2012 International Conference on Network and Computational Intelligence (ICNCI 2012)
IPCSIT vol. 46 (2012) © (2012) IACSIT Press, Singapore

Analysis and Evaluation on Weighted Instant Messaging Network

Lin LU [+], Jun-yong LUO, Yan LIU and Ming-tao LI

Information Science and Technology Institute
Zhengzhou, Henan, 450002, China

Abstract. People registered on an instant messaging (IM) system shared their state online and exchange instant messages by text, audio, video, file transfer or even some other expanding forms in the future. While their communication via IM system, a new type of social network appeared on the Internet. Comparing to the email network, IM supports multi-media communication on data exchange between users. And evaluating the relationship between two connected users had to consider multiple factors synthetically.

This paper discussed the characteristics on tight relation on user communication with a new design taking into account three aspects in communication. And we proposed a compositive method to achieve the edges weighting in an IM network. Edges were weighted and ranked under different assumptions as well as a synthetical one. As a result, the final IM communication sequence showed a rather different statistical behavior than any single factor based. The research showed a new thought in SNA (social network analysis) and had an improved space to be upgrade.

Keywords: instant messaging (IM), Weighted social network, Social Presence, TOPSIS

1. Introduction

Instant messaging is a kind of near-synchronous communication. With a fast network, transmission times are fractions of a second and the experience is of near-synchronous interaction. Like chat, IM allows users to type messages into a window, but like the phone, it is based on a dyadic "call" model. Most IM systems also provide awareness information about the presence of others. In Tencent's Instant Messenger (QQ), the user creates a "buddy list" of people to monitor. A buddy list window shows whether buddies are currently online, their emotional state, and whether they are active or busy. Users can call their buddies in the buddy list and talked with them after response. These are the two mainly aspects service, presence and instant message exchange, which IM supplied to register people.

With the instant messages sent and received between buddies, we can extract a special network from a specific IM system. Either in which form the message reached, transit or direct, there are two user nodes in an IM network Diagram with an edge linked. By this way, we'll get an IM network and use existed SNA (Social Networks Analysis) technology to research. While IM system supplied different service from email and webpage, where are more forms in communication such as text, voice, video and files transferring and any other different types. So general method was not good enough to solve the multi-messaging problems. In this paper, we focused on the communication evaluating in IM network.

2. Related Work

Most of the studies used the frequency of effective communication among users as a weight for analysis. For example, A Barrat[1] researched on the IM network in 2004 just with the frequency for standard with closeness for impact. In 2005, Yihjia Tsai[2][3] studied in the e-mail communication network (ECN) and defined weights for times of communication between email users, as the same with the count of letters sent

[+] Corresponding author. Tel.: 18611421760
E-mail address: anna5260@sina.com.

and receipt through an edge of two-way communication between two users email. In 2007, JP Onnela[4] established a weighted network according to the mobile phone call records and for each side existed two weights to reflect the social interaction strength. These two weights were the total call duration and cumulative call time. In 2010, Zhao Yuan-ping[5] worked in information dissemination process using IM user's connectivity as a user authority to judge the credibility of the user comments for message propagation modeling in the IM networks. They introduced fitness parameter to calculate the edge weight in evaluating users' closeness in communication.

IM communication has some important features of traditional face-to-face interpersonal communication, such as instantaneous perception of buddies online state, edited personal signatures, rich emotional images and so on. With the popularity and its social relationship with its buddies, we evaluated relationships should consider more indexes rather than merely through the communication frequency.

In this paper, we discussed the characteristics on close relation between people with a new design taking into account three aspects in communication. And we proposed a compositive method based on the TOPSIS theory to achieve the edges weighted in an IM network. We compared the edge weight and the order under different assumptions. As a result, the final ranked IM communication sequence showed a rather different statistical behavior from any single element based.

3. IM Network Weighting Analysis

As any other kind of communication analysis, there are two user-points at the end of one edge emblematizing a communicated relationship in a network topology graph $G(V,E)$, where $V=\{1,2,...\}$ was the nodes set and $E=\{e(a,b)\,|\,a,b\in V, a\neq b\}$ was the edges set. As shown in Fig.1, it was an example of social network and also one was our experiment in real implement.

