Exploring Social Networks with Topical Analysis

Jiyeon Jang **Jinhyuk Choi** **Gwan Jang*** **Sung-Hyon Myaeng**

Division of Web Science and Technology
***Department of Computer Science**
KAIST, South Korea
{jiyjang, demon, gjang, myaeng}@kaist.ac.kr

Abstract

As the number of social networking services (SNS) and their users grow, so does the complexity of individual networks as well as the amount of information to be consumed by the users. It is inevitable to reduce the complexity and information overload, and we have embarked exploring topical aspects of SNS to form refined topic-based semantic social networks. Our current work focuses on conversational aspects of SNS and attempt to utilize the notions of topic diversity and topic purity between two users sharing conversations. This topic-based analysis of SNS makes it possible to show different types of users and their conversational characteristics. It also shows the possibility of breaking down a huge "syntactic" social network into topic-based ones based on different interaction types, so that the resulting semantic social networks can be useful in designing various targeted services on online social networks.

Introduction

With the growing popularity of SNS and the resulting complexity of the networks, there has been a surge of research on their structural properties such as the size, density, degree of distribution, community structure, link predictability, and information diffusion. The analyses mainly focus on connectivity-based properties of social networks, i.e. *syntactic social networks*, which are formed by explicit connections among users (e.g. "follower-following" relationships in Twitter, "friend" relationships in Facebook).

While the complexity of explicit social networks deserves continuous investigations, we focus on the topical aspects of SNS and attempt to find a way to form *semantic social networks* by paying attention to topicality of individual conversations. This type of semantic social networks can identify smaller and more intact relationships among the users of SNS over the syntactic social networks. With the initial goal of helping SNS users deal with information overload incurred by the large number of tweets in Twitter in the timeline, our investigation concentrates on building ego-centric networks centered around individuals. Individual networks can be integrated to form semantic social networks for the whole and help identifying topically based online community groups.

This paper explores whether and how we can form semantic social networks based on conversations in Twitter. We analyzed topics of tweets exchanged between a particular user and all the connected "friends" (i.e. *following* and *follower* nodes in syntactic social network) in Twitter and attempted to generate an ego-centric network based on the topics. Instead of simply identifying the topics being discussed between two users, *me* (i.e. the center of a network) and a *friend*, we attempted to characterize the relationships between a center and all the friends by introducing two concepts: topic diversity and topic purity. Topic diversity in a relationship indicates the extent to which the relationship shares a variety of topics. Topic purity on the other hand measures the extent to which the shared topics are concentrated in a small number of topics regardless of the number of topics that have been the subject of conversations (i.e. diversity) between the two users.

To show the feasibility of constructing meaningful semantic social networks and delineate the patterns of the ego-centric relationships, we analyzed more than 4.5 million tweets that form more than 1.3 million conversations shared between 1,414 users (i.e. centers) and their conversational partners (friends). The number of unique partners involved was 263,638. That is, we attempted to form 1,414 different semantic social networks whose shapes vary depending on which salient topics to use.

Related Work

Analyzing social networks has been a topic of great interest in the data mining research community. Most of them focus mainly on structural properties of the

networks such as size, density, degree of distribution, or community structure (Mislove et al. 2007; Kempe et al. 2003; Kwak et al. 2010; Cha et al. 2010).

A new line of research on online social networks has emerged beyond analysis of syntactic social networks, mainly focusing on the contents flowing over syntactic networks (Paul et al. 2011; Hong & Davison 2010; Liben-Nowell & Kleinberg 2007; Sousa et al. 2010; Weng et al. 201). Weng et al. (2010), for example, found influential users in Twitter for a specific topic. They extracted topics from the contents each user generated and then computed topical similarity among the users. Topical similarities among the users as well as the link structures of the social networks were used to extend the PageRank algorithm. Sousa et al. (2010) focused on whether the motivation of user interactions is social or topical. They extracted three topics – "sports", "religion", and "politics" – based on keywords from the replied contents each user generated.

While the previous research mainly focuses on characterizing each user, our work characterizes each relationship between two users established by conversations using topical aspects of the conversations. And categorize the relationships in terms of diversity and purity. Instead of focusing on individuals, we focus on the relationships.

