with the structure of social networks [Golbeck 2006].

Recently, some researches related to social network-based recommendation systems are being carried out, Gupta et al. pointed that the user's decision will be influenced by their friends in the network [Gupta et al. 2008]. Golbeck used the weight of trust to make recommendation for users [Golbeck 2006]. However, privacy issue is a serious concern when using the factor of trust relation which is not easy to be obtained in a social network. Sometimes, the relations which users declared publicly are not always the real relations [Huberman et al. 2009].

Furthermore, Pham et al. adopt social relationships to identify similar communities and then make recommendation through these communities [Pham et al. 2011]. Nowadays, some researchers are focusing on how to improve recommendation by using the factor of trust. Pitsilis et al. pointed out that trust relation is useful in applying the concept of clustering for recommendation [Pitsilis et al. 2011]. Seth and Zhang proposed a method by using strength-of-weak-ties to bring diverse recommendations to target user [Seth and Zhang 2008]. They address that the weak ties connecting two groups in a network means there are some differences between these two groups. However, the dataset used in their experiment is a quite small scale dataset and a user can only be clustered into only one cluster. Golder and Yardi focused on people who are in the second-degree network of the target user, trying to find out who are sharing interests or having a relation of reciprocity [Golder and Yardi 2010].

However, most of those researches focus on the data from Facebook.com or the user profiles in Orkut. A few researchers are focusing on the message content-and the network structure in microblogs. Therefore, in this paper, Plurk (which is the microblog with highest usage in Taiwan) will be utilized as the data source and the platform to implement the propose recommendation system.

Currently, most recommendation systems are designed to find friends or followers for users. In this paper, we suggest to recommend a list of the target users for companies. Our purpose is to use this recommendation system to understand the features of users and in that way support the companies' marketing strategies. Accordingly, we will define different product characteristics from users in Plurk. If a company plans to promote a product, our recommendation system will find the appropriate target users who have the most similar SNA measurements.

3 The Concept of the Proposed Approach

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

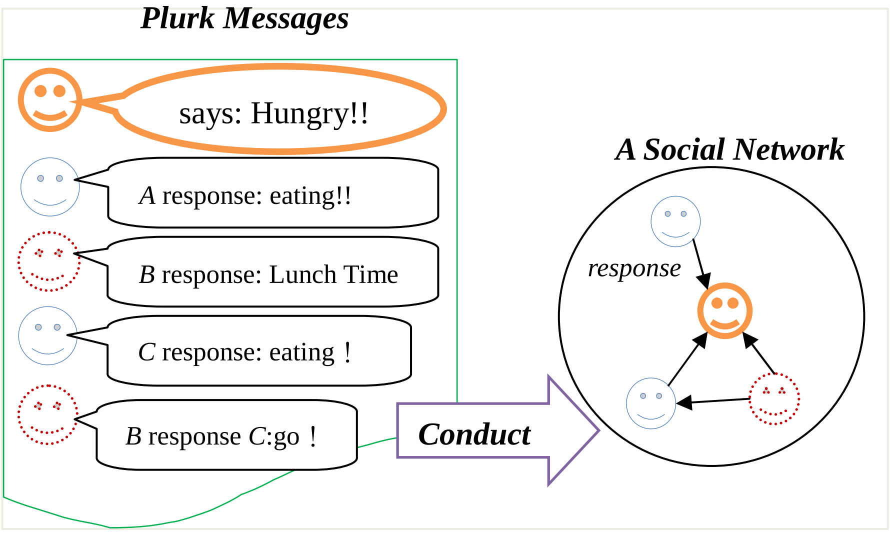


Figure 3: The concept of the proposed approach

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 messages and social data will be pre-processed and the meaningful keywords will be kept. Furthermore, the target marketing theory will be used to pre-define the categories of products and target users to conduct the relationship between product features and social network structure of users. In Figure 3, the system is 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 is used in the recommendation system. The process of the three parts will be introduced in detail below.

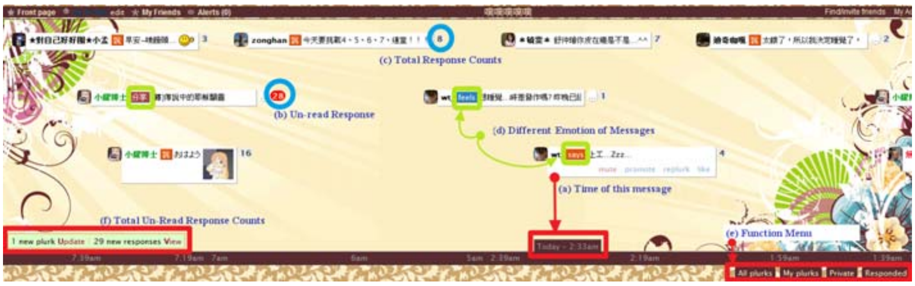
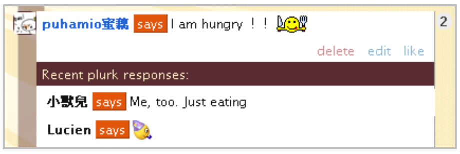


