Extraction and Visualization of Implicit Social Relations on Social Networking Services

Meesun Song Wonkyu Lee Junghwan Kim

Media Solution Center
Samsung Electronics, Suwon-Si, Gyeonggi-Do, Korea
{ms45.song, wkyu.lee, junghwani.kim}@samsung.com

Abstract

Today social network services like blogs, communities, and social networking sites dominate the web. As Web 2.0 has evolved this way, analyzing social networks has become a promising research issue. There have already been several researches on social network analysis based on users' activities in social services. Most of them focus on the links among the users such as citation, trackback, and comment. However, few studies have analyzed the relations within message threads. In general, they considered the one-way relationship from a comment writer to a post writer. Since users communicate with each other primarily by posting comments one after another, the message threads are key to analyzing latent social relationships. In this paper, we propose a novel method to extract the social relations hidden behind message threads. To evaluate our algorithms, we developed an evaluation system and measured the performances. In addition, since the typical node-edge diagram for social network visualization is not intuitive or readable, we also introduce a novel visualization and interaction method suitable for social relation exploration. Further, we expect our work will help enhance social recommendations, advertisements and personalization.

Introduction

Social media and social networking services such as personal blogs, online communities, Facebook, Twitter, and wiki are now the mainstream services of Web 2.0. As the popularity of social networking has dramatically increased, discovering knowledge from these social services has become a challenging and primary research issue. There have been several researches on social network analysis and two main approaches have been studied: one is influential user discovery (Shin, Xu, and Kim 2008; Nauerz and Groh 2007; Java et al. 2006; Agarwal et al. 2008; Java et al. 2006) and the other is social network construction (Furukawa, Matsue, and Ohmukai 2007; Karamon, Matsuo, and Ishizuka 2008; Lin et al. 2006; Nauerz and Groh 2007). The former approach concentrates on finding the most influential users in communities or blogospheres by analyzing their connections and activities, whereas methods for discovering a social network of users are studied in the latter approach.

Although studies are classified into two approaches, most studies primarily examined the links among users' activities, such as citation, trackback, and comment, to extract the social relation. In addition to the explicit links such as citation or trackback, the one-way relation from a comment writer to a post writer is assumed. However, often users make arguments or comments within a comment thread and the main post writer may not be included in the context of the current comment thread. In other words, there are complex relationships between the comment writers and the post writer or among the comment writers, which are not limited to a one-way connection. In order to alleviate this problem, we focused on the latent relationship within the message thread.

The primary communication method of social networking services is posts. By writing posts and comments, people share their thoughts, opinions, and real time status. People tend to write messages and replies to people they have close friendships with or are arguing with. Also people write lots of comments one after another when they are having an in-depth conversation. Based on this tendency, we made a model on the latent social relationship among users by examining and analyzing the message threads. In this paper, we propose a method to extract the latent social relationship from a social networking service by analyzing the users' activities. The users' writing patterns are especially considered to examine the intensity of the conversation and the strength of the social relation.

We applied our algorithms to a Facebook dataset and developed an evaluation system to appraise the proposed algorithms. The experimental result shows that the proposed method using a weighted harmonic rule with a root-included sliding window fits best for social relation extraction.

In addition, we introduce visualization and various interaction methods for exploration and exploitation of the social relationship. Most of the previous works presented the extracted social networks as a node-edge graph (Shin, Xu, and Kim 2008; Nauerz and Groh 2007) or didn't consider any visualization method. Since lots of nodes and edges make the graph too complicated, it is often difficult to understand the social network in a node-edge graph. We suggest a dynamic radial graph to visualize the eco-centric social network data. Using an animated transition and a distortion technique, we were able to achieve convenient graph exploration.

This paper is organized as follows. The related works are discussed in the next section. We then explain our extraction and visualization methods for social relations in detail. Finally we describe the evaluation method and performance results, followed by conclusions.