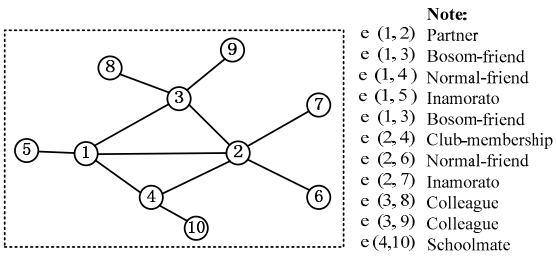

Figure 1 an Example of IM social network

3.1. Multiple Descriptions

- *Kinetic behavior*

Under normal circumstances, if two users communicated more frequently, they had a higher activity of communications. And also if there was a large time span from their earliest connection to the last one that we inspected, it has more possibility that they were a pair of old friends. As the multimedia communications in the instant messaging system behaviour (such as audio and video chat) is difficult to capture, in this paper, we extracted the text based instant messages and the session initiate messages in the multimedia connection stage to record and calculate the number and the time span of multimedia communications instead of their true P2P messages.

- *Social Presence impact*

In a general sense, if the presence in communication between two users is stronger, their relationship must be more intimate. In1976, Short, Williams and Christie first proposed the Social Presence Theory[6], which depicted the different effects when people communicate with Face-to-face, voice and text based interactions. And the conclusion was that images, voices and the texts were weakened in turn in present feelings. A high degree of social presence will have a higher sense of integration into the people, also has a higher degree of intimacy. In this way, we defined in network applications that IM users also had a social presence in Video, audio, file transfer, text chat, weighted from strong to weak in the medium of instant messaging between the user communicated behaviors. At contrary to mail communication, IM network supports multimedia communications, so the weight assessment in user communication needs to distinguish among different types in the IM communication.

- *Social community*

It is more objective to take the social circle of different users as an impact factor in evaluating the strength of edges.

Generally speaking, if a user's social contacts are more extensive, the smaller tightness averaged on a single contact communication. Conversely, if a user's social range is smaller, then a contact with it more closely. In another point, if there are more common contacts of the two users, their social circle is more overlap, the more closely the relationship between two people.

3.2. Indexes calculation

As shown in Fig 2, each edge was impacted by a few elements, the end nodes, neighbor nodes and neighbors' edges. Attribute information extracted by index calculation as follows:

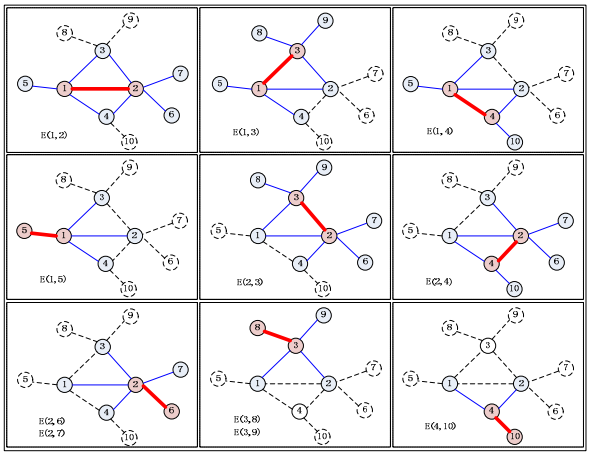

Figure 2 Edges extracted based on Fig.1

- *Communication Times*

Times stand for the frequency of people contact in a specific period, the more times in exchanges, the more closely between users. In calculation, the times value on edge $e(a,b)$ is by formula (1).

$$comm\_times_{ab} = send_{ab} + receive_{ab} \tag{1}$$

where $send_{ab}$ and $receive_{ab}$ signs the directions of the messages from a to b and b to a respectively.

- *Communication Duration*

The longer the duration of communications in the same type of behavior, the greater in the user closeness. For short text chat, one contact length can be gained in margin calculation, by the timestamp of the last message the first time in more than one message in the same session. Types of multimedia audio and video and file transfer session, due to P2P transmission and proprietary protocol, though had difficulties in capture, it still would be account by the establish and terminate timestamp of the control message in connection instead of the calculation real dialogue transport.

- *Communications time span*

In common, we all have 'Old Friend' that some people we don't contact frequently but still very good friend. In this sense, the longer that inter-node communication time span, the relationship more closely. Node a , b interaction time span as follows:

$$Span_{ab} = latest\_date_{ab} - earliest\_date_{ab} \tag{2}$$

$latest\_date_{ab}$ is the newest date when node a and b exchanges, and $earliest\_date_{ab}$ is the oldest in history during the monitored period.