Figure 4: T he interface of the Plurk system

Figure 4 illustrates the interface of the Plurk system which is a timeline that displays the messages received in a chronological order. This mechanism is here referred to as “Time River”, where users can drag to see all the messages. The “Time River” shows the following elements as well as information: (a) time of this message, (b) unread message, (c) total response counts, (d) different emotion: the emotion that users would like to describe the category of the message, such as loves and wonders, (e) function menu: displayed to the boot that users can switch to display different types of messages, and (f) total un-read responses counts.

3.1.1 Data Collection

The message on microblogs is very short, simple, conversational and being updated quickly, which totally differs from that of traditional blogs. As the messages usually feature in short and conversational, the data from the messages of microblogs are more inclined to show users’ real preference.



puhamio蜜穢 says I am hungry ！！

delete edit like

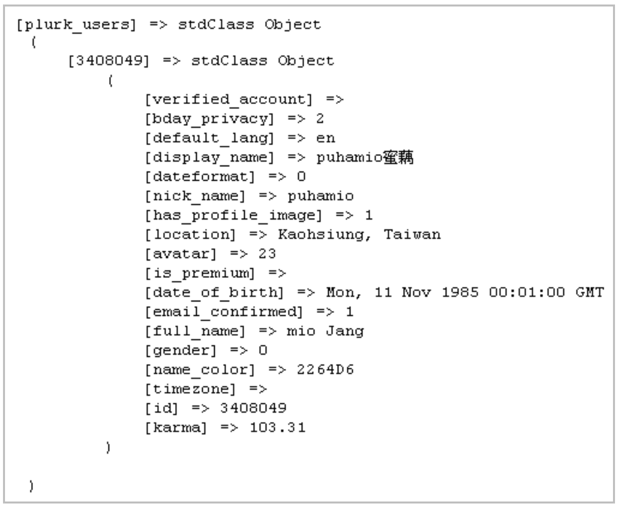
Recent plurk responses:

小獸兒 says Me, too. Just eating

Lucien says

Figure 5: A sample of the message and response in Plurk

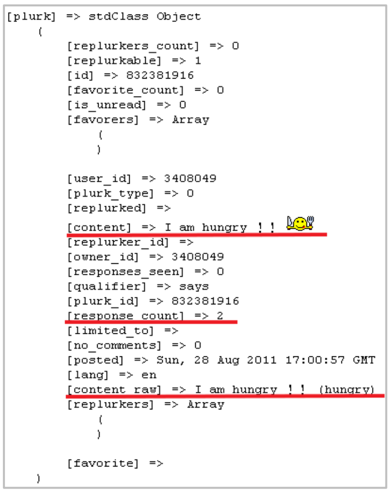
Figure 5 exhibits an example of the messages and interactions in Plurk between author, *id: puhamio*, and her friends and followers. The content may contain not only text but also individual images, emotion pictures, uploaded pictures and links. In this figure, two responses follow the original post by *id: puhamio*.



```
[plurk_users] => stdClass Object
 (
     [3408049] => stdClass Object
         (
             [verified_account] =>
             [bday_privacy] => 2
             [default_lang] => en
             [display_name] => puhamio蜜耩
             [dateformat] => 0
             [nick_name] => puhamio
             [has_profile_image] => 1
             [location] => Kaohsiung, Taiwan
             [avatar] => 23
             [is_premium] =>
             [date_of_birth] => Mon, 11 Nov 1985 00:01:00 GMT
             [email_confirmed] => 1
             [full_name] => mio Jang
             [gender] => 0
             [name_color] => 2264D6
             [timezone] =>
             [id] => 3408049
             [karma] => 103.31
         )

 )
```

Figure 6: The data structure of the raw data of the message (1)



```
[plurk] => stdClass Object
    (
        [replurkers_count] => 0
        [replurkable] => 1
        [id] => 832381916
        [favorite_count] => 0
        [is_unread] => 0
        [favorers] => Array
            (
            )

        [user_id] => 3408049
        [plurk_type] => 0
        [replurked] =>
        [content] => I am hungry !!
        [replurker_id] =>
        [owner_id] => 3408049
        [responses_seen] => 0
        [qualifier] => says
        [plurk_id] => 832381916
        [response_count] => 2
        [limited_to] =>
        [no_comments] => 0
        [posted] => Sun, 28 Aug 2011 17:00:57 GMT
        [lang] => en
        [content_raw] => I am hungry !! (hungry)
        [replurkers] => Array
            (
            )

        [favorite] =>
    )
```

Figure 7: The data structure of the raw data of the message (2)

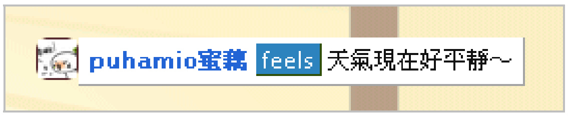
Owing to an increasing number of applications are trying to use the data in Plurk, API (Application Programming Interface) is also provided to help researchers to acquire the messages. Figure 6 and 7 is the source data of the message in figure 5 which is provided by Plurk API. The data contain two parts of information-*plurk* and *users*. By viewing the API, we can acquire the user's personal information, such as karma value, location and unique id. We can also get the information of each message in detail such as response counts and favourite counts.