Related Works

There have been several works about the influential user discovery. Shin et al. (Shin, Xu, and Kim 2008) defined new features for describing users' social activities and showed that the Cross Reference (CR) feature, which measures a user's popularity by the amount of exchanged comments works best in discovering power users. They used the harmonic mean of the number of comments from user A to B and the number of comments vice versa to calculate the popularity. Their work also suggested a novel interface for effective exploration of power users based on CR rank. However, they considered only the one-way relation from the comment writer to the post writer, and the proposed visualization method was still based on a node-edge graph. Agarwal et al. (Agarwal et al. 2008) used the number of in-links, the number of comments, the number of out-links, and the length of a post to identify the influential bloggers.

Several other researches focused on the social network construction. Furukawa et al. (Furukawa, Matsue, and Ohmukai 2007) analyzed social networks in terms of four metrics: citation, comment, trackback, and blogroll. In addition, readership relations derived from the user log data was analyzed and characterized. They also predicted the readership relation from the four kinds of social networks. Karamon et al. (Karamon, Matsuo, and Ishizuka 2008) proposed an algorithm to systematically identify important network-base features, such as graphical distance, common neighbors, the number of links and structure equivalence, to analyze user behavior efficiently. Lin et al. (Lin et al. 2006) focused on community extraction. They suggested the mutual awareness feature with action type, frequency and time of occurrence to discover the communities. Nauerz and Groh (Nauerz and Groh 2007) derived the social network and the expert users using web usage mining, tagging behavior analysis and explicit social network.

These previous works primarily used the number of links such as citation, trackback, comment and blogroll to analyze social network. However, few of them have examined the latent relations presented in message threads.

Extraction of Social Relations

Base model: Frequent Set Mining Algorithm

There are two forms of message board as shown in Figure 1. When a post is written, comments are appended to the corresponding post. Type (a) ignores the hierarchical relation among comments, whereas (b) shows the hierarchical relation.

When you consider the message thread shown in Figure 1, there may be a relation between user 2 and user 3. However, the previous researches only considered the relation between user 2 or user 3 and user 1. In general, it is very common that the arguments or discussions occur within a message thread by writing comments. There could be relations among comment writers in addition to the relationship with the post writer. Therefore, discovering the hidden or latent relation in a message thread is an important metric for capturing social relations.

It is relatively easy to detect the relation between user 2 and user 3 in a form (b) since the relations between comments are explicitly defined. However, even in this case users sometimes do not give comments directly to the reference message or comment. Furthermore, a lot of social networking sites only offer a message board of type (a).

In order to extract these hidden relationships, we took an approach from the frequent set mining techniques (Liu 2006). The typical application of frequent set mining is the market basket analysis. The algorithm extracts frequent item sets from shoppers' purchase lists. For example, when a frequent item set, milk and bread, is discovered, we can conclude that milk and bread are frequently bought together and those items are closely related. Frequent set mining is also used in discovering users with similar behavior patterns by analyzing co-occurring web pages and portlets (Nauerz and Groh 2007). We took this approach to discover sets of users that appeared frequently together by examining users' writing patterns. Similar to the market basket analysis for detecting the relationship among purchases, we applied Apriori, an association rule mining algorithm (Agrawal and Srikant 1993), to extract the relationship among the users within message threads.

A message thread is considered as a transaction in an association rule mining algorithm with duplicated users being removed, so that user sets that frequently appear together in those transactions are discovered and the frequency is used as a social relation score for the user set. This extracted user set implies users who have a close relation since they share similar interests or they communicate frequently.

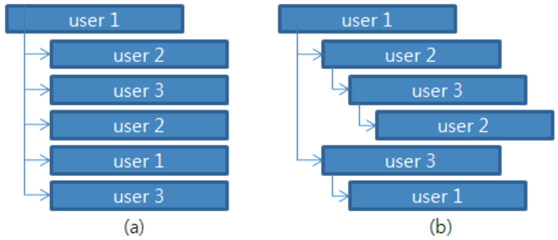

Figure 1: Two forms of a message thread

Weighted Harmonic Rule (WHR)

The limitation of the basic frequent set mining algorithm is that it only considers the existence of users in a transaction, so it does not capture the intensity of the conversation.

