Semantic Social Networks Constructed by Topical Aspects of Conversations: An Explorative Study

Jiyeon Jang[†‡] **Jinhyuk Choi**[†] **Gwan Jang**[‡] **Sung-Hyon Myaeng**[†]

[†]Division of Web Science Technology [‡]Department of Computer Science
Korea Advanced Institute of Science and Technology, 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 net works as well as the amount of information to be consumed by the users. Users of SNS exchange short and instantane ous messages interactively, which can be seen as conversa tions. We explore this conversational aspect of SNS and show how refined topic based semantic social networks can be formed in order to reduce the complexity and infor mation overload. Among other possibilities, we use the no tion of topic diversity and topic purity of SNS conversations between two users and show different types of social rela tionships can be identified in that they break down a huge "syntactic" social network into topic based ones based on different interaction types. Resulting semantic social net works can be useful in designing various targeted services on online social networks.

Introduction

Many social networking services (SNS) such as Twitter, Flickr, or Facebook have emerged over the past few years. A unique characteristic of SNS is that the users are able to form their own online social networks for different purposes such as content sharing, communication, and news reporting (Java et al. 2007). 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 (Kumar, Novak, and Tomkins 2006; Mislove et al. 2007; Liben-Nowell and Kleinberg 2007; Kwak et al. 2010; Cha, Haddadi, and Benevenuto 2010). These 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 and "friend" relationships in Facebook).

While the complexity of explicit social networks deserves continuous investigations, a new line of research on online social networks has emerged mainly focusing on the contents flowing over syntactic networks (Hong and Davison 2010; Magnani et al. 2011; Sousa, Sarmento, and Rodrigues 2010; Weng et al. 2010). Weng et al. (2010), for example, found influential users in Twitter for a specific topic. They extracted topics from the contents generated by each user and computed topical similarities among the users, which were then used together with the link structures of the social networks 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.

This paper explores whether and how we can form *se mantic social networks* based on topicality of conversations in Twitter. We analyze topics of tweets exchanged between a particular user and all the connected friends in Twitter and attempt to generate an egocentric network based on the topics. Instead of simply identifying the topics being discussed between two users, such as me (i.e. the center of a network) and a friend, we attempt 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 on 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.

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. That is, 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, it is defined in this paper as a thread of sequential replies preceded by the initiating tweet in Twitter. Figure 1 shows an example of a conversation in Twitter. In the figure, two conversations exist; a thread of

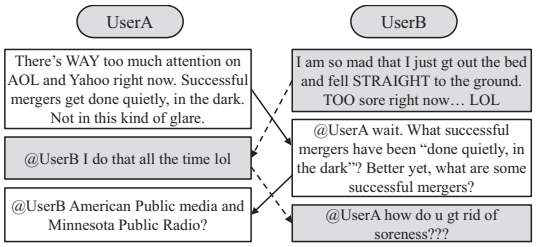


Figure 1: An example of a conversation.

white boxes and that of gray boxes. Note that a conversational partner of User A is User B and vice versa.

To identify topics for all the relationships centered around a user, we use Latent Dirichlet Allocation (LDA), which is a commonly used method for topic modeling (Blei, Ng, and Jordan 2003; Griffiths and Steyvers 2004; Steyvers and Griffiths 2007). LDA models 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. We set hyper-parameters α and β to 0.1 and 0.01, respectively, which were commonly used in the past (Kim and Oh 2011). 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 a conversation (document)-topic matrix for a user, which contains a topic distribution for each conversation, we can represent each conversation C_i as follows:

$$\mathrm{C}_i = (t_{1i}, t_{2i}, t_{3i}, \dots, t_{Ki}),$$

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 distributions, $\mathrm{MTD}(u, u_p)$, of a relationship between two users, a user u and a conversational partner u_p, is computed as follows:

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

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 $|\mathrm{C}_j|$ 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.

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 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 has been commonly used for topical similarity measurement for its superiority (Blei, Ng, and Jordan 2003; Weng et al. 2010; Kim and Oh 2011).

Topic distance matrix of a user can be constructed by calculating topic dissimilarities among all topics identified for a relationship 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, D_{ij} represents the topic dissimilarity between two topics t_i and t_j, and D_{ij} is calculated as $\mathrm{D}_{ij} = JSD(t_i, t_j) = \frac{1}{2}\ (KL(t_i||M) + KL(t_j||M))$, 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.

Topic diversity should be high when dissimilar a variety of topics are strongly represented in topic distribution. As such, we multiply $\mathrm{MTD}(u, u_p)$ and $\text{Topic Dist}\ (u)$ to form a vector where each element indicates how strong the corresponding topic is in comparison with other topics. Topic diversity can be measured by taking an average of the distinctiveness of individual topics. $\text{Topic Diversity}(u, u_p)$ of a relationship between two users can be computed as:

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

where $\mathrm{AVG}(.)$ is an average value of elements of a vector.

Topic purity indicates the tendency that a relationship (the conversations carried out by two users) focuses on specific topics. If two users exchanged tweets on a single topic such as local politics only, for example, their topic purity is maximal. Even if they talked about many different topics occasionally, their topic purity would be quite high if they tended to elaborate on a particular topic more frequently. 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. It is entirely possible for a relationship with higher topic diversity to have higher purity than others with lower topic diversity.

Since the topic purity detects whether there are a small number of outstanding topics, we chose a simple method of taking the maximum value of elements in MTD. This is because our interest is 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 topic to represent topic purity. Thus, Topic Purity(u, u_p) of a relationship between two users 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, and $\text{MTD}(\text{u}, \text{u}_p) = (\text{MTD}_1(u, u_p), , \ldots, \text{MTD}_k(u, u_p))$.

