Identify Intimate Social Relationship in Blogsphere

Li Han[1,2], SongXiang Cen[2], Jian Ma[2], and Yan Ma[1]

[1] Beijing University of Posts and Telecommunications, China
hanli_bupt@126.com, mayan@bupt.edu.cn
[2] Nokia Research Center Beijing, China
csx008@gmail.com, jian.j.ma@nokia.com

Abstract. Weblog has become a kind of important information resource on the Internet for it can make people publish individual experiences and opinion easily. Blogsphere has evolved vast and complex social network through blogrolls, citation, reading, comments and other social activities. More attentions are paid on related research for the prevalence of blog. The paper proposes the definition of intimate relationship based on comments links to analyze social network in the blogsphere. Blog social network is modeled as weighted directed graph and every weight depends on the frequency of social relationship between bloggers. And then blog intimate friends circle can be extracted with identification algorithm proposed by the paper. Blog data from one of famous blog sites (Hibaidu) in China is collected and taken as our study object. Through analyzing the blog dataset, we found that the intimate relationships are kept more stably and different bloggers maintain intimate relationship with the same blogger but they have no social interactions, which is named as the phenomena of familiar strange friends. These findings will contribute to the understanding of blogsphere and providing better blog services for users.

1 Introduction

Weblog has become a prominent social medium on the Internet. But so far, blog has no recognized definition. According to "Glossary of Internet Terms" [1], basically a blog is a journal that is available on the web. The activity of updating a blog is "blogging" and someone who keeps a blog is a "blogger". Blogs are typically updated daily through software that allows people with little or no technical background to update and maintain the blog. Postings on a blog are almost always arranged in chronological order. Weblog has many particular features compared with regular web pages so specific attentions should be paid on weblog. At the same time, as a kind of popular social medium, Blogsphere gestates gigantic social network through social activities, such as blogrolls, citation, comments, reading and so on. Increasing interests from research and industrial communities about blog harness much research work along with the prevalence of weblog. Ongoing research in the area includes mining blog content [2] [3] [4] [5] and analysis of blog social network. Our work is more related to the latter. Traditional analysis of social network always carried out based on static graph [2]. But there is variability in the social network so subsequent works pay more attention on social network evolution. Now most of the blog research has focused on

blogrolls and citation links [6] [7] [8]. But comments links is also an important social relationship, the focus of this paper will paid on comments links.

The study of intimate relationships is a relatively new area within the field of social psychology. Intimate relationship is a particular close interpersonal relationship, which plays a central role in our overall human experience. According to Miller *et al* [9], the intimate relationship consists of people that we are attracted to, whom we like and love, and those who we marry and provide emotional and personal support. This paper extends intimate relationship to characterize social relationship in the blogsphere, more specifically, the definition of intimate relationship will be proposed based on comments link to weight social relationship among bloggers. We think only when there are the same interest things among bloggers, they will keep intimate relationship. So the research of intimate relationship in the blogsphere will help understanding of blog social network and propose new added services based on findings, such as, personalized blog search engine or recommendation of blog.

The paper is organized as follows. Section 2 reviews related work on blog community and information diffusion. Section 3 gives the definition of intimate relationship and intimate friends circle and proposes corresponding identification algorithm. Section 4 conducts experiment based on dataset from China Hibaidu website. The summarization and the direction of future work are given in the final section.

2   Related Work

Analysis of blog social network can be applied to research of blog community and information diffusion of blogsphere. Kumar *et al*. [8] aimed to experimentally observe and model the highly dynamic and temporal community structure. So they defined time graphs to extend the traditional notion of evolving directed graph, capturing link creation as a point phenomenon in time. Their algorithm for community extraction based on time graph consisted of pruning and expansion. The blog graph would be first scanned for all vertices of degree at most two. Vertices of degree zero and one were removed and vertices of degree two were checked to determine whether their two neighbors were connected. If so they were passed through as a seed in the expansion step and the resulting community was output. The aim of the expansion step was to grow the seed into a set of nodes that constituted a potential community. They analyzed the evolution of structural properties of blog time graph and the results showed that the macroscopic structure and the microscopic structure of the graph began to change dramatically. They also presented analysis of bursty behavior within the blog communities and showed that this burstiness was a fundamental property of link creation in blogspace. Tseng *et al* [10] aimed to capture the community landscape on a specific topic and allow users to explore these important blog communities. When users identified a query of interest and fetched relevant entries, and then the impact scores of relevant entries were calculated. The ranking scores for the blogs were derived from the entry scores. The community structure could be extracted through rank-based connectivity analysis. Given different thresholds of ranking score, different slices of the community structure were given as a set of clusters. Tomographic clustering algorithm generated an ordered sequence of blogs, which could be drawn into a curve called mountain view with blog scores on the vertical axis. Mountain view could