- *Social Presence Score*

To mark a presence score according to the social presence theory for each contact edge in G. For the presence strength from strong to weak, given 25 points, 20 points and 15 points respectively for the type of video, audio and file transfer per time. Additionally, one texted chat message worth 1 point so that a short conversation consisted of a few message means more.

By this way, the Presence Score obtained between node a and b in formula (3):

$$Score_{presence} = \sum_{1}^{t} (value_t \times count_t) \tag{3}$$

- *Communications Extension*

The edge directly connected to the node is a neighbor node of a user. Suppose a user's social relationship averaged in each edge it directly connected with, in another word, a node had more neighbors then his communication social circle was larger but the weight averaged on each connection would be smaller. On the contrary, nodes with less neighbors, each edge would share more of the relationship between the weights.

In this means, we compute the weight using formula (4), in which S, T are the respective neighbor set of node a and b . $|S|$ and $|T|$ are the number of nodes in S and T respectively.

$$extension = |S \cup T| \tag{4}$$

And here we use $s+t-1$ for it, while some other way using $s \times t-1$ with reasons, too.

- *Communications Overlap*

If the two users had a lot of friends in common, the more overlap of their social circle, the more closely the relationship between these two people. In formula (5), S and T are the same meaning with (4):

$$overlap = \frac{|S \cap T|}{|S \cup T|} \tag{5}$$

4. Topsis based Method

4.1. TOPSIS about

The TOPSIS (Technique for Order Preference by Similarity to Ideal Solution)[7] proposed by CL.Hwang and K.Yoon was first published in 1987. The basic thought is based on the normalized raw data matrix, finding the optimal solution in the limited scenario and the worst solution, and optimal vectors respectively, then calculate the distance between the objects and the optimal solution and the worst programs of all evaluation. Evaluation of the object and the optimal solution is relatively close to the degree is to be the basis of the evaluation of the pros and cons.

Edge-based multi-attribute evaluation means to give a formal description on the comprehensive evaluation of the edge of the instant communication communications network: If the j^{th} indicator of the i^{th} edge is x_{ij}, ($i = 1,...,m$, $j = 1,...,n$). Based on TOPSIS theory we can construct a comprehensive evaluation function as formula (6):

$$y=\mathrm{f}(\mathbf{w},\mathbf{x}) \tag{6}$$

Where $\mathbf{w} = (w_1, w_2, \cdots, w_n)'$ is the target weight vector, $\mathbf{x} = (x_1, x_2, ..., x_n)$ is the assessed-value vector of each edge indexes for estimate, and y is the comprehensive evaluation value of the edge x .

4.2. Indicators and pretreatment

1) The availability selection. Select indicators to achieve extraction in fact.

In so many indicators, we have responding formula to calculate. But in Indicator “extension”, communication duration, we though also had method to compute on it, its sense had taken use in other indicators, for last time and accounts in conversations. So here we didn’t employ it, just using other 5 indicators would be enough.

2) The applicability selection. Indicators remove if its distinction is small and vague.

Our 5 indicators all had distinguishing ability in evaluating and ranking. And additional remove was not in need.

3) The trend of processing selection.

The indicators selected in this article are monotonous. Except the communications link indicator “extension”, all of indicators are excellent indicators. To solve this problem, we used reciprocal value of each in the calculation of the ranking. After finishing all the indicators in line with the greater the assessed value, the more important.

4.3. TOPSIS based Evaluation

Step1: Evaluation matrix $A_{m\times n}$ normalization

In matrix $A_{m\times n}$, in a row $x_{i\bullet}$ are indicators' values on an edge sequence while in a column $x_{\bullet j}$ those are the values on different edges under one indicator.

$$A_{m\times n}=\begin{bmatrix} Index_1 & Index_2 & \cdots & Index_n & \\ x_{11} & x_{12} & \cdots & x_{1n} & e_1 \\ x_{21} & x_{22} & \cdots & x_{2n} & e_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} & e_m \end{bmatrix} \tag{7}$$

Normalizing the matrix $A_{m\times n}$ for z_{ij} as formula (8):

$$z_{ij}=x_{ij}\Big/\sqrt{\sum_{i=1}^{m}x_{ij}^{\ 2}} \tag{8}$$

Step2: Confirming the decision matrix E

Under the j^{th} index, calculate the weighting percentage of edge i according to formula (9):

$$p_{ij}=x_{ij}\Big/\sum_{i=1}^{m}x_{ij} \tag{9}$$

Using the entropy weighting for the index, calculated as formula (10):

$$e_j=-k\sum_{i=1}^{m}p_{ij}\times\ln(p_{ij}) \tag{10}$$

Where $k=\dfrac{1}{\ln n}>0$, $e_j>0$. For a specific j, the smaller the complexity in x_j, then the greater e_j is. When x_{ij} are all equal, $e_j=e_{max}=1$.