In Plurk message, a lot of noise data exist. Firstly, we make a list in table 1 covering the top-10 stop-words which will be eliminated in our experiment. Besides that, lots of robots have been designed to make responses to users' messages automatically. These messages made by the robots will be removed as well since the research is focusing on the real people's thoughts and behaviors.

	Terms in Chinese	**Meanings in English**
1	可以	Can
2	哈哈	Laugh
3	現在	Now
4	LOL	An emotion signal
5	知道	Know
6	大家	Everyone
7	真的	Real
8	什麼	What
9	今天	Today
10	沒有	Nothing

Table 1: The Top-10 terms that will be recognized as noises in our experiment

For content processing, CKIP (Chinese Knowledge Information Processing), a Chinese morphological analysis tool, is brought into play. It can help to define the parts of speech of each term. For example, Figure 8 is a message about "The weather is peaceful now." After the process of CKIP, we can get a string "天氣(Na)現在(Nd) 好(Dfa) 平靜(VH).", translation in English is "Weather(Na) Now(Nd) Good(Dfa) Peaceful(VH)." The contents in the parentheses show the part of speech of each term. Therefore, from the list in table 1, "現在" means now and will be removed.



puhamio蜜糖 feels 天氣現在好平靜~

Figure 8: A sample Plurk message

3.1.2 Database

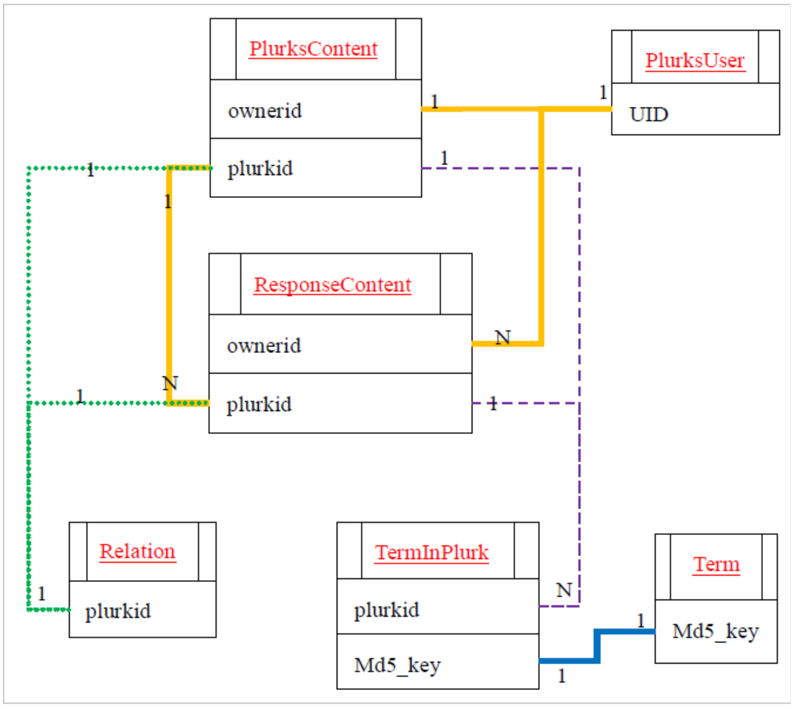


Figure 9: The E-R diagram of the database

After the completion of data collection and extraction, the output will then be stored into a database. The database is designed according to the structure of the message in microblogs. Two fields are covered in the tables 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*. On the other hand, the field *Relation* will be employed to store the relationships in a message. The terms of each message will be extracted and stored in the fields *TermInPlurk* and *Term*. Figure 9 illustrates the E-R (Entity Relationship) diagram of the database.