be seen as the contour of the community structure since the curve showed upper bounds of paths within social network. Peaks and valleys of the view depicted representive blogs as community authorities and community connectors respectively. Kazunari [11] proposed the concept of a latent weblog community, which was a meeting space for bloggers who wrote about similar or closely related topics but did not know each other. To extract latent weblog communities, a new method to partition a bipartite graph into subgraph was proposed, which removed the highest frequency edges to go through the shortest paths between the weakest pairs. Because the method tried to divide node pairs which had the weakest relation among all nodes paris, the authors named it as Weakest pair algorithm. Experimental results explained that the weakest pairs algorithm was more effective than shortes path betweenness in extracting latent weblog community from Blogspace in terms of information loss and structure completeness. Backstrom *et al* [12] aimed to resolve the basic questions about evolution of social community based on source data from LiveJournal and DBLP and found that the propensity of individuals to join communities and the development of communities depended in subtle ways on the underlying network structure.

Another problem of information diffusion in the blogsphere is also related to our work. Gruhl *et al* [13] studied the dynamics of information propagation in the blogsphere and presented a characterization of topic propagation from individual to individual respectively at macroscopic and microscopic level. Adar *et al* [14] defined a set of features appropriate for infection inference in the blogsphere and constructed infection tree to visualize the information flow.

Our works analyze blog social network and identify intimate friendship from Blogsphere. We extend intimate friendship in the social psychology and firstly give the definition of intimate relationship for the blogsphere. Based on the definition, intimate relationships are extracted by the weight of social interaction among blogger. Previous research paid more attention on the number of neighbours in the blogsphere when finding blog communities. But we take not only the number of social relationship but also the intensity and frequency of social interaction between blogers into consideration. Additionally, as a kind of social medium from Internet, Blog has specific regional for different culture. The paper utilizes real blog dataset from China website HiBaidu as research object. The findings are more suitable for the status of China Blog than research works based on other blog datasets.

3   Model Blog Social Network

Blogsphere can be described by blog graph, a directed graph $B(V, E)$, where V is the set of all the blogger and a directed edge $(v_x, v_y) \in E$ suggests there is social relationship between blog v_x and blog v_y. The social relationship can be citation, comments, blogroll and trackback but the paper will pay more attention on comments relationship. So there will be a directed edge from v_x to v_y when blog v_x give
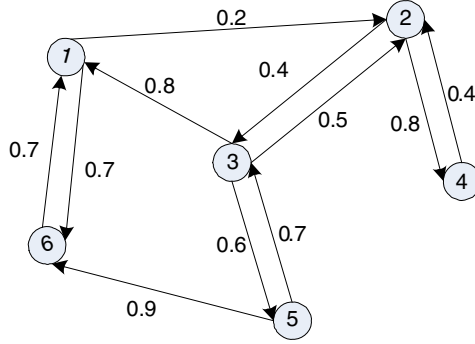
Fig. 1. Weighted Directed Blog Graph

comments to any post of blog v_y during studied time, where v_y is called the head and v_x is called the tail. But in the blog graph, v_y can be named as blogger and v_x is named as reviewer. In order to identify intimate friendship in the blogsphere, directed blog graph is extended to weighted directed blog graph. The label or weight of directed edge from v_x to v_y denotes the percentage of comments gived by v_x to blog v_y in the total number of posts of blog v_y.

