LETTER

Discovery of Information Diffusion Process in Social Networks*

Kwanho KIM[†], *Nonmember*, Jae-Yoon JUNG[††a)], *Member*, *and* Jonghun PARK[†], *Nonmember*

SUMMARY Information diffusion analysis in social networks is of significance since it enables us to deeply understand dynamic social interactions among users. In this paper, we introduce approaches to discovering information diffusion process in social networks based on process mining. Process mining techniques are applied from three perspectives: social network analysis, process discovery and community recognition. We then present experimental results by using a real-life social network data. The proposed techniques are expected to employ as new analytical tools in online social networks such as blog and wikis for company marketers, politicians, news reporters and online writers.

key words: *information diffusion process, online social networks, social network services, process mining*

1. Introduction

With the advent of social network services such as Twitter, Facebook and Tumblr in the last decade, information is able to be diffused rapidly through social networks. As users publish information to others and republish the publications of others, information is forwarded from users to users, which makes an information diffusion process. Thus, information diffusion data for a user can be considered as the historical logs of actual interactions among users. For instance, in social network services, information which can be regarded as posts such as user's statuses, news articles, bookmarks and photos is diffused by using functions like "tweet" and "retweet" on Twitter and "blogging" and "reblogging" on Tumblr.

Analysis of information diffusion among publishing and republishing activities in social networks is significant from two perspectives. First, we can understand overall social behaviors among the users in a social network service. Second, a user can recognize his/her own social interactions as the following issues: which users are influencing or influenced by the user, how the user's posts are propagating to other users, and what communities are organized around the user. Moreover, those findings can be further utilized for the various applications such as social marketing [1], public opinion mining [2], and information distribution analysis [3].

There have been many existing studies on social behaviors, interaction patterns, user's influence, information epidemic, and information propagation by investigating social links such as friends and followers [4], [5]. However, the previous approaches rarely reflect the dynamic nature of social interactions over social networks, since social interactions are not limited to static relations among users. For example, users in social network services such as Twitter and Tumblr can interact with other users regardless of static relationships. Therefore, we expect that analyzing information diffusion which reflects the actual interactions among users can help to deeply understand the dynamic nature of social interactions.

In this paper, we present an innovative approach to application of process mining to information diffusion analysis in social network services. Process mining techniques support process-oriented analysis of historical data by employing data mining and machine learning techniques. Specifically, we propose an information diffusion model and attempt to analyze the information diffusion process over the ego-centric network for a user. Moreover, for the ego-centric network, we analyze the social network and the communities based on the relations among users and their republishing history. To achieve the objectives, we adopt three representative techniques, such as social network mining, α-algorithm, and fuzzy analysis [6]. The techniques provide different perspectives (i.e. social network, process, and community) of information diffusion process in social networks.

We regard homogeneous events across users (e.g. republishing the same post) as a process instance. The assumption is reasonable since the republications are repeatedly made among users in a social network and republications related to the same topics are likely to follow the similar diffusion process in the same community.

The experimental results show that such analysis techniques for information diffusion process can be utilized as new analytical tools on social networks to understand the ego-centric network for a user.

2. Information Diffusion Process

In this section, we first present an information diffusion model for social network services and then introduce three process mining techniques for the discovery of information diffusion process on social network services.

Manuscript received December 13, 2011.

[†]The authors are with Seoul National University, Gwanak-ro, Gwanak-gu, Seoul, 151–744, Republic of Korea.

[††]The author is with Kyung Hee University, Yongin, Gyeonggi-do, 446–701, Republic of Korea.

[*]This research was supported by the Korea Sanhak Foundation (No.400-20110159), and partly by the National Research Foundation of Korea (NRF) grants funded by the Korea Government (MEST) (Nos.2011-0003560, 2011-0004423, and 2011-0030814) and Engineering Research Institute at Seoul National University.

a) E-mail: jyjung@khu.ac.kr (Corresponding author)

DOI: 10.1587/transinf.E95.D.1539

2.1 Information Diffusion Model

A social network is a tuple $S = \langle U, F\rangle$ where $U = \{u_i \mid i = 1, \ldots, N\}$ represents a set of N users and $F \subseteq U \times U$ is a finite set of post subscription relations between two users. For example, $S = \langle\{u_1, u_2, u_3\}, \{(u_2, u_1), (u_3, u_1)(u_3, u_2)\}\rangle$ represents a social network consists of three users in which u_2 subscribes to the posts of u_1, and u_3 subscribes to the posts of u_1 and u_2. Note that $(u_i, u_i) \notin F$ for $u_i \in U$.