Analysis of Semantic Social Networks

Dataset

We chose Twitter to collect the conversational data. To detect conversations, we used the "Reply" options although Twitter allows users to react to tweets of other users by "Favorite" and "Retweet" as well as "Reply".

To collect our dataset, we first randomly sampled 2,036 users who use English, have more than 3,200[1] tweets in total, and have at least one conversation between September 9th, 2011 and October 4th, 2011. We then 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. We repeatedly followed the chain of replies to recover the complete set of conversations. After collecting all the conversations, we duplicated a conversation into multiple copies if more than two users were involved in it so that each conversation in our dataset has only two users.

In order to ensure we had enough data for topic extraction, we identified the users with more than 400 conversations, removed the conversations whose length is less than 2 tweets, and removed special characters and stop words. Consequently, our dataset contains 1,414 users with their 263,638 unique conversational partners, and 1,338,002 conversations including 4,582,461 tweets in total.

Characterizing Topic-based Relationships

We define a semantic social relationship R as follows:

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

A semantic social relationship exists between a user (u) and a conversational partner (u_p). Each relationship has its topic distribution vector $\vec{P}$ by computed MTD, topic diversity TD, and topic purity TP. 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.

We first analyzed the overall trend of all the relationships in terms of their topic diversity and purity values. 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 values are about 0.77 and 0.22, respectively. Moreover, the relationships in the rang-

[1] Twitter API limits that the maximum number of tweets to show is 3,200 even if a user wrote more than 3,200 tweets

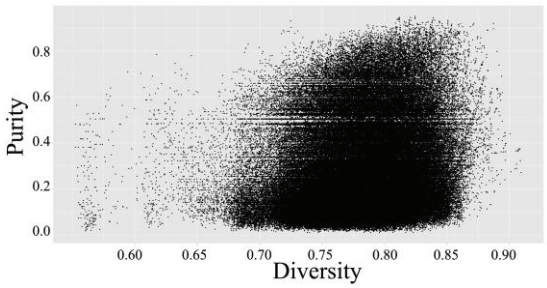


Figure 2: Distribution of topical social relationships.

es 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 high purity, 17% toward high diversity and low purity, 13% toward low diversity and low purity, and 7% toward low diversity and high purity. This analysis suggests that about a half of the relationships can be categorized by patterns of diversity and purity combinations although the majority of the conversational relationships tend to talk about diverse topics without a small number of salient, concentrated topics. It is encouraging, however, to observe that a quite large number of relationships have high purity and carries out conversations with a focus.

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 as in Figure 3. Note that the four samples are chosen in such a way that the numbers of conversations and tweets are almost the same across the four 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 lower 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-follower connections or based on conversational relationships, the types of interactions based on topic dive-

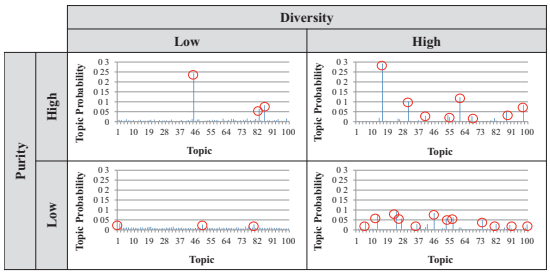
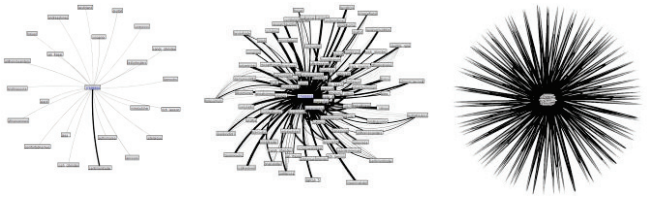


Figure 3: Topic distributions for a sample of relationships for different categories.



(a) (b) (c)

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

rsity and purity, and a particular topic.

The biggest advantage of semantic social network comes from the fact that we can identify sub-networks by selecting topics on relationships and the types of relationships. Figure 4 (a) shows a network of conversational partners on a particular topic[2] with high diversity and high purity relationships. 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 analyses, 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.

Discussion

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 each of her partners. In order to characterize different types of topical interactions, we introduced the notions of topic diversity and topic purity that can be computed for individual relationships. Using these measures, social relationships of users can be characterized in terms of their conversational behaviors or styles in online interactions with the "friends".

The resulting social networks can be used in various applications. 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. On the other hand, if a user shows high purity in the relationships, it is likely to make him/her happy by delivering the contents from the people whose profile matches.

There are several avenues we plan to explore for future research. We are going to investigate user patterns based on their relationships, include other interactions such as retweets and favorites, and analyze temporal aspects of topics since user interests would change over time. Also, 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. Another direction is to compare and combine syntactic and semantic social networks for a synergy (Li et al. 2011). Still another avenue to explore is a variety of applications that can be made possible by using semantic social networks.

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

[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

References

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

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

Griffiths, T. L. and Steyvers, M. 2004. Finding Scientific Topics, Pro ceedings of the National Academy of Sciences, v.101, 5228 5235.

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

Java, A., Song, X., Finin, T., and Tseng, B. 2007. Why We Twitter: Un derstanding Microblogging Usage and Communities. Proc. Web KDD/SNA KDD.

Kim, D. and Oh, A. 2011. Topic Chains for Understanding a News Cor pus, 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, JASIST, 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.

Magnani, M., Montesi, D., Nunziante, G., and Rossi, L. 2011. Conversa tion Retrieval from Twitter, Proc. ECIR.

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

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

Steyvers, M. and Griffiths, T. L. 2007. Probabilistic Topic Models, T. Landauer, D. McNamara, S. Dennis, and W. Kintesch, editors, Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum, In Press.

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