The relation between user 1 and user 2 can be inferred from both message thread (a) and (b) in Figure 2. In addition, user 1 and user 2 shared a conversation more intensively in case (a) than (b). However, the frequent set mining approach infers the same relation score. We need to take the intensity of a conversation into consideration to more precisely predict the social relation.

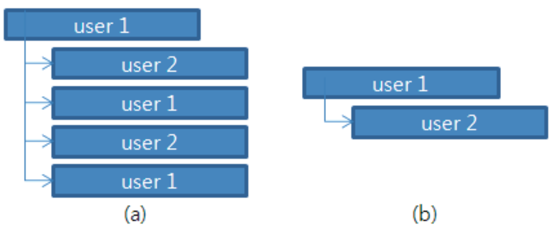

Figure 2: Examples that show the difference of the conversation intensity

As mentioned earlier, people usually have a conversation or an argument using comments. If two users appear in one message thread several times, the probability that the two users are having a conversation is high. To capture this intensity of a conversation in a message thread, a user's occurrence in a message thread is used as the weight for each user. If a user group has an equally high weight, then they are probably having a conversation with each other. To extract the user groups whose occurrences are equally high and appear frequently together, we added weight to the basic frequent set mining algorithm and the harmonic mean was applied to the weight of each user. Because of the property of the harmonic mean, we discovered user sets in which each user had an equally high weight.

We first defined a transaction as a set of users who appear in same message thread as follows:

$$\mathrm{T}_1 = \{u_1, u_2 \ \ldots\}, \mathrm{T}_2, \ldots, \mathrm{T}_\mathrm{n}$$
$$\mathrm{W}_1 = \{\mathrm{W}(u_\mathrm{j}) | \forall u_\mathrm{j} \in \mathrm{T}_1\}, \mathrm{W}_2, \ldots, \mathrm{W}_\mathrm{n} \ ,$$

where $\mathrm{T_i}$ is a set of users who wrote a post or comments in i^{th} message thread and $\mathrm{W_i}$ is a set of $\mathrm{W}(u_\mathrm{j})$, which is the number of occurrences of user u_j in i^{th} message thread, for all $u_\mathrm{j} \in \mathrm{T_i}$. For example, consider a message thread in which user u_1 wrote the main posting and users u_2 and u_3 wrote the comments and user u_1 replied to users u_2 and u_3. In this message thread, the weight of user u_1 is 3 and that of users u_2 and u_3 is 1, 1 respectively.

To evaluate the social relation score of users, there are two steps. First, calculate the relation score within each transaction and then calculate the overall score among all the transactions.

To calculate the relation score of each transaction, we apply the harmonic mean to the weight of each user. For instance, in order to evaluate the relation score between users u_1 and u_2 in $\mathrm{T_i}$, the harmonic mean of two users' weights in $\mathrm{T_i}$ message thread is calculated.

$$\mathrm{Score}_{u_1 u_2}(\mathrm{T_i}) = \mathrm{H}(\mathrm{W}(u_1), \mathrm{W}(u_2)) = \frac{2 * \mathrm{W}(u_1) * \mathrm{W}(u_2)}{\mathrm{W}(u_1) + \mathrm{W}(u_2)} \ ,$$

where $\mathrm{H}(\cdot)$ is a harmonic mean function.

If we apply this equation to the examples in Figure 2, the relation score of user 1 and user 2 is (2*3*2)/(3+2), that is 2.4 in case (a) and the relation score is 1 in case (b). On the other hand, the basic frequent set mining algorithm produces a score of 1 as the relation score in both cases. We can see that the harmonic mean reflects the intensity of the conversation.