Topical Analysis of Conversations

In order to investigate the relationships of the users, we focus on the conversational contents rather than analyzing in isolation the contents individual users generated. Therefore, we analyze topicality of the tweets shared by two users or conversational partners, not those written by a single user. While a conversation can be defined in various ways depending on the types of SNS, they are defined in this paper as a thread of sequential replies in Twitter. Figure 1 shows an example of a conversation in Twitter. Note that a conversational partner of User A is User B and vice versa.

Topics are identified by applying Latent Dirichlet Allocation (LDA) to a collection documents (conversations) for a user. Basically all the words in the tweets are used except for stop words. Once topic distributions are computed for the collection of conversations centered around a user, the result can be used to compute a topic distribution for each relationship between the user (i.e. me) and a conversational partner (i.e. a friend) by taking a mixture of topic distributions corresponding to the conversations shared by the pair of users. Since topic distributions are now available for all the relationship pairs centered around the user, it becomes possible to build an ego-centric semantic social network by selecting a particular topic whose probability is higher than a threshold. We describe the process below as well as the way topic diversity and topic purity are computed. Topic diversity and topic purity are important characteristics of a relationship that can be used to further constrain the semantic social network that is constructed by using a set of topics.

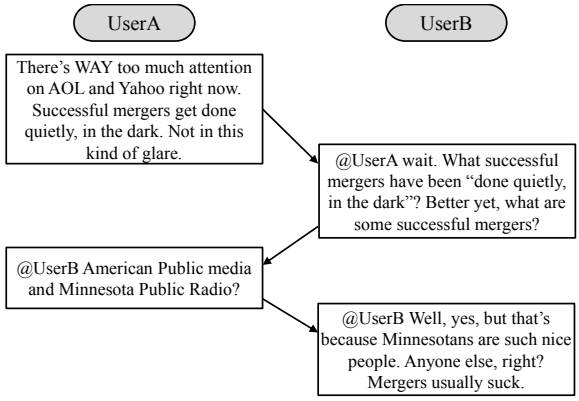


Figure 1: An example of a conversation.

To identify topics for all the relationships centered around a user, we use LDA and model each document (i.e. conversation in this work) as a mixture of topics, each of which is represented as a probability distribution over words, and each word is treated as chosen from a single topic. In LDA, a word document co-occurrence matrix can be decomposed into two parts; document-topic matrix and topic-word matrix. The number of topics we extract is 100.

Document-Topic Matrix for a user shows topic distributions of all the conversations the user has shared with others since we regard one conversation as one document. If two users share only one conversation, the relationship has only one topic distribution; otherwise, it has multiple topic distributions. *Topic-Word Matrix* shows a word distribution in each topic and hence can be used to compute similarities among topics. Given the two matrices and key topics derived from them, we can compute topic diversity and topic purity between two users based on the shared conversations.

Having constructed a document-topic matrix for a user, which contains a topic distribution for each conversation, we can represent each conversation C_i as follows:

$$C_i = (t_{1i}, t_{2i}, t_{3i}, \dots, t_{Ki}),$$

where K is the number of topics and t_{ki} is a probability of k^{th} topic of conversation C_i. When there are multiple conversations for a relationship, we compute a composite topic distribution that embraces all the topic distributions for the purpose of understanding the topics covered between the two users. Mixture of topic distribution, $\mathrm{MTD}(u, u_p)$, of a relationship between two users, a user u and a conversational partner u_p, is computed as follows:

$$\mathrm{MTD}(u,\ u_p) = \left(\frac{\sum_{m=1}^{N} t_{1m} * |C_m|}{\sum_{i=1}^{K}\sum_{j=1}^{N} t_{ij} * |C_j|}, \frac{\sum_{m=1}^{N} t_{2m} * |C_m|}{\sum_{i=1}^{K}\sum_{j=1}^{N} t_{ij} * |C_j|}, \ldots, \frac{\sum_{m=1}^{N} t_{Km} * |C_m|}{\sum_{i=1}^{K}\sum_{j=1}^{N} t_{ij} * |C_j|}\right),$$

where N is the total number of conversations in the relationship, K is the number of topics, t_{ij} is probability of i^{th} topic of conversation j, and $|C_j|$ is the length of conversation j, which is the number of tweets in each conversation. Since the number of characters is limited in a tweet, it makes sense to use the number of tweets as an important factor as it indicates how eagerly two users were engaged in a conversation. MTD essentially represents a composite topic distribution for a relationship across multiple conversations.