Next, we will discuss what can be named as intimate relationship in the blogsphere based on weighted directed blog graph. The degree of intimate relationship between bloggers depends on the weight. There will be threshold value set, and if the weight of a directed edge is above the threshold value, it can be thought that there is intimate relationship between the blogger and reviewer. If the threshold value is set to 0.6, there are two bidirectional intimate relationships between blog 1 and 6, blog 3 and 5 as Figure 2. But the relationships from blog 3 to 1, from blog 5 to 6, and from blog 2 to 4 can not be thought as intimate relationship for its undirectivity.

3.1   Definition of Circle of Intimate Relationship

Let $V = \{v_1, v_2, ... v_x, ... v_n\}$ be the set of blog node, N be the set of neighbour blog nodes, such as $N(v_x) = \{v_a, v_b, ... v_i, ... v_z \mid (v_x, v_i) \in E, or, (v_i, v_x) \in E\}$ be the neighbour set of blog node v_x. The blog dataset $D = R \cup W = \{r_1, w_1, r_2, w_2, ... r_m, w_m\}$ is a blog relationship dataset of blogger reviewing relationship itemset R and weight itemset W. Every item pair (r_i, w_i) be $(r_i, w_i) = \{v_x, v_y, w_{xy}(v_x, v_y)\}$, $w_{xy}(v_x, v_y)$ means the first blogger v_x give comments to the second blogger v_y and the number of comments is w_{xy} percent of

the number of posts published by the second blogger . Every $r_i = (v_x, v_y)$ is ordered pair and has the unique corresponding value of w_i .

Definition 1 (Intimate Relationship)

There is intimate relationship in the $p = (v_x, v_y) \in R$ if $w_{xy}(v_x, v_y) > \sigma$ and $w_{yx}(v_y, v_x) > \sigma$, where σ is the threshold value. And $p = (v_x, v_y)$ can be named as intimate relationship pair, which is not ordered pair. So the intimate relationship is bidirectional.

According to social psychology, intimate relationship comes into being gradually so there is a kind of social relationship, which can not be divided into intimate relationship but will be likely to become intimate relationship in the future. We name this relationship as promising intimate relationship and give its definition in the definition 2. Decay factor α is utilized to reduce criteria for classifying.

Definition 2 (Promising Intimate Relationship). There is promising intimate relationship in the $p = (v_x, v_y) \in R$ if $w_{xy}(v_x, v_y) > \alpha\sigma$ or $w_{yx}(v_x, v_y) > \alpha\sigma$. So promising intimate relationship is directive.

Definition 3 (Intimate Friends Circle). Intimate friends circle is made up of center and edge. $q = \{v_{mx}, v_{my}, ... v_{mz}\}$ is the center of intimate friends circle, when $\forall v_{ma}, v_{mb}$, v_{ma} and v_{mb} must be intimate relationship pair. The edge of circle $NE(q) = \{v'_{ux}, v'_{uy}, ... v'_{uz}\}$, $\forall v'_{ux}, \exists v_{my} \in q$, there must be promising   intimate relationship between v_{ux}' and v_{my}. So $q \cup NE(q)$ constitutes intimate friends circle, and $|q|$ is the size of intimate friend circle.

Definition 4 (Intimate Friends Circle Set)

$Q_{ifc} = \{Q_1, Q_2, ... Q_m\}$ is the intimate friends circle set for the blogsphere, every item Q_i is $q_i \cup NE(q_i)$, q_i is the center of intimate friends circle and $NE(q_i)$ is the edge of circle.

3.2   Identify Intimate Friends Circle Set

The definition of intimate friends circle set has been given above. How to identify intimate friends circle set from the vast and complex blog social network, the corresponding algorithm is proposed in Table 1. According to the above definition, every intimate friends circle is made up of center and edge. There may be no edge under certain value of threshold and decay factor but circle must have a center. For blogsphere, the quantity of intimate friends circle depends on the number of center.