3.1.3 Keyword Extraction

Due to the data from microblogs are un-structured, the keywords in the message must be extracted firstly by performing 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, TF (Term-Frequency) and TF-IDF (Term Frequency-Inverse Document Frequency) are basic measurements which can be applied to measure the keyword frequency. TF means the term frequency in a document. TF-IDF represents the importance of words among all documents. The two measurements are defined as the formulas below:

$$TF = freq(i,j) / Maxfreq(l,j) \quad (4)$$

$$IDF = log(N/ni) \quad (5)$$

$$TF\text{-}IDF= TF×IDF \quad (6)$$

	Term	Meaning in English	Frequency in This message (document).	Frequency in all dataset	TFIDF
1	**戴佩妮**	Name of a singer	**4**	1	**0.70644**
2	**薔薇**	Rose	**4**	1	**0.70644**
3	**youtube**		**4**	**13**	**0.36445**
4	**2010**		**2**	1	**0.35322**
5	作品	Creations	1	1	0.176611
6	Live		1	1	0.176611
7	Concert		1	1	0.176611
8	新歌	New Song	1	1	0.176611
10	演唱會版	Concert version	1	1	0.176611
11	2008 年	Year	1	1	0.176611
12	發行	Publish	1	1	0.176611
13	tanya	Name of a signer	1	1	0.176611
14	點播	Pick a song	1	1	0.176611
15	水水	Pretty	1	1	0.176611

Table 2: The TF-IDF value of a message in Plurk

In table 2, we rank the TF-IDF value in a descending order that allow us to choose the most important terms in the message. According to the results, the top-4 terms are related to music, such as the name of the singer or Youtube. Therefore, TF-IDF is helpful to extract the keywords from Plurk messages.

In this stage, we have been replacing, removing and retaining the terms. In retaining terms, Chinese terms usually contain two characters in order to be meaningful. Therefore, we have to retain some meaningful terms which just carries only one character in Chinese. For example, “eat” in Chinese is one character, so we keep this term. In the stage of replacing, the hyperlinks belonging to Youtube.com will be replaced to "Youtube", due to those hyperlinks don't contain any meanings.

Most of the messages in Plurk contain emotional pictures as showed in table 3. The emotional picture is very helpful for us to understand the meaning of the content. For example, if the raw content is hungry, the topic of underlying message might be related to the discussion of food.

Emotional Picture	Raw Content
	(music)
	(hungry)
	(mmm)

Table 3: Emotional pictures and the raw content in Plurk

3.1.4 The SNA Modules

In order to extract the important characteristics from Plurk, the measurements in SNA are adopted in this paper on the grounds of the features and network structure of Plurk.

The responses in Plurk are used to generate directed social networks. The ties in a network denote the response relation in those messages. The direction of an arrow means the target of user's response. Most of the SNA measurements are meaningful for describing social networks. Generally speaking, in a directed graph, there are two different types of *degree* which are named *in-degree* and *out-degree*. *Out-degree* usually means expansiveness [Scott, 2000]. A user with large out-degree implies that this user makes responses to the others very frequently. That is to say, actors have higher out-degree are inclined to help the companies more easier to spread the promotion information.

On the other hand, *in-degree* means receptivity or popularity [Scott, 2000]. A user with a large amount of in-degree relation indicates that this user receives many responses from other users. This actor is easier to has others' attentions comparatively and then to be more influential.

Accordingly, the company's image building and word-of-mouth strategy could be implemented by locating this actor. Formulation 7 and 8 are used to calculate the value of in-degree and out-degree. In the formulations, d_i denotes in-degree and d_o denotes out-degree [Scott, 2000]:

$$d_I = \frac{\sum_{i=1}^{g} d_I(n_i)}{g} \qquad (7)$$

$$d_O = \frac{\sum_{i=1}^{g} d_O(n_i)}{g} \tag{8}$$

Another useful SNA measurement is density which is the average proportion of lines in a graph. The density is symbolized by △ in formula 9. For there are *g* nodes and maximum possible number of lines could present by g(g-1)/2 and *L* is the number of lines in this graph. If the density is equal to 1, it is called a *complete* graph. In other words, those users in the graph made responses to all others users in the same network. This measure also can be used to evaluate the *cohesiveness* [Scott, 2000].

$$\Delta = \frac{L}{g(g-1)/2} \tag{9}$$

Therefore, the higher density of a network means users in the network have strong passion and pay more attentions to it.

C*loseness, betweenness* are the measurements that have already been discussed in section 2. Some measurements may have some meanings in common, such as a higher density also have a higher closeness. Therefore, if this rule is not in a characteristic vector, we can discuss more about this interesting phenomenon. In this paper, we propose to use the SNA measurements to define the characteristic vector of each product from user's social networks.

3.2 Target Definition Based on The Concept of Target Marketing

The main purpose of this paper is to conduct the user's characteristic vector of different products with the support of the theory of target marketing. As shown in Figure 10, user and product are treated as different target in marketing. For example, different users may have different interests. The user who is keen on music may be prone to update messages related to the subject of music, such as "Linkin Park" or "Mariah Carey", or have some links such as YouTube. Therefore, 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 arrangement such as 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 the 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.

Formula 10 and 11 are two feature vectors to identify the features of product category and user. For each $User_i$, $Keyword_n$ is the keywords that extracted from the user's messages, and Cat_m indicates the user's product preference. SNA_p represents the measurements of the social networks structure. Each category of product is pre-defined as Cat_j, the ranked keyword *x* is represented as $Keyword_x$. The social networks measurement SNA_y is denoted as ranked yth SNA measurement of a product category.