Topic diversity (TD) in a relationship is introduced as a way of measuring the degree to which a relationship shares a wide range of topics. A high TD value means the two users have conversed over many different topics. A low value means their conversations stayed in more or less coherent topics. TD can be measured in terms of similarity among the topics for a relationship. In our framework, topical similarity can be computed using topic-word matrix which consists of word distributions for individual topics identified. Among several similarity metrics we can choose from, we opted for JS Divergence because it is commonly used for topical similarity measurement for its superiority (Blei, Ng, and Jordan 2003; Weng et al. 2010; Kim and Oh 2011).

Dissimilarity D_{ij} between two topics t_i and t_j can be calculated as:

$$D_{ij} = JSD(t_i, t_j) = \frac{1}{2}\left(KL(t_i||M) + KL(t_j||M)\right),$$

where $M = \frac{1}{2}(t_i + t_j)$ and $\mathrm{KL(P||Q)} = \sum_i P_i \log\frac{P_i}{Q_i}$. KL stands for KL Divergence. After calculating topic dissimilarities among all topics identified for a relationship, topic distance matrix of a user u can be expressed as follows:

$$\text{Topic Dist}\ (u) = \begin{bmatrix} D_{11} & D_{12} & \cdots & D_{1K} \\ D_{21} & D_{22} & \cdots & D_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ D_{K1} & D_{K2} & \cdots & D_{KK} \end{bmatrix},$$

where K is the number of topics and D_{ij} represents the topic dissimilarity between two topics t_i and t_j.

Since topic diversity should be high when dissimilar topics are highly represented in topic distribution, we multiply $\mathrm{MTD}(u, u_p)$ and $\text{Topic Dist}\ (u)$ to result in a vector where each element indicates how distinct the corresponding topic is in comparison with other topics. By taking an average of the distinctiveness of each topic, topic diversity can be measured. Therefore, $\text{Topic Diversity}(u, u_p)$ of a relationship between a user u and a conversational partner u_p can be computed as:

$$\text{Topic Diversity}\ (u,\ u_p) = \mathrm{AVG}\left(\mathrm{MTD}(u,\ u_p) \times \text{Topic Dist}\ (u)\right)$$

where $\mathrm{AVG}(.)$ is a scalar product of the vector and its unit vector.

Topic purity indicates the tendency a relationship or the conversations carried out by two users focuses on narrow topics. If two users exchanged tweets on local politics only, for example, their topic purity is maximal. Even if they talked about many different topics occasionally but tended to get into conversations on a particular topic, their topic purity would be also quite high. The more uniform a topic distribution, the lower topic purity. Note that a relationship may have higher purity even with a greater number of salient topics than another with less number of topics. A relationship with higher topic diversity can still have higher purity than others with lower topic diversity.

Since the topic purity detects whether there are a small number of outstanding topics, a natural choice for a metric would be entropy; once we obtain MTD or a composite topic distribution for a relationship, entropy can be computed in a straightforward way. However, we chose a much simpler method of taking the maximum value of elements in MTD. This is because our interest was to identify a relationship that has an outstanding topic. Given that the sum of all the probability values in MTD is 1, it is sufficient to use the maximum probability value of the outstanding topics to represent topic purity. Thus, $\text{Topic Purity}(u, u_p)$ of a relationship between a user u and a conversational partner u_p can be calculated as:

$$\text{Topic Purity}\,(u, u_p) = Max\left(\text{MTD}(u, u_p)\right),$$

where $\sum_{k=1}^{K} \text{MTD}_k(u, u_p) = 1$, K is the number of topics.

Analysis of Semantic Social Networks

Dataset

We chose Twitter to collect the conversational data because of its openness, availability, and activeness. Since Twitter allows its users to upload their tweets and react to tweets of other users by a few options such as "Favorite", "Retweet", and "Reply". To detect conversations, we used the "Reply" option.