Table 1. Algorithm to identify intimate friends circle set

Algorithm Identify Intimate Friends Circle Set

Input : Blog relationship dataset D
       Threshold σ Decay factor α
Output: Intimate friends Circle Set Q_{ifc}
1: Initialize 0 intimate friendship pair set P
2: For each $(r_i, w_i) = (v_x, v_y, w_{xy}(v_x, v_y)) \in D$
3:    if $w_{xy}(v_x, v_y) > \sigma$
4:      Find $(r_i^{'}, w_i^{'}) = (v_y, v_x, w_{yx}(v_y, v_x))$
5:        if $w_{yx}(v_y, v_x) > \sigma$
6:          add (v_x, v_y) to P
7: end
8: Initialize $j = 1$
9: for $p_i = (v_x, v_y) \in P$
10:    add v_x, v_y to q_j
11:    if $\exists v_z$ satisfy $(v_x, v_z) \in P \,\&\, (v_z, v_y) \in P$
12:      add v_z to q_j
13:    for each $v_{x^{'}} \in q_j$
14:      if $\exists(r_i, w_i) = (v_{x^{'}}, v_a, w(v_{x^{'}}, v_a)) \,\&\, w(v_{x^{'}}, v_a) > \alpha\sigma$
15:        add v_a to $NE(q_j)$
16:      if $\exists(r_i^{'}, w_i^{'}) = (v_a, v_{x^{'}}, w^{'}(v_a, v_{x^{'}})) \,\&\, w^{'}(v_a, v_{x^{'}}) > \alpha\sigma$
17:        add v_a to $NE(q_j)$
18: $j = j + 1$
19:    add $q_j \cup NE(q_j)$ to Q_{ifc}
20: end

4 Experiment

To analyze intimate relationship in the blogsphere, we crawled blog data from one of the famous blog website Hibaidu (http://hi.baidu.com) as our research object and collected a set of 3484 blogs containing 536351 posts until sep 11th 2008. Baidu space is a Chinese blog publication platform whose web pages are well organized with unified templates. This enables us to extract the posts, comments and other information such as published date simply by regular expressions. Because we pay more attention on the comments relationship in the blogsphere, we extract corresponding reviewing relationship from every post and construct blog social relationship dataset D as the description in the section 3. Every item includes two blog nodes (A, B), which means that blog A give comments to blogger B during studied time, and there will be a directed line from A to B.
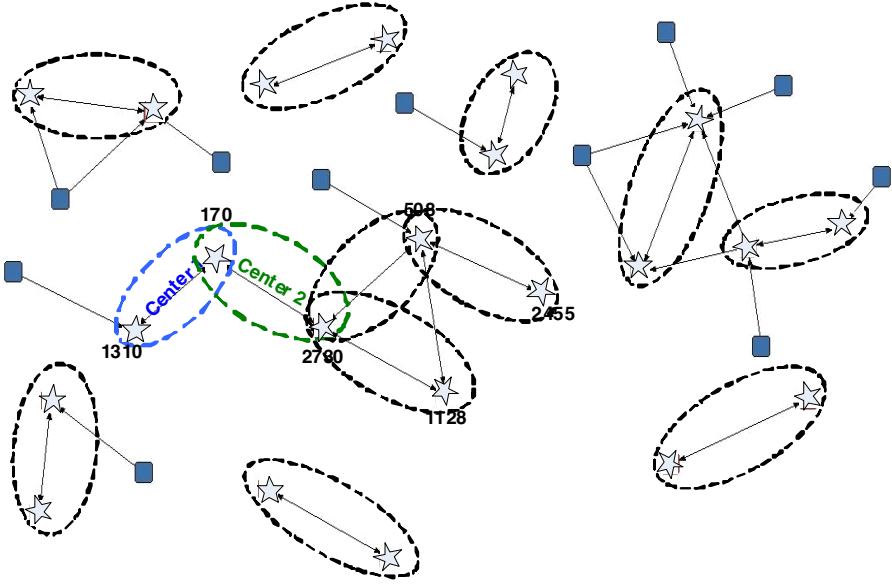
Fig. 2. Intimate Friends Circle Set