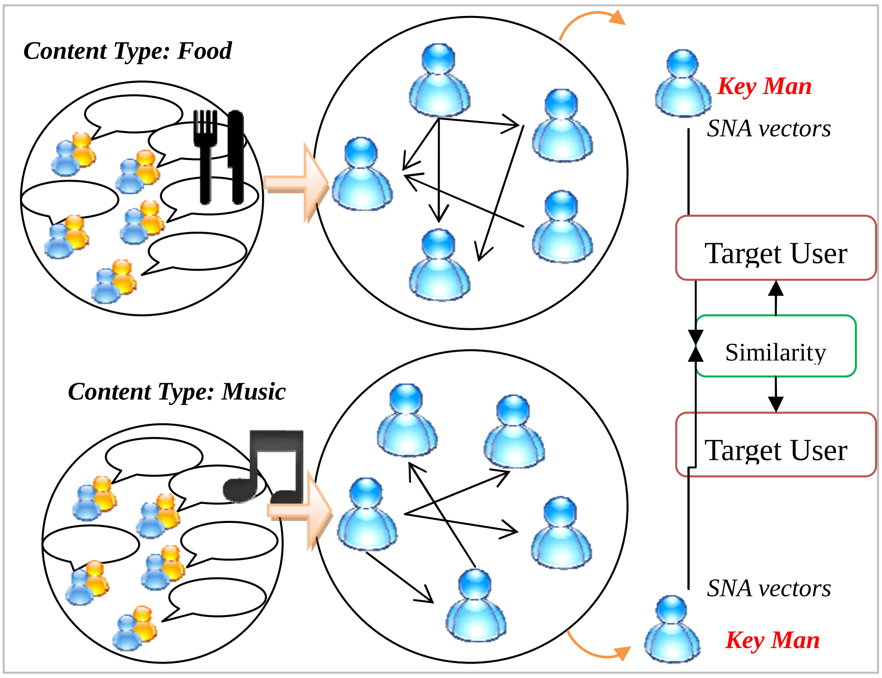


Figure 10: The process for feature understanding and recommendation

$$Cat_j = \{Keyword_1,\ldots,Keyword_x,\ SNA_1\ldots,SNA_y\ \} \qquad (10)$$

$$User_i = \{Keyword_1,\ldots,Keyword_n,\ Cat_1\ldots,Cat_m,SNA_1\ldots,SNA_p\ \} \qquad (11)$$

For example, we are using the relationships of responses to generate the social graph. In the social networks, most of the messages are related to music. The keywords of the messages are extracted by TF-IDF and then the vector Cat_{music} will be conducted. At the same time, those SNA metric are also calculated based on the social networks. For $User_A$ and $User_B$, their keywords and SNA measurements: *degree*, *closeness*, *betweenness* and *density* are described in (12) and (13). The $\boldsymbol{Cat_{default}}$ is an uncertain value until the category has been decided by using the similarity between SNA measurements.

$$User_A = \{youtube,\ mayday,\ song,\ \boldsymbol{Cat_{default}},\ 31.03,\ 58.06,\ 3.6,\ 0.61\} \qquad (12)$$
$$User_B = \{youtube,\ jay,\ song,\ Taiwan,\ book,\ \boldsymbol{Cat_{default}},\ 29.53,\ 59.16,\ 3.4,\ 0.51\} \qquad (13)$$

Therefore, our propose method uses Plurk messages to generate social graph and calculate those social networks measurements. Generally speaking, we combine the microblogs content with the social relation for recommendation. Although those content in the microblogs are very noisy and hard to be analysed, they are still useful for recommendation. According to our method, we have pre-processed these messages and grouped messages with different type of products together. Then, the vector of product will be defined. Since the microblogs is a highly dynamic web service and the messages are updated very quickly, we have pre-defined the SNA vectors of products to help us to find the target users from the huge dynamic social relations.

In other words, user's interests could be changed over time, but the social structure of different products would not be changed. Therefore, a social recommendation system can help to traverse the target users extensively by using the social relationships. Furthermore, we can find the *key man* in each social graph and using this *key man's* vectors to find the target users who have similar SNA vectors, as shown in Figure 10. A *key man* can be an opinion leader or an actor has higher influence. According to the research result of Cho et al., they suggest that marketers can use *distance centrality* to identify an opinion leader which can effectively make diffusion [Cho et al. 2011]. On top of that, *sociality* and *centrality* have been proved as the most effective measurement to get higher diffusion. Therefore, these SNA measurements can be treated as the baseline for us to find similar users for recommendation.