The more complexity the indicator x_j,the bigger in its otherness coefficient $g_j=1-e_j$, and the better effect of g_j in clearly differentiate the role of the indicators.

According to the formula above to fix on the normalized weight coefficient w_j:

$$w_j=g_j\Big/\sum_{h=1}^{n}g_h \tag{11}$$

With $y_{ij}=w_j\times z_{ij}$, we got the weighted value of edge evaluation matrix, described as follows:

$$E=\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{bmatrix} \tag{12}$$

Step3: Euclidean distance calculation

The best value of the indicators was taken as a positive ideal solution y^+ and the worst value as the negative ideal solution y^-, respectively. Where the Euclidean distance s_i^+ from y_i to y^+ was shown in formula (13):

$$s_i^+=\sqrt{\sum_{j=1}^{n}(y_{ij}-y_j^{\ +})^2} \tag{13}$$

Similarly, the Euclidean distance s_i^- from y_i to y^- was shown in formula(14):

$$s_i^-=\sqrt{\sum_{j=1}^{n}(y_{ij}-y_j^{\ -})^2} \tag{14}$$

At last, the value of comprehensive assessment was evaluated in formula (15):

$$E(i)=\frac{S_i^-}{S_i^+ + S_i^-} \tag{15}$$

The bigger $E(i)$ valued on the edge, the more closely in communication relationship between the corresponding users.

5. Experiments

5.1. Data

First, there are 20 campus volunteers in our department took part in a basic information questionnaire. And the experiments set up a port mirroring at the campus gateway, in a one-month (30 days) period, on the campus network of 20 campus volunteers to use the IP address tracking, captured the communications data under the MSN protocol in texted chat messages and control link messages in voice, video, file transfer types.

5.2. Guide Line

According to the cases provided by the 20 campus volunteers, about their MSN user ID, age, country of origin, major, interests, clubs, hobbies, their campus community, in touch with the relatives and friends MSN logo and chat time and other personal information to establish a personal information database. According to the similarity of the user attribute information in the user questionnaire survey and the actual relationships between users provided in prior, we picked up 11 people in the questionnaire with relationships in diversity as experiment database and result guide line shown in the Table 1 as shown in accessories.

5.3. Result Comparison

To validate and compare the TOPSIS-based ranking effective, we sorted our experiments in 3 groups, testing in active degree, Presence strength and Social circle aspects and all indicators above respectively as shown in Tab.1 and Fig 3.

Ex 1. Choose the times and time span of the communication as the principle respectively, the ranking result showed in line "c_times" and "t_span" respectively.

Ex 2. Choose the presence scoring principle, different types of communication such as text, file transfer, voice and video types weighted according to the formula(3) score. And contactors' tightness ranking showed in "presence".

Ex 3. Choose the size and overlap degree of the social circle as the principle respectively, relations ranking numeration showed in "extension" and "overlap".

Ex 4. Ranking edges with all indicators based on TOPSIS method, and the result list showed in "final" curve.

All the results in the 4 groups of experiments were shown in Figure 3 and the accurate values were shown in Table 2 as shown in accessories.

At the contrary to the Figure 1 in which positions are equivalent in the original e(2,6) and e(2,7), but their weighted value had a great difference in the multimedia communications appraisal. Beside the frequency and amount factors in user communication, we added media and social circle factors for a compositive adjudgement. It not only showed difference in ranking sequence from a ranking by a single index, but also more similarly reflect a real relation in tightness.