Based on the blog social network, we extract intimate friends circle from Hibaidu blog dataset with our algorithm under $\sigma = 0.6$ and $\alpha = 0.5$. The final results are given by Figure 2. In order to describe our findings, every blog is given to a unique serial number in the following parts. It is easily concluded that there is few indeed intimate relationship in the blogsphere. A blog can belong to two or three intimate friends circle concurrently but no so many social activities exist among his initimate friends within different communities, for example blog 170 and 1310 form the center of circle 1, blog node 170 and 2730 form center 2 but there is no intimate relationship between node 1310 and 2730. Similar situation also exists between 170 and 1128, 170 and 508, 2730 and 2455. We name the phenomenon as familiar strange friends. But there is a pair of blog node 1128 and 508, which form different intimate friend circle with node 2730, but have more social interaction. In order to study the phenomena of familiar strange friends further, we pay more attention on the social relationship between node pair 1128 and 508. Here we adopt two indicators $SR(B,R)$ and $DP(B,R)$ to weight the social relationship between blog nodes as formula (1) and (2).

$$SR(B,R) = \frac{number\_comment\_(R\_to\_B)}{number\_post(B)} \tag{1}$$

$$DP(B,R) = \frac{number\_comment\_(R\_to\_B)}{number\_comment(R)} \qquad (2)$$
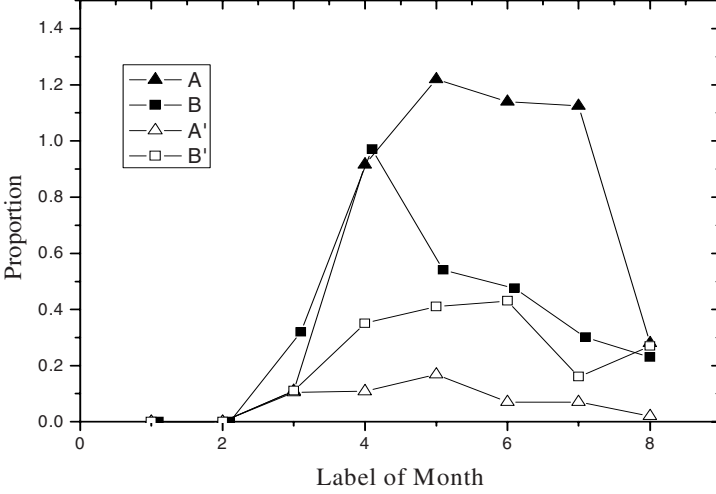
Fig. 3. Social Relationship between Blog Node 1128 and 508

$number\_post(B)$ denotes the number of post of blog B during studied time, $number\_comment(R)$ denotes the number of comments of reviewer R and $number\_comment\_(R\_to\_B)$ is the number of comments from reviewer R to blog B. Line A and $A^{'}$ denote $SR(1128,508)$ and $DP(1128,508)$ respectively, and line B and B' is $SR(508,1128)$ and $DP(508,1128)$ in Figure 3.

It can be found that the social relationship between 508 and 1128 will be kept once the relationship has been established. But it is found that there is changing tendency in the relationship during May to July 2008. In order to uncover the reason, we study the activity of blog node 1128 and 508 in the blogsphere. For every blog node, the degree of activity is measured according to the number of post and comments from blogger as formula (3).