4 An Empirical Study

4.1 The pre-processed Data

In this section, we will demonstrate an empirical study to show how the proposed approach works, and Plurk is the platform for the empirical study. The pre-processed raw data from Plruk will then be stored in database. The structure of the database is shown in Table 4 and Table 5. Table 4 is the database structure of the user and Table 5 is the database structure of the messages in Plurk.

RowName	Meaning	Example
id	Users unique id	3182573
nick_name	User' s account name	mynameishey
karma	Usage value	93.23
gender	0=female , 1=male	0
location	User location	Kaohsiung, Taiwan

Table 4: The database structure of user profile

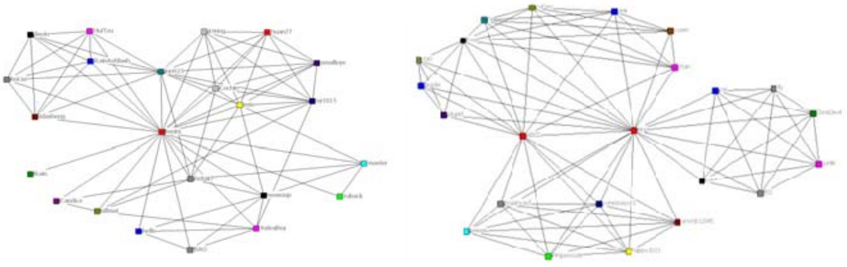
rowName	meaning	example
userID	Target user	3182573
ownerID	Message author	3182573
qualifier	The type of post (is, says, asks, ...)	says
content_raw	Message content	Good night
plurkID	The unique id of the message	589257961
responesCount	Response count	3
favorCount	The count of who likes this message	1
favorers	The list of who likes this message	3408049
posted	Timestamp	Thu, 23 Dec 2010 14:34:49 GMT

Table 5: The database structure of Plurk message

4.2 Keywords and The Social Structure

In this empirical study, we choose the keywords that are related to *Food (ie: the name of a restaurant)*, *3C (such as ipad)*, and a dataset which has been extracted accordingly. Then, UCINET is used to generate the social graph and calculate the SNA measurements to help the understanding the features of products and social network structure.

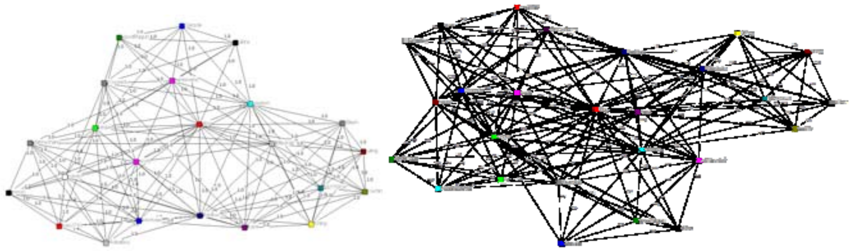
In Figure 10, the clusters in the target network A and B can be defined very easily through the visualized social network graph. Figure 11 (a) is the network graph based on the message type of food of target A and Figure 11 (b) is the network graph of target B. Figure 12 (a) is the network graph based on the message type of 3C of target A and Figure 12 (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.



(a) *(b)*

Figure 11: Type of Food in Target A and B (a) Target A with 23 actors and 9 messages (b) Target B with 23 actors and 4 messages

The details of SNA measurements are shown in Table 6. In Table 6, we can find in the type of Food, the SNA values of target A and B are very close. As the same phenomenon, in the type of 3C the values are also close with each other. For the type of 3C, the measurements of density and closeness both point that actors are close to each other and closer than the type of Food. Also, in the type of 3C, the average geodesic path length (Avg. Dist.) is shorter than that of the type of Food. The analysis results indicate that users in the type of 3C are more condensed. On the other hand, the higher density means that more homogeneous users in this product type. From Table 6, a big gap of those SNA measurements can also be found between type Food and 3C. In other words, the SNA measurement can be applied to define the characteristics vector of different products based on the social network features. The analysis results indicate that the proposed method is reasonable and can be discussed further in the future.



(a) (b)

Figure 12: Target B in the type of 3C (a) 23 actors (b) 4 messages

Type	**Food**		**3C**	
SNA \ User	Target A	Target B	Target B with 23 actors	Target B with 4 messages
Degree	31.225	34.783	51.799	48.000
Closeness	60.260	61.475	68.547	66.937
Betweenness	3.275	3.106	2.296	2.167
Density	0.631	0.643	0.726	0.714
Avg Dist	1.866	1.854	1.769	1.781
Norm Dist	0.798	0.803	0.842	0.837

Table 6: The value of SNA measurements of target A and B of the type of Food and 3C

5 Conclusion and Discussion

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 presented in this paper demonstrates how the recommendation system works based on real data collected from Plurk. From the analysis results, we can find the difference significantly of the SNA measurement between different products. 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.