Once the relation score in each transaction is calculated, the total relation score that covers all the transactions is evaluated as follows.

$$\mathrm{TotalScore}_{u_1 u_2} = \mathrm{H}\left(\mathrm{Score}_{u_1 u_2}(\mathrm{T}_1), \ldots, \mathrm{Score}_{u_1 u_2}(\mathrm{T_m})\right) * \mathrm{m} = \frac{\mathrm{m}}{\sum_{\mathrm{i}=1}^{\mathrm{m}} \frac{1}{\mathrm{Score}_{u_1 u_2}(\mathrm{T_i})}} * \mathrm{m} \ ,$$

where m is the number of co-occurring transactions. The total relation score among them is also estimated by the harmonic mean over the relation scores of all the co-occurring message threads and then it is multiplied by the number of message threads in which the users appeared together. Since the harmonic mean value only captures the average intensity in message threads, the number of co-occurring message threads is multiplied. This would discover the users who had intensive conversations in many message threads.

Transaction division methods

There are various ways to determine a transaction when applying the weighted harmonic rule mining. The following describes the three methods that are used to determine a transaction.

Normal As described above, basically one message thread is one transaction. The transaction includes the user who wrote a main post and users who wrote comments to the post.

Sliding Window When the interval of the posting time between the comments is short, the corresponding users may be more closely related to each other. Thus, we applied the sliding window to the proposed model. We broke the thread into small threads with the size of the sliding window and then applied the weighted harmonic rule mining. (b) in Figure 3 shows how the sliding window divides a message thread into several transactions. We expect that this would find user groups that appear frequently together and have conversations sequentially.

Root-included sliding window The sliding window focuses on the sub group regardless of the main post writer. Even though the distance between the comments and the main post is far, it is more likely that the comment is a message related to the main post. In addition to the sliding window, we added the main post writer to each of the divided small threads to infer the relations of the sub group and the main post writer. The decomposed message thread using this method is shown as (c) in Figure 3. The total relation score is calculated with the same equation as the above model.

Visualization of Social Relations

To visualize the extracted social relations, a force directed graph and radial layout are commonly used. However, the forced-directed graph is limited in visualizing the distance between the nodes. The radial layout presents the relation between the main node and all the other nodes, but it does not show the relation of neighbor nodes. Since the extracted social relations need to be visualized in easily understandable and interactive form, we propose a dynamic radial graph to cope with the limitations of previous visualization techniques.

(a) in Figure 4 show the dynamic radial graph with the extracted social relationship from a Facebook dataset. The basic layout is similar to the radial layout. The main user is mapped to the red circle, while users who have a friend relation with the main user in Facebook are mapped to yellow and users who have no friend relation are mapped to white. If the node is placed closer to the center, the user has a closer relationship with the main user. The neighbors placed based on their relative relationship. This enhances the readability of the social network.

We applied a distortion technique to solve the occlusions between the nodes. (b) in Figure 4 shows the effect of the distortion technique. Although the three nodes that are close to the center are hardly recognizable, the nodes are clearly distinguishable after applying the distortion.

When the main user is changed, the transition of a graph occurs smoothly and the context is preserved. (c) describes an animated transition. Since the context is maintained, tracking the visual item is easy even when the nodes are moved.

A user can manipulate the graph in the following three ways: by changing the time to change the dataset according to a selected duration; by changing the filter value for the threshold, which is the minimum relation score for a user to appear in the graph; and by changing the distortion value with the slider button.

Evaluation

To determine whether our algorithm fits best for social relation extraction, we evaluated the weighted harmonic rule mining and the transaction division methods in the following process.

Data Set

The data set used in our experiment was the Facebook status data from Jan. 01, 2008 to Aug. 31, 2009 retrieved from a volunteer's Facebook account. The data includes her own status updates and comments and all of her friends' status updates from their wall which is a user's profile page and their comments.

The number of total postings was 1783 and the number of comments was 2575. The total number of users involved in this dataset was 281.

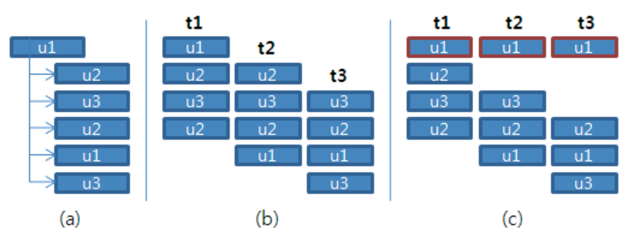

Figure 3: Transaction division methods. (b) and (c) show how the message thread (a) is decomposed using a sliding window of size 4 and a root-included sliding window, respectively.