In a social network S, a publishing and republishing activity of a user is denoted as $e = \langle p, r, t\rangle$ where p means the post, $r \in F \cup (U \times \{\phi\})$ is the republishing relation for p between two users, and t is the timestamp of the activity. If $r = (u_2, \phi)$, it means that u_2 published a new post p. For instance, $e_1 = \langle p, (u_2, u_1), t\rangle$ represents that u_2 republishes post p of u_1 at time t. We let a set of republishing relations for post p as $R(p) = \{r \in U \times U \mid \exists e = \langle p, r, t\rangle\}$.

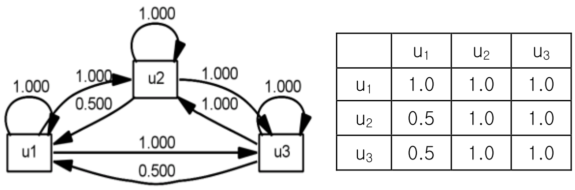
Finally, we develop an information diffusion history model from a set of publishing activities, which represents the publication or republication cases of all the posts on a social network service. An information diffusion history H is a set of pairs $(p, R(p))$ and formalized as $H = \{(p, R(p)) \mid \exists p.\forall r \in R(p).\exists e = \langle p, r, t\rangle\}$.

Prior to introducing the techniques, let us compare information diffusion process and business process. The process mining assumes that there exist one or more process models in a given event history logs. Similarly, we can assume that there are one or more information diffusion processes in an ego-centric social network. It is because that the posts of a certain user contain one or more topics and the posts will be diffused according to the topics on the social network.

2.2 Social Network Analysis

We employ the previous social network analysis method proposed in [7], [8]. The method focuses on discovering the relations among users who have participated together in the same process by utilizing process mining techniques. Since information diffusion history in social network services has already contained explicit information of such as users and republishing history, and it is straightforward to apply the social mining to the analysis of information diffusion process in social networks.

By adopting the technique for a given information diffusion history, we can obtain the sociography such as an example shown in Fig. 1. The arc weights in the graph can be interpreted as how often two users publish (or republish) the same information (i.e. posts) together. In detail, the matrix shows how much percentage a user in the column republished the posts that another user in the row republished. For instance, $(u_2, u_1) = 0.5$ means that u_1 republished half of posts republished by u_2. In summary, a sociography shows the relations and shared interests among users in a social network.



	u1	u2	u3
u1	1.0	1.0	1.0
u2	0.5	1.0	1.0
u3	0.5	1.0	1.0

Fig. 1 Example of social network analysis.

2.3 Process Discovery

The problem of information diffusion process discovery can be defined as how to map information diffusion history to a process model such that the model is "representative" for the behaviors in the history. There are many existing techniques for discovering a process model from historical event log. Among them, we consider α-algorithm [9] which is one of the first process discovery algorithms.

Given an information diffusion history H, the α-algorithm produce an information diffusion process model based on Petri net by analyzing four kinds of ordering relations: direct succession, causality, unrelated, and parallel, denoted by $>_H$, $\rightarrow_H$, $\#_H$, and $\|_H$, respectively. For an information diffusion history H, a republication relation r is equivalent to the direct succession relation of the algorithm. Therefore, we can describe four ordering relations of the α-algorithm for information diffusion process discovery as follows:

- $u_2 >_H u_1$ if and only if there is a republication relation $r = (u_2, u_1)$ in H.
- $u_2 \rightarrow_H u_1$ if and only if $u_2 >_H u_1$ and not $u_1 >_H u_2$.
- $u_2 \#_H u_1$ if and only if not $u_2 >_H u_1$ and not $u_1 >_H u_2$.
- $u_2 \|_H u_1$ if and only if $u_2 >_H u_1$ and $u_1 >_H u_2$.

Based on the result of investigating those ordering relations in the information diffusion history H, we can generate a Petri net-based information diffusion process by using the α-algorithm.

2.4 Community Recognition

Since, different from general process mining, our main focus is on users rather than tasks, we employ the fuzzy mining algorithm to discover communities which are groups of related users and understand the relations of the users in and over the communities [10].

In the fuzzy mining algorithm, relationships among users are obtained by calculating relative significance between users which is defined as:

$$rel(u_1, u_2) = \frac{1}{2}\frac{fr(u_1, u_2)}{\sum_{u_x \in U} fr(u_1, u_x)} + \frac{1}{2}\frac{fr(u_1, u_2)}{\sum_{u_x \in U} fr(u_x, u_2)}$$

where $fr(u_1, u_2)$ represents the frequency of observing the two users' relation.