To collect our dataset, we crawled public timeline[1] of Twitter from September 29th, 2011 to October 4th, 2011, so as to sample users randomly. Then we examined all the tweets and the users of the tweets whether or not they satisfy the following conditions:

- Each tweet crawled must be written in English
- The total number of tweets of a user identified from the crawled tweets should be over 3,200.

We randomly sampled 2,036 users among those who satisfy the conditions and collected all the conversations they were engaged in.

In order to track all the conversations of the users, we identified the tweets that were replied to some other tweets. Then, we repeatedly followed the chain of replies to recover the complete set of conversations. After collecting all the conversations, we duplicated a conversation to multiple copies if more than two users were involved in it so that each conversation in our dataset has only two users.

Before we constructed semantic social networks for analysis, we refined our dataset further. In order to ensure we had enough data for topic extraction, we identified the users with more than 400 conversations. In addition, we removed the conversations whose length is less than 2 tweets. The volume of dataset used in our experiment is described in Table 1.

Total number of users	1,414
Total number of conversations	1,338,022
Total number of tweets in conversations	4,582,461
Total number of unique conversational partners	263,638

Table 1: The volume of dataset used in our analysis.

[1] http://twitter.com/public_timeline provides the 20 most recent tweets in Twitter. This public timeline is cached for 60 seconds.

Characterizing Topic-based Relationships

We define a semantic social relationship R as follows:

$$R = < u, u_p, \vec{P}, TD, TP >$$

A semantic social relationships exist between a user (u) and a conversational partner (u_p). Each relationship has its topic distribution vector $\vec{P}$ computed by MTD, topic diversity, and topic purity. In the current experiment, each user pair has 100 topic-specific relationships since $\vec{P}$ contains a topic probability for each topic of 100 topics that were extracted in this study.

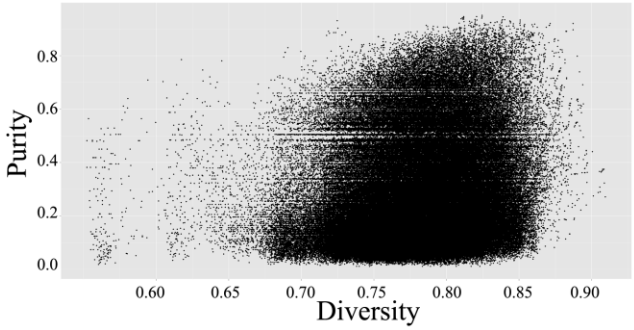


Figure 2: Topical social relationship Distributions

We first analyzed the overall trend of all the relationships in terms of their topic diversity and purity values. As in Figure 2 where topic diversity and purity values for relationships are plotted, we can see that the relationships lean toward high diversity and low purity since the median values of topic diversity and purity are about 0.77 and 0.22, respectively. Moreover, the relationships in the ranges of 0.76 and 0.78 in topic diversity and 0.19 and 0.25 in topic purity, which hardly show tendencies, account for about 40% of all relationships. The rest can be divided into four categories: 24% of the relationships have a tendency toward high diversity and purity, 17% toward high diversity and low purity, 13% toward low diversity and purity, and 7% toward low diversity and high purity.

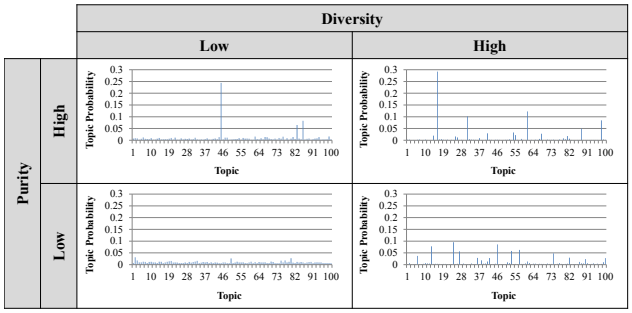
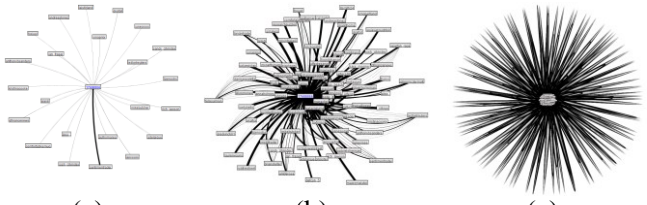


Figure 3: A sample of relationships for different categories.