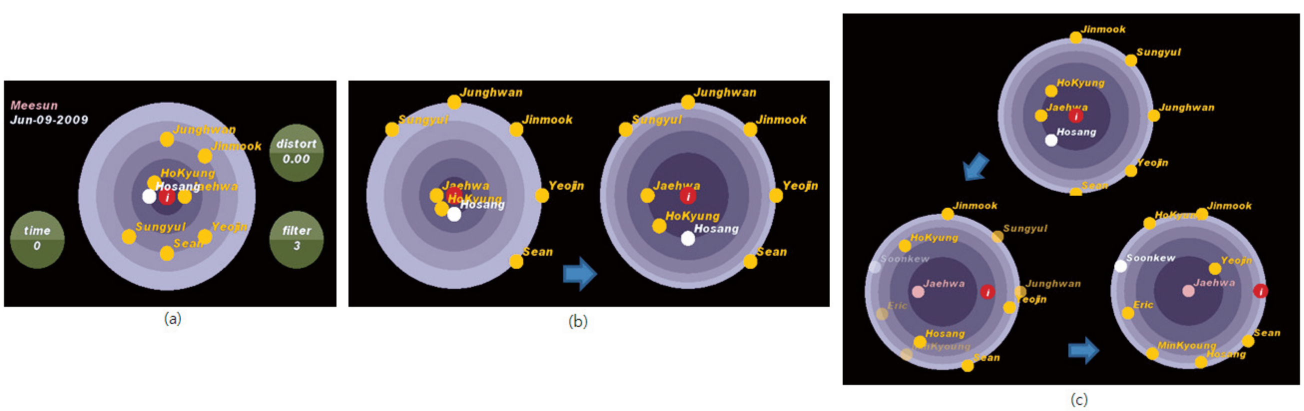

Figure 4: Visualization and interaction of the social relation. (a) represents the dynamic radial graph. (b) shows the distortion technique to avoid collision and (c) shows animated transition when changing the main user.

Evaluation System

To discover the implicit relationship hidden in message threads, people were needed to actually look into the posts and comments and understand their meaning. Therefore, we developed a web site that helped testers to evaluate the relationship easily, and we asked them to evaluate users' relationships based on the posts and comments of the users. These collected results were used as the ground truth in our evaluation.

On the main page of the site, all the posts and comments during a randomly-selected week and a randomly-selected main user were displayed. Then testers were asked to discover the top five users closest to the main user based on the messages they shared in a decreasing order of closeness. To avoid a sparse evaluation result, we limited the range of the selected duration to five weeks in which the number of posted messages was high. Testers were required to understand the context of the messages and find the users who had close relations to the given main user.

We ran the evaluation in two phases. At first, we provided a randomly-selected dataset and a randomly-selected main user. In this phase, 13 testers participated and 58 test cases were evaluated. Next, we selected the nine test cases that the most testers evaluated, and then asked the testers to evaluate the selected test cases again.

Results

About eight evaluation results were collected for each of the nine test cases. To merge evaluation results collected from different testers, we ordered the results by the count the testers agreed on and also kept the original order that testers specified.

The four methods, frequent set mining, weighted harmonic rule mining (WHR) with normal transaction division method, WHR with sliding window of size 3, and WHR with a root-included sliding window were evaluated. The sliding window size was set to 3 based on the performance of each window size as shown in Figure 5. We executed the four methods and compared the results with the user-evaluated result.

We first calculated the Precision, Recall, and F-measure scores which are the widely used measures in information retrievals and classification tasks. We extracted the top five users with the highest relation score from each proposed method and then calculated the scores based on the top five user- evaluated results. Table 1 shows that WHR with the root-included sliding window method showed higher performance than the other methods.