Through the results of the fuzzy mining, we can be provided for the simple visualization of the discovered relations among communities and hierarchical view of users and communities. Since information diffusion involving many users is often complex and ad-hoc, we abstract the discovered process to obtain comprehensible model by adjusting the level of relative significance between users.

Note that some process mining techniques like fuzzy mining and trace clustering were not developed for structured business process, but for less-structured or ad-hoc processes in real-life. Therefore, the fuzzy mining is useful to effectively simplify and filter the complicated interactions on social networks by capturing the communities.

3. Experimental Results

For the three purposes described in Sect. 2: social network analysis, process discovery and community recognition, we conducted three experiments by applying the suggested methods. For the experiments, we collected dataset from Tumblr which is one of the most popular micro-blog services in the world. The service has roughly 9.9 billion posts in 28 million blogs. In the experiments, we aimed at providing substantial knowledge for the information diffusion process for a single user. This kind of social network analysis is often called "ego-centric network" analysis. An ego-centric network includes as its nodes the ego user and all nodes to which the ego user has a connection within some specified path length.

We chose a user "sunsurfer" among the most active users to construct an ego-centric network in order to demonstrate the results of the experiments. We then collected the posts which the ego user has published or republished in recent 3 weeks and that have been republished more than 10 times at August 30, 2011. As a result, 73 posts were collected and the posts were totally republished 2446 times by 433 different users. Note that we used a process mining tool ProM 5.2 [11] for the experiments.

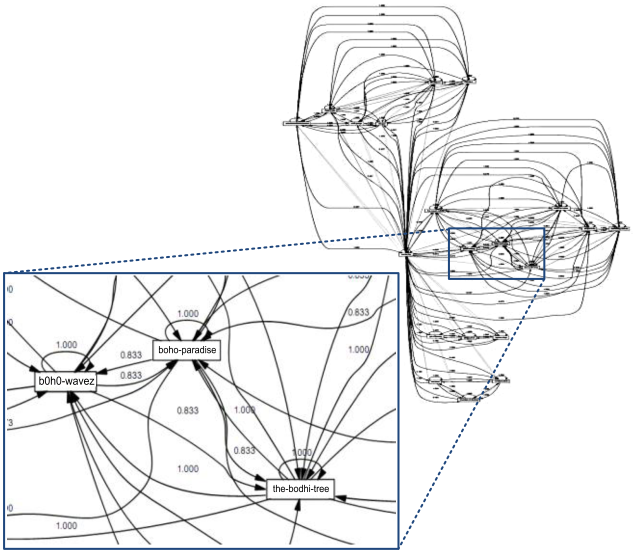


Fig. 2 Discovered ego-centric network by analyzing the relations among users.

First, social network analysis for information diffusion was performed on the ego-centric network, and the result is shown in Fig. 2. Due to the space limitations, the diagram depicts social interactions among only the top 20 frequent users in the network. We can recognize roughly four groups in the whole diagram and the detailed interaction of some users in the zoomed-in figure.

Next, an experiment for discovering an information diffusion process model was made by utilizing the α-algorithm. Figure 3 shows the discovered network obtained by analyzing the history of republications related to the ego user. The network consists of the ego user and the top 25% users in terms of their relation weights against the ego user. The result shows how the posts published or republished by the ego user were diffused by other users in the social network. In the zoomed-in figure, we can expect that if a post published or republished by the ego user reaches to "lonelyoceans", the post will be probably transmitted again to "fountain-of-fortune" via two intermediate paths: "exotic-tea" and "exotic-c", or "em-bers".

Finally, we applied the fuzzy mining algorithm to find communities for the ego user. Figure 4 shows the result by

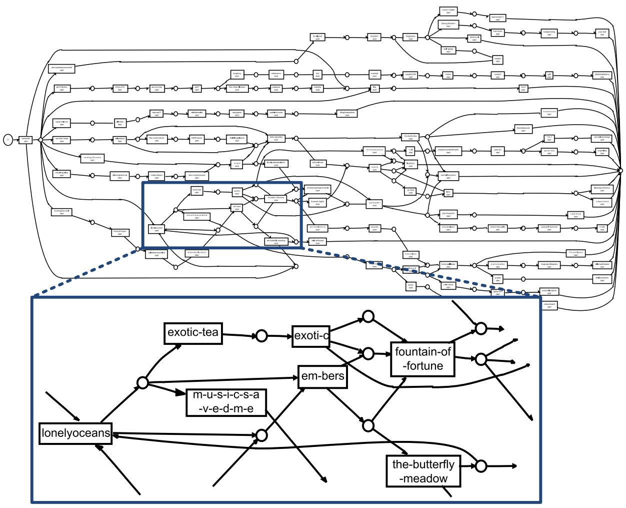


Fig. 3 Discovered information diffusion model for an ego-centric network.