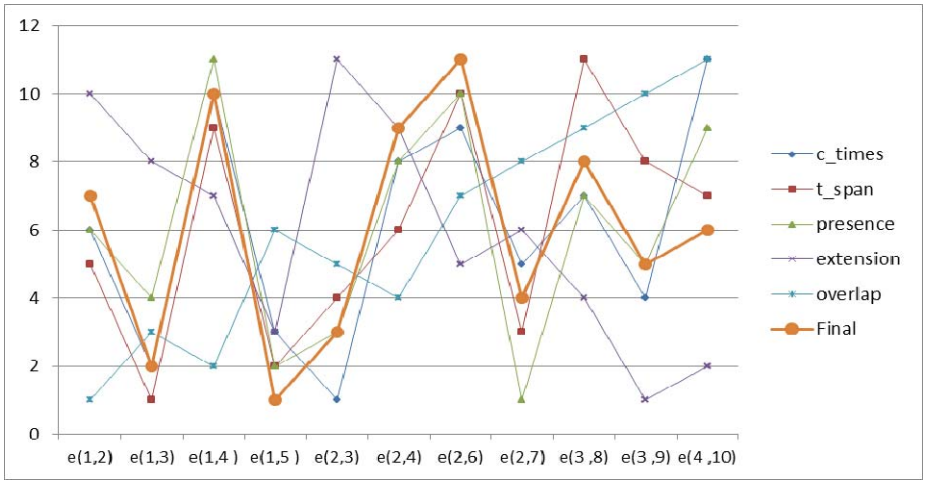

Figure 3 experiment based on Fig.1

6. Conclusions

The contribution of our work is a supplementary to social network in multimedia analysis. It made a basis work for weighted network analysis and a good preparing for advanced research on key nodes detection, community discovery and other many studies in weighted IM network.

References

[1] ABarrat, MBarthelem, A Vespignani. *Modeling the evolution of weighted networks*. Physical Review E, 2004.

[2] Yihjia Tsai, Cheng-ChinLin, Ching-ChangLin. *Characteristics of Weighted Email Communications*. Proceeding of ICICS, 2005: 399-403.

[3] Staniford S, Paxson V, WeaverN. *How to own the Internet in your spare time*. Proceeding of the 11th USENIX Security Symposium, August 2002: 149-167.

[4] JP Onnela, J. Saramaki, J. Hyvonen. Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences, 104(18):7332-7336, 2007.

[5] Zhao Yuan-ping. *Research on Modeling of Internet Topology and Information Propagation in Instant Messaging System*. Beijing University of Posts and Telecommunications.2010.

[6] Guna wardena. *Social Presence Theory and Implications for Interaction and Collaborative Learning in Computer Conferences*. International Journal of Educational Telecommunications, 1995 (2):147-166.

[7] CT Chen. *Extensions of the TOPSIS for group decision-making under fuzzy environment*. Fuzzy sets and systems, 2000 (114): 1-9

Accessories

Table 1 data of surveys used in experiments

Edge(m,n)	*Comm _count*	*Time_ span*	*Social Presence Score*					*Exten-sion*	*Over-lap*	*Remark*
			Text (1)	File (15)	Voice (20)	Pmg (25)	total			
e(1,2)	39	26	29	10	0	0	129	8	2/7	Partner
e(1,3)	136	30	133	0	0	3	208	7	1/6	Bosom-friend
e(1,4)	7	16	7	0	0	0	7	6	1/5	Normal-friends
e(1,5)	73	30	63	6	5	10	473	4	0	inamorato
e(2,3)	154	27	152	0	3	0	212	8	1/7	Bosom-friend
e(2,4)	16	20	16	0	0	0	16	7	1/6	Club-membership
e(2,6)	9	10	9	0	0	0	9	5	0	Normal-friends
e(2,7)	46	25	19	10	17	3	534	5	0	inamorato
e(3,8)	28	9	25	3	0	0	55	4	0	colleague
e(3,9)	66	17	57	9	0	0	147	3	0	colleague
e(4,10)	5	20	4	1	0	0	14	3	0	schoolmate

Table 2 experiment results with Fig.3 in detail

Rank / *Edge*	*Comm_times*	*Time_span*	*Social-Presence*	*Extension*	*Overlap*	*Topsis-Final*
e(1,2)	6	5	6	10	1	7
e(1,3)	2	1	4	8	3	2
e(1,4)	10	9	11	7	2	10
e(1,5)	3	2	2	3	6	1
e(2,3)	1	4	3	11	5	3
e(2,4)	8	6	8	9	4	9
e(2,6)	9	10	10	5	7	11
e(2,7)	5	3	1	6	8	4
e(3,8)	7	11	7	4	9	8
e(3,9)	4	8	5	1	10	5
e(4,10)	11	7	9	2	11	6