Although the root method performed better generally, the precision and recall scores were not high enough to conclude that it performs best in extracting the hidden relationships. Thus, we executed another evaluation. In social relation extraction, the ordering of the results and the results that the most testers selected are more important. Accordingly, we computed the matching percentages by first giving the highest weight to people whom the most

Methods	Precision	Recall	F-Measure
Frequency	0.62	0.58	0.59
WHR(Normal)	0.70	0.68	0.67
WHR(Window)	0.68	0.66	0.65
WHR(Root)	**0.72**	**0.70**	**0.69**

Table 1: Performances of the proposed algorithms

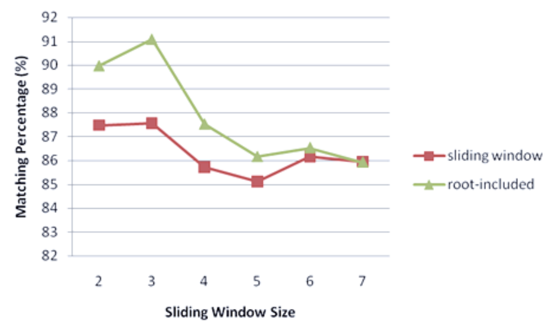

Figure 5: Matching percentages between the algorithms' derived results and the user-evaluated results based on various sliding window sizes

testers agreed on and then by giving weights based on their diminishing order of acquaintance. For example, the result that matched the top two users was better than the result that matched the bottom two users. The results are summarized in Figure 6. Graph (a) shows the results for each test case and (b) shows the averaged matching percentages of all the test cases. The WHR performed better than the frequent set mining. Further, the WHR with the root-included sliding window performed best among the four algorithms; The matching percentage was over 90%. The precision and recall value was not quite notable but the matching percentage was outstanding. This implies that results from the root method matched the order of the users that the most testers agreed on.

Special cases like test cases 5 and 8 in Figure 6 show relatively low performances compared to the other test cases. This is because either the given main user of those cases was a very active person who wrote comments to all of his friends or a new-comer who rarely posted. Thus, the results of the testers did not match well and the result of the algorithm also didn't match well. However, the root method was still more acceptable, even for these extreme cases.

We also tested the arithmetic mean instead of the harmonic mean for the WHR with the root-included sliding window. We applied that average equation in two cases. One was to calculate the relation score within a transaction. The other was to calculate the score among all the transactions. We first substituted the latter harmonic mean to the arithmetic mean and then substituted both. Apparently using the harmonic mean results in a better performance, as Figure 7 shows.

As described above, the weighted harmonic rule mining with the root-included sliding window fits best for the extraction of implicit social relations in message threads.

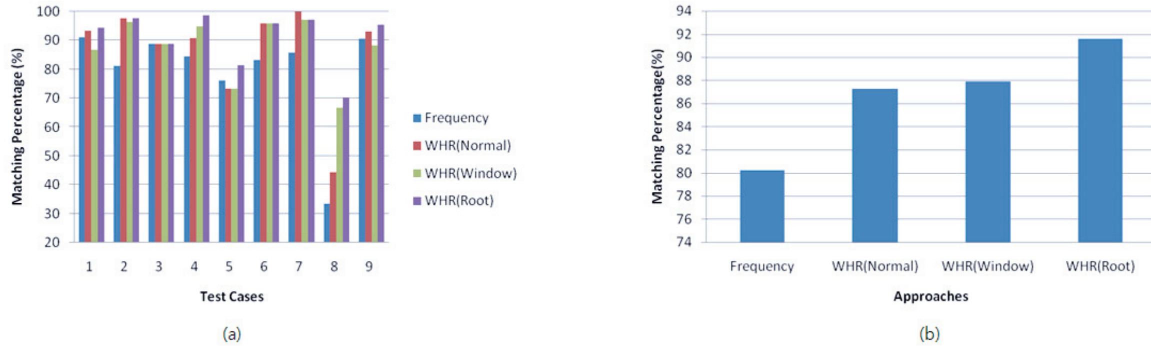

Figure 6: Matching Percentages. (a) shows the results for each test case and (b) averages them out.