To get a sense of the characteristics of the relationships belonging to each of the four categories, we select one sample for each and illustrate what the topic distributions look like in Figure 3. Note that when the four samples were chosen, we ensured the numbers of conversations and tweets are almost same across the found cases. We can recognize the high diversity relationships on the right have more peaks than those on the left. High purity relationships in the upper row, on the other hand, have higher peaks than those in the low row. Reciprocally, the graph patterns indicate that the two measures, diversity and purity of a topic, seem appropriate in characterizing conversational relationships.

Semantic vs. Syntactic Social Networks

The main differences between semantic and syntactic social networks lie in the size and richness of the relationships. The size of a social network can be reduced simply by considering whether a relationship is purely based on following and follower connections or based on conversational relationships. It can be further reduced by considering the types of interactions based on topic diversity and purity. For example, a network can be formed by only considering the conversational partners whose relationships have high topic diversity and purity. Furthermore, a much simpler network can be formed by considering a particular topic. An example would be an ego-centric network for a user and the partners who have shared conversations on 'finance'.



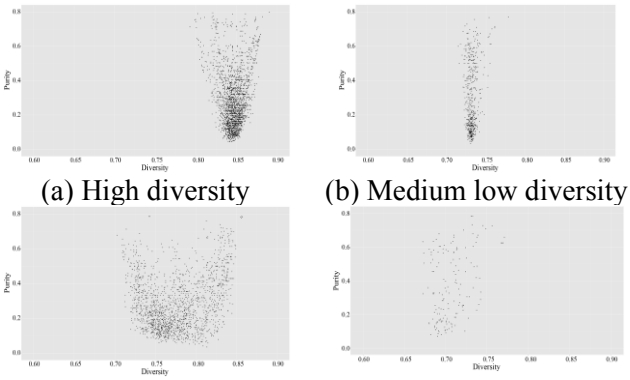
(a) (b) (c)

Figure 4: Different networks created for a user and the partners depending on the number of topics considered

The biggest advantage of semantic social network comes from the fact that we can identify sub-networks by selecting topics on relationships. Figure 4 (a) shows a network of conversational partners on a particular topic[2]. At the center is the node for the user who is connected to about 20 conversational partners by an edge. The thickness of an edge indicates intensity of the topic in conversations with the partner. As topics are added, the network becomes denser as can be seen in (b). Since a relationship between the user and a particular partner can have up to 100 edges corresponding to the maximum number of topics in our current implementation, the network becomes much more complex when no topic selection is done. The 'core' at the center in Figure 4 (c) represents all the partners, which are heavily concentrated in a small region while each spike means a topic-labeled arc that links the user and a partner. Since there can be up to 100 links between the user and a partner, the visualization package[3] we used show them this way.

Characterizing Users

Users can be characterized based on their behaviors reflected on the types of their conversational relationships. Figure 5 shows four different types of users sampled from our data, characterized by the tendency of the conversational relationships they had. The user shown in (a) has a tendency of having relationships with high diversity with varying purity whereas the user in (b) tends to stay in a small number of topics (low diversity) across all the relationships but vary widely in purity. The user in (c) is shown to have very diverse types of relationships. Compared to the other users, the user in (d) does not have as many relationships but tend to stay in a smaller number of topics with relatively higher purity, indicating that s/he would enjoy focused conversations on rather limited topics with a small number of friends.



(a) High diversity (b) Medium low diversity

(c) Widely scattered (d) Low diversity and high purity

Figure 5: Examples of relationship distributions of four different users. Each dot represents a relationship.

[2] The topic in this figure is on 'finance', which is actually represented by a set of words {banks, allessio, rastani, financial, loans}.