$$DA(B) = number\_post(B) \times number\_comment(B) \qquad (3)$$

Figure 4 gives the degree of activity for blog node 1128 and 508 from January to August. We analyze the social interaction between node 1128 and 508 by combining Figure 3 with 4. Node 508 keeps his attention about node 1128 since he gives comments to node 1128. The attention has decreasing tendency on August but the degree
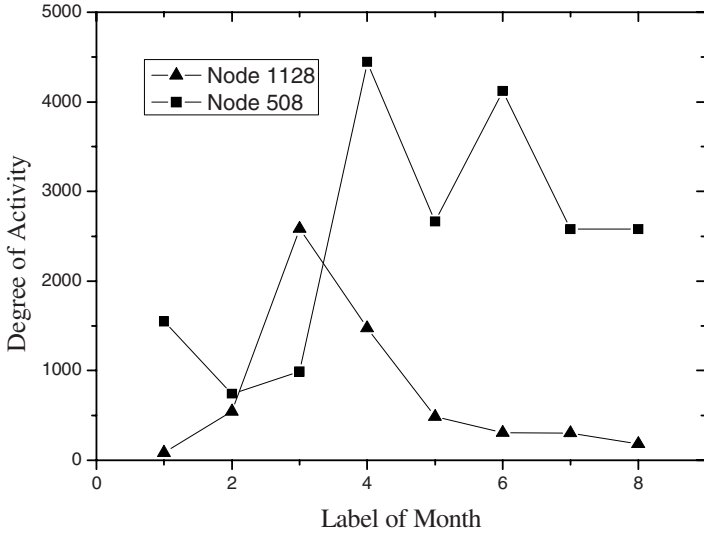
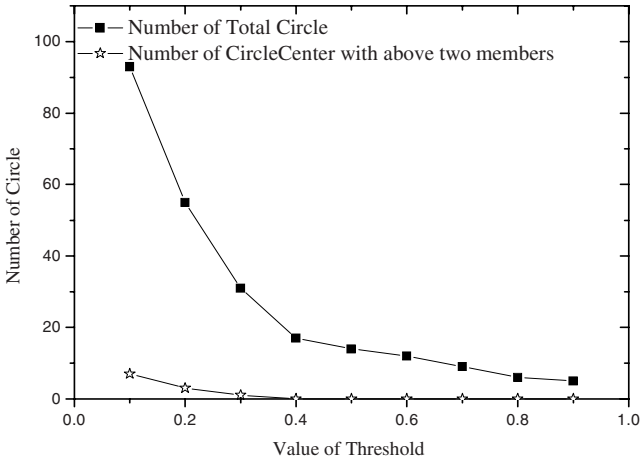**Fig. 4.** Degree of Activity of blog node 1128 and 508

Fig. 5. Number of Circle in the Blogsphere

of activity of node 508 also drops. Node 1128 pays less attention on node 508 from May to August at the same time the active degree of node 1128 has the similar tendency during the same time. Based on the observed result, it is believed that there should be more social relationship between these familiar strange friends. We will provide recommendation among these familiar strange friends found and follow the status of relationship to verify the conjecture in the future research work.
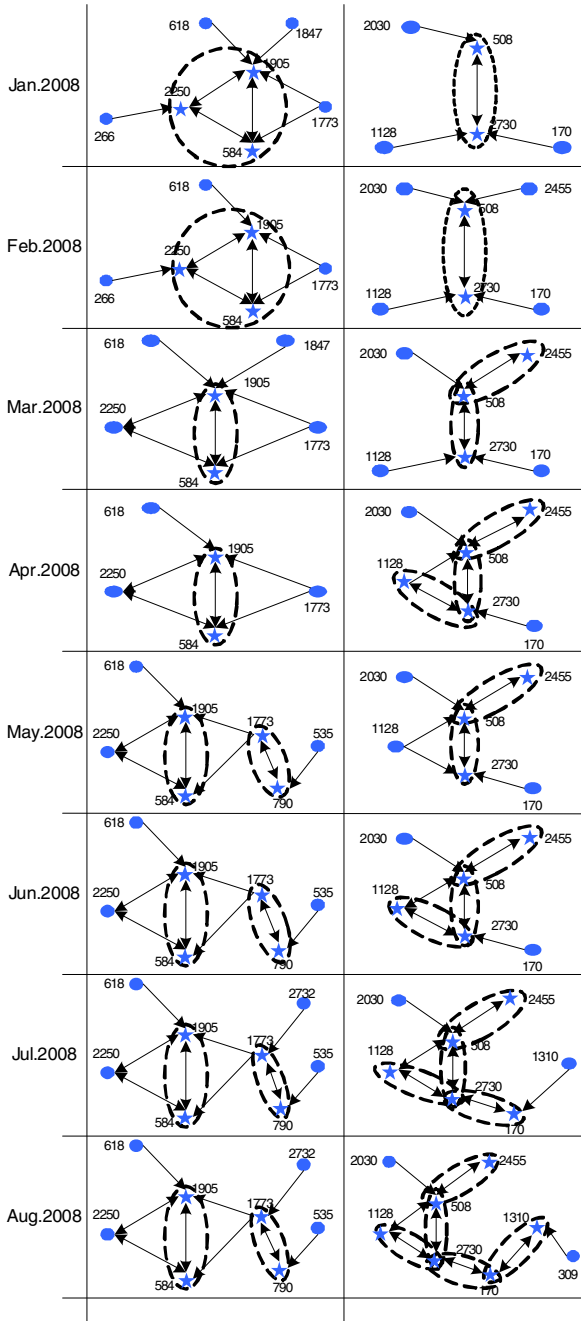
Fig. 6. Evolution of Intimate Friends Circle in Blogsphere