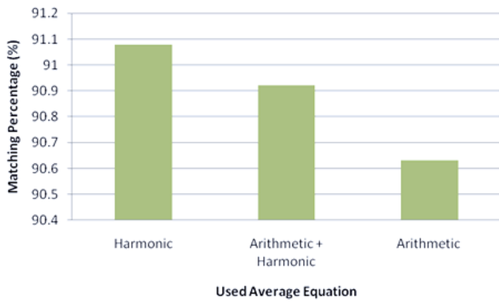

Figure 7: Matching percentage comparison when using different average equations.

Conclusion

In this paper, we proposed several approaches for extracting the latent social relationship from message threads. We evaluated the algorithms with a Facebook dataset. The weighted harmonic rule mining with a root-included sliding window showed the best performance. The visualization and interaction methods for these extracted social networks were also introduced. This would enhance the usability of social network data.

Even though we used the data from a social networking site in the experiment, our proposed model can be applied to any blogs, communities. The mobile web reorganization, friend recommendations, and social advertisements can be processed on the basis of discovered implicit social relations. Our method can also be combined with other previous social relation detection algorithms that do not take into account the complex relationship within message threads. This will result in performance improvement.

References

Liu, B. 2006. *Web Data Mining – Exploring Hyperlinks, Contents and Usage Data.* Springer.

Agrawal, R. and Srikant, R. 1994. Fast Algorithms for Mining Association Rules. In *Proceedings of the 20th VLDB conferences*, 487-499. Santiago, Chile.

Shin, H., Xu, Z., and Kim, E., 2008. Discovering and Browsing of Power Users by Social Relationship Analysis in Large-scale Online Communities. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 105-111.

Furukawa, T., Matsue, Y., and Ohmukai, I. 2007. Social Networks and Reading Behavior in the Blogosphere. In *Proceedings of ICWSM'07*, 51-58.

Karamon, J., Matsuo, Y., and Ishizuka, M. 2008. Generating Useful Network-based Features for Analyzing Social Networks. In *Proceedings of AAAI '08*, 1162-1168.

Lin, Y., Sundaram, H., Chi, Y., Tatemura, J., and Tseng, B. 2006. Discovery of Blog Communities based on Mutual Awareness. In *Proceedings of WWW2006.*

Nauerz, A. and Groh, G. 2008. Implicit Social Network Construction and Expert User Determination in Web Portals. In *Proceedings of AAAI '08*.

Java, A., Kolari, P., Finin, T., and Oates, T. 2006. Modeling the Spread of Influence on the Blogosphere. In *Proceedings of WWW2006.*

Agarwal, N., Liu, H., Tang, L., and Yu, P. S. 2008. Identifying the Influential Bloggers in a Community. In *Proceedings of WSDM 2008*, 207-218.

Wei Wang, W., Jiong Yang, J., and Yu, P. S. 2000. Efficient Mining of Weighted Association Rules (WAR). In *Proceedings of SIGKDD 2000*, 270-274.

Feng Tao, F., Fionn Murtagh, F., and Mohsen Farid, M. 2003. Weighted Association Rule Mining using Weighted Support and Significance Framework. In *Proceedings of SIGKDD 2003*, 661-666.

Costabello, L., and Goix, L. W. 2008. Time Based Context Cluster Anlysis for Automatic Blog Generation. In *Proceedings of WWW2008.*

Heer, J. 2007. Animated Transitions in Statistical Data Graphics. IEEE Transactions on Visualization and Computer Graphics 13(6), 1240-1247.

Herman, I., Melançon, G., and Marshall, M. S. 2000. Graph Visualization and Navigation in Information Visualization: A Survey. IEEE Transactions on Visualization and Computer Graphics 6(1), 24-43.

Heer, J. and Boyd, D. 2005. Vizster: Visualizing Online Social Networks. In *Proceedings of the 2005 IEEE Symposium on information Visualization*, 32-39. INFOVIS. IEEE Computer Society, Washington, DC, 5.

Leung, Y. K. and Apperley, M. D. 1994. A review and taxonomy of distortion-oriented presentation techniques. ACM Trans. Comput.-Hum. Interact. 1, 2 (Jun. 1994), 126-160.