[3] http://jung.sourceforge.net. JUNG: Java Universal Network/Graph Framework,

Summary and Future Work

Our study is on discovering and exploring a new type of social networks – semantic social networks – based on topical aspects of conversations between a user and its partners. To elicit topics from Twitter conversations, we applied LDA, a widely used topic modeling tool. In order to characterize different types of topical interactions, we introduced the notion of topic diversity and purity that can be computed for individual relationships. Using these measures, users can be classified or characterized in terms of their conversational behaviors or styles in online interactions with "friends".

We focused on how semantic social relationships can be established in an ego-centric social network and explored ways to utilize such networks. We showed a way of categorizing users using their conversational behaviors based on different combinations of topic diversity and purity measures of the established relations. The categorization can help not only understanding the way an individual interacts with his/her online friends but also making it amenable to group users who show similar behaviors.

We also showed how semantic social networks constructed in the proposed way can alleviate the complexity of networks and information overload in SNS, which should be faced by the entities providing the services and actual users. Social networks can be reduced to much smaller semantic networks by specifying one or more topics of interest while finding new meaningful connections that are not available in syntactic networks.

In addition to the obvious benefits of semantic social networks, they can be used in a more application-oriented manner. For example, the patterns of the topical interactions identified for individual users can be used to filter out or recommend contents in SNS. This kind of service can be refined further by understanding how diverse or pure the past interactions have been. For the users showing high diversity in the relationships, for example, the service may not want to adhere to the history of the topics covered in the conversations so much.

There are several avenues we plan to explore for future research. We are currently investigating further on different ways to analyze topic-based user patterns. For instance, we are applying more sophisticated linguistic processing for noisy data. Other issues include what would happen if we use retweets or favorites in extracting topics and how to analyze temporal aspects of topics since user interests would change over time.

A natural extension to the current framework targeted at ego-centric networks is to integrate individual networks to build general semantic social networks that include a group of people, if not the entire population. Another direction is to compare and combine syntactic and semantic social networks for a synergy. Few studies have examined both of structural properties and semantic properties of online social networks (Li et al. 2011). Still another avenue to explore is a variety of applications that can be possible by using semantic social networks.

Acknowledgements This research was supported by WCU (World Class University) program under the National Re-search Foundation of Korea and funded by the Ministry of Education, Science and Technology of Korea (Project No: R31-30007)

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet Allocation, Journal of Machine Learning Research, v.3, 993-1022.

Cha, M., Haddadi, H., and Benevenuto, F. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy, Proc. ICWSM.

Hong, L. and Davison, B. D. 2010. Empirical study of topic modeling in Twitter, Proc. 1st Workshop on Social Media Analytics (SOMA).

Kempe, D., Kleinberg, J., and Tardos, E. 2003. Maximizing the spread of influence through a social network, Proc. KDD.

Kim, D. and Oh, A. 2011. Topic Chains for Understanding a News Corpus, Proc. CICLING.

Kumar, R., Novak, J., and Tomkins, A. 2006. Structure and Evolution of Online Social Networks. Proc. KDD.

Kwak, H., Lee, C., Park, H., and Moon, S. 2010. What is Twitter, a Social Network or a News Media? Proc. WWW.

Li, D., Ding, Y., Sugimoto, C., He, B., Tang, J., Yan, E., Lin, N., Qin, Z., and Dong, T. 2011. Modeling Topic and Community Structure in Social Tagging: the TTR-LDA-Community Model, Journal of the American Society for Information Science and Technology, 62(9), 1849-1866.

Liben-Nowell, D. and Kleinberg, J. 2007. The link prediction problem for social networks, Journal of the American Society for Information Science and Technology, 58(7), p.1019-1031.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. 2007. Measurement and and Analysis of Online Social Networks, Proc. IMC.

Paul, S. A., Hong, L., and Chi, E. H. 2011, Is Twitter a Good Place for Asking Questions? A Characterization Study. Proc. ICWSM.

Sousa, D., Sarmento, L., and Rodrigues, E. M. 2010. Characterization of the Twitter @replies Network: Are User Ties Social or Topical? Proc. SMUC.

Weng, J., Lim, E. -P, Jiang, J., and He, Q. 2010. TwitterRank: finding topic-sensitive influential twitterers, Proc. WSDM.