According to the above definition, every circle has unique center, which depends on the value of threshold, so we study the number of circle when changing threshold value. From the results of Figure 5, the simple interaction social relationship is common so the number of circle will drop significantly when value of threshold change from 0.1 to 0.4. It is also concluded that the number of circle center with above two members is very few.

In order to study the process of intimate friends circle, we extract two intimate friend circle groups, which will include more than one circle, to visualize the formation of circle set as Figure 6. The value of threshold is 0.6 and decay factor is 0.5. The intimate relationship in the center of intimate friends circle can be kept stably than that of circle edge, so the basic intimate relationship in the intimate friends circle has the slow evolution.

5   Conclusions

The paper pays more effort on analyzing the social relationship among bloggers in the blogsphere. The intimate relationship depends on frequency of interaction between blogger. Through studying the blog dataset from Hibaidu, we found the constancy of intimate relationship and the phenomena of familiar strange friendship. The finding of phenomena will help to characterize social network in the blogsphere. At the same time, individual blog recommendation service can be provided for users through identifying intimate relationship and the results of mining content of blog.

References

1. Matisse's Glossary of Inernet Terms,
   http://www.matisse.net/files/glossary.html
2. Bar-llan, J.: An Outsider's View on Topic-oriented Bloggiing. In: Proceeding of the Alt.Papers Track of the 13th International Conference on World Wide Web, paper 28-34 (May 2004)
3. Mishhe, G.: Experiments with Mood Classification in Blog Posts. In: Style 2005-1st Workshop on Stylistic Analysis of Text for Information Access, at SIGIR 2005 (2005)
4. Mei, Q., Liu, C., Su, H., Zhai, C.: A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs. In: Proceedings of 15th International Conference on World Wide Web (2006)
5. Ni, X., Xue, G.-R., Ling, X., Yu, Y., Yang, Q.: Exploring in the Weblog Space by Detecting Informative and Affective Articles
6. Adamis, L., Glance, N.: The political blogosphere and the 2004 U.S. election: divided they blog. In: LinkKDD 2005 Conference, Chicago, Illinois (2005)
7. Adar, E., Adamic, L.: Tracking infromation epidemics in blogsphere. In: Web Intelligence 2005, Compiegne, France (2005)
8. Kumar, R., Novak, J., Raghavan, P., Tomkins, A.: On the bursty evolution of blogspace. In: Proceedings of 12th International Conference on World Wide Web (WWW), pp. 568–576 (2003)
9. Miller, R.S., Perlman, D., Brehm, S.S.: Intimate Relationship. McGraw-Hill, Toronto

10. Tseng, B.L., Tatemura, J., Wu, Y.: Tomographic clustering to visualize blog communities as mountain views. In: Proceedings of 2nd Annual Workshop on the Weblogging Ecosystem (2005)
11. Ishida, K.: Extracting latent weblog communities: A partitioning algorithm for bipartite graphs. In: Proceedings of 2nd Annual Workshop on the Weblogging Ecosystem (2005)
12. Backstrom, L., et al.: Group formation in large social networks: membership, growth and evolution. In: Proceedings of 12th ACM SIGKDD, pp. 44–54. ACM Press, New York
13. Gruhl, D., Guha, R.V., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogsphere. SIGKDD Explorations 6(2), 43–52 (2004)
14. Adar, E., Adamic, L.A.: Tracking information epidemics in blogspace. In: Web Intelligence (2005)