From the literatures, it is a very complex work and the performance is poor when analyzing the content of microblogs. However, the content of microblogs is still being considered as very useful resources. In our approach, we are trying to keep the content as an important material to train and generate the SNA vectors. Based on those vectors from the content of microblogs the accuracy of clustering can then be increased and which are useful for us to find similar target users.

In the future, we are going to implement the recommendation system based on the approach that developed in this paper. However, there are still some challenges for use to deal with. For example, how to identify the key mans from the discovered clusters. Furthermore, how to design an ideal mechanism as the core of the recommendation system is also what we will focus in the future.

References

[Adomavicius and Tuzhilin 2005] Adomavicius, G., and Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 2005, 734-749.

[Billsus et al., 1998] Billsus, Daniel, and Pazzani, M. J.: Learning collaborative information filters, *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998.

[Breese et al., 1998] Breese, J. S., Heckerman, D., and Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering, *Proceedings of the 14th conference on Uncertainty in Artificial Intelligence*, *461*(8), 1998, 43-52.

[Cho et al., 2011] Cho, Y., Hwang, J., Lee. D.: Identification of Effective Opinion Leaders in the Diffusion of Technological Innovation: A Social Network Approach, *Technological Forecasting and Social Change*, 2011.

[Ediger et al., 2010] Ediger, D., Jiang, K., Riedy, J., Bader, D. A., and Corley, C.: Massive Social Network Analysis: Mining Twitter for Social Good, *2010 39th International Conference on Parallel Processing*, 2010, 583-593.

[Godbole and Godbole 2007] Godbole, N., Skiena, S.: Large-Scale Sentiment Analysis for News and Blogs, *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007, 219-222.

[Golbeck 2006] Golbeck, J.: Generating Predictive Movie Recommendations from Trust in Social Networks, *In Proceedings of the 4th International Conference on Trust Management*, *3986*, 2006, 93-104.

[Golder and Yardi 2010] Golder, S. A., Yardi, S.: Structural Predictors of Tie Formation in Twitter: Transitivity and Mutuality, *Proceedings of the Second IEEE International Conference on Social Computing*, 2010, 88-95.

[Goodreau 2007] Goodreau, S. M.: Advances in Exponential Random Graph (p*) Models Applied to a Large Social Network, *Social Networks*, *29*(2), 2007, 231-248.

[Gupta et al., 2008] Gupta, A., Jain, R., and Song, S.: Movie Recommendations Using Social Networks. *Stanford University Stanford, CA,2008.*

[Hannon et al., 2010] Hannon, J., Bennett, M., and Smyth, B.: Recommending Twitter Users to Follow using Content and Collaborative Filtering Approaches, *Proceedings of the fourth ACM conference on Recommender systems*, 2010, 199-206.

[He and Chu 2010] He, J., Chu, W. W. (2010). A Social Network-Based Recommender System (SNRS). *Data Mining for Social Network Data*, *12*, 2010, 47-74.

[Ho et al., 2011] Ho, C. T., Li, C. T., and Lin, S. D.: Modeling and Visualizing Information Propagation in a Micro-blogging Platform, *2011 International Conference on Advances in Social Networks Analysis and Mining*, 2011, 328-335.

[Hsu et al., 2010] Hsu, C. L., Liu, C. C., Lee, Y. D.: Effect of Commitment and Trust Towards Micro-Blogs on Consumer Behavioral Intention: A Relationship Marketing Perspective, *International Journal of Electronic Business*, *8*(4), 2010, 292-303.

[Huberman et al., 2009] Huberman, B. A., Romero, D. M., and Wu, F.: Social networks that matter: Twitter under the microscope, *First Monday*, *14*(1), 2009.

[Java et al., 2007] Java, A., Song, X., Finin, T., and Tseng, B.: Why We Twitter: Understanding Microblogging Usage and Communities, *Proceedings of the 9th WebKDD and 1st SNAKDD 2007 workshop on Web mining and social network analysis*, 2007, 56-65.

[Jun et al., 2006] Jun, T., Kim, J. Y., Kim, B. J., and Choi, M. Y.: Consumer referral in a small world network, *Social Networks, 28*(3), 2006, 232-246.

[Jung 2005] Jung, J. J.: Visualizing Recommendation Flow on Social Network, *Journal of Universal Computer Science*, *11*(11), 2005, 1780-1791.

[Jung 2008] Jung, J. J.: Ontology-based Context Synchronization for Ad Hoc Social Collaborations, *Knowledge-Based Systems*, 21 (7), 2008, 573-580.

[Jung 2009] Jung, J. J.: Social grid platform for collaborative online learning on blogosphere: a case study of eLearning@BlogGrid, *Expert Systems with Applications*, *36(2)*, 2177-2186, 2009.

[Jung 2010] Jung, J. J.: Reusing Ontology Mappings for Query Segmentation and Routing in Semantic Peer-to-Peer Environment, *Information Sciences*, 180 (17), 2010, 3248-3257.

[Jung 2012a] Jung, J. J.: Attribute selection-based recommendation framework for short-head user group: An empirical study by MovieLens and IMDB, *Expert Systems with Applications*, 39 (4), 2012, 4049-4054.

[Jung 2012b] Jung, J. J.: Evolutionary Approach for Semantic-based Query Sampling in Large-scale Information Sources, *Information Sciences*, 182 (1), 2012, 30-39.

[Lee et al., 2011] Lee, C. H., Yang, H. C., Chien, T. F., and Wen, W. S.: A Novel Approach for Event Detection by Mining Spatio-temporal Information on Microblogs, *2011 International Conference on Advances in Social Networks Analysis and Mining*, 2011, 254-259.

[Lento et al., 2006] Lento, T., Welser, H., Gu, L., and Smith, M.: The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop Weblogging System, *In Proceedings of the 15th International World Wide Web Conference*, 2006.

[Matsuo et al., 2007] Matsuo, Y., Mori, J., Hamasaki, M., Nishimura, T., Takeda, H., Hasida, K., and Ishizuka, M.: POLYPHONET: An Advanced Social Network Extraction System from the Web, *Web Semantics Science Services and Agents on the World Wide Web*, *5*(4), 2007, 262-278.

[McCarthy and Perreault 1990] McCarthy, E. Jerome , Perreault, W. D.: Basic Marketing: A Managerial Approach, *University of California*, 1990, 72-73, Irwin (Homewood, IL).

[Mika 2005] Mika, P.: Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks, *Web Semantics Science Services and Agents on the World Wide Web*, *3*(2-3), 2005, 211-223.

[Mooney and Roy 1999] Mooney, R. J., Roy, L.: Content-Based Book Recommending Using Learning for Text Categorization, *Proceedings of the fifth ACM conference on Digital libraries DL 00*, August 1999.

[Pal and Counts 2011] Pal, A., Counts, S.: Identifying Topical Authorities in Microblogs, *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, 45-54.

[Pazzani and Billsus 1997] Pazzani, M., Billsus, D.: Learning and Revising User Profiles: The Identification of Interesting Web Sites, *Machine Learning*, *27*(3), 1997, 313-331.

[Pham et al., 2011] Pham, M. C., Cao, Y., Klamma, R., and Jarke, M.: A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis, *Journal of Universal Computer Science*, *17*(4), 2011, 583-604.

[Pitsilis et al., 2011] Pitsilis, G., Zhang, X., and Wang, W.: Clustering Recommenders in Collaborative Filtering Using Explicit Trust Information, *IFIP Advances in Information and Communication Technology*, *358*, 2011, 82-97.

[Resnick et al., 1994] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J.: GroupLens: an Open Architecture for Collaborative Filtering of Netnews, *In Proceedings of the 1994 ACM conference on Computer supported cooperative work*, 1994, 175-186

[Sarwar et al., 2001] Sarwar, B., Karypis, G., Konstan, J., and Reidl, J.: Item-based Collaborative Filtering Recommendation Algorithms, *Proceedings of the tenth international conference on World Wide Web WWW 01*, 2001, 285-295.

[Scott 2000] Scott, J.: Social Network Analysis: A Handbook, *London: Sage.*

[Seth and Zhang 2008] Seth, A., Zhang, J.: A Social Network based Approach to Personalized Recommendation of Participatory, *Conference on Weblogs and Social Media (ICWSM 2008)*.

[Ting 2008] Ting, I. H.: Web Mining Techniques for On-Line Social Networks Analysis, *Proceedings of the 5th International Conference on Service Systems and Service Management*, 2008, 696-700.

[De Valck et al., 2009] De Valck, K., Van Bruggen, G. H., and Wierenga, B.: Virtual Communities: A Marketing Perspective, *Decision Support Systems*, *47*(3), 2009, 185-203.

[Wang et al., 2011] Wang, K. Y., Thongpapanl, N., Wu, H. J., and Ting, I. H.: Identifying Structural Heterogeneities between Online Social Networks for Effective Word-of-Mouth Marketing, *2011 International Conference on Advances in Social Networks Analysis and Mining*, 2011, 418-422.

[Wilson 1989] Wilson, D. S.: Levels Of Selection: An Alternative to Individualism in Biology and the Human Sciences, *Social Networks*, *11*(3), 1989, 257-272.

[Yamaguchi et al., 2011] Yamaguchi, Y., Amagasa, T., and Kitagawa, H.: Tag-based User Topic Discovery Using Twitter Lists, *2011 International Conference on Advances in Social Networks Analysis and Mining*, 2011, 13-20.

[Zhao and Rosson 2009]Zhao, D., Rosson, M. B.: How and Why People Twitter: The Role That Micro-Blogging Plays in Informal Communication at Work, *GROUP 09 Proceedings of the ACM 2009 international conference on Supporting group work*, 2009, 243-252