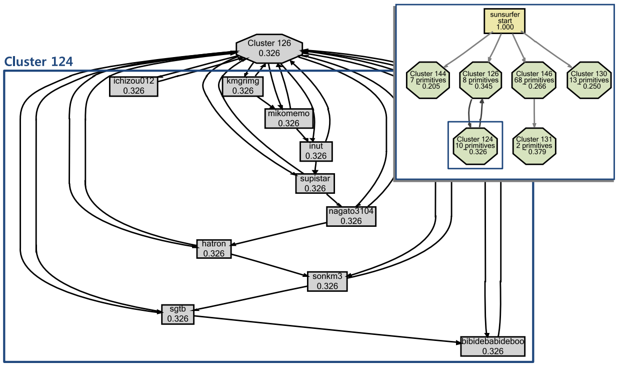


Fig. 4 Discovered communities in an ego-centric network.

setting relative significance level to 0.6. The figure represents the six communities composed of the top 25% users are mainly related to the ego user. The posts of the ego user tend to be diffused to four communities (clusters 144, 126, 146, and 130) and then diffused again to two other communities (clusters 124 and 131). In particular, some posts are republished from cluster 124 to 126. The zoomed-in figure for cluster 124 shows in detail how 10 users in the cluster are related to each other. Through the analysis, we can recognize the posts of the ego user cannot be diffused to two clusters 124 and 131 without the republications of clusters 126 and 146, respectively.

4. Conclusion and Discussion

As social network services are growing rapidly, a huge amount of information is being diffused on the Web. In this paper, we proposed methods for discovering information diffusion process which reflect actual interactions among users in social networks by using process mining techniques. Through the experimental results of a real-life dataset, we illustrated that the proposed approach can provide interesting results based on republication of a popular blog service.

It is expected that the proposed methods will be employed for developing new analytical tools for the specialized users such as company marketers, politicians, news reporters and online writers. The approach can also be applied to various different social network services, since most of the services such as Twitter and Facebook provide open application programming interfaces (API), which can be used for collecting social interaction history among users.

For our future work, we consider information diffusion analysis from an ego-centric network to the sociocentric network to discover information diffusion patterns. Furthermore, it can be a challenge to extend the analysis results to the cases of multiple interconnected services like the blogosphere.

References

[1] J. Leskovec, L.A. Adamic, and B.A. Huberman, “The dynamics of viral marketing,” ACM Trans. Web, vol.1, no.1, pp.1–39, May 2007.

[2] L.A. Adamic and N. Glance, “The political blogosphere and the 2004 U.S. election: Divided they blog,” Proc. 3rd Int. Work. on Link Disc., pp.36–43, New York, NY, 2005.

[3] P. Sriprasertsuk and W. Kameyama, “Information distribution analysis based on human’s behavior state model and the small-world network,” IEICE Trans. Inf. & Syst., vol.E92-D, no.4, pp.608–619, April 2009.

[4] M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummad, “Measuring user influence on twitter: The million follower fallacy,” Proc. 4th Int. Conf. on Weblogs and Social Media, pp.10–17, Washington, DC, 2010.

[5] T.Y. Berger-Wolf and J. Saia, “A framework for analysis of dynamic social networks,” Proc. 12th ACM Int. Conf. on Knowl. Disc. and Data Mining, pp.523–528, New York, NY, 2006.

[6] W.M.P. van der Aalst, Process mining: Discovery, conformance and enhancement of business processes, Springer, Berlin, 2011.

[7] S. Wasserman and K. Faust, Social Network Analysis: Methods and applications, Cambridge University Press, Cambridge, 1994.

[8] W.M.P. van der Aalst and M. Song, “Mining social networks: Uncovering interaction patterns in business processes,” Lect. Notes Comput. Sci., vol.3080, pp.244–260, June 2004.

[9] W.M.P. van der Aalst, T. Weijters, and L. Maruster, “Workflow mining: Discovering process models from event logs,” IEEE Trans. Knowl. Data Eng., vol.16, no.9, pp.1128–1142, July 2004.

[10] C. Günther and W.M.P. van der Aalst, “Fuzzy mining – adaptive process simplification based on multi-perspective metrics,” Lect. Notes Comput. Sci., vol.4714, pp.328–343, Sept. 2007.

[11] W.M.P. van der Aalst, T. Weijters, A.J.M.M. Weijters, B.F. van Dongen, A.K. Alves de Medeiros, M. Song, and H.M.W. Verbeek, “Business process mining: An industrial application,” Inf. Syst., vol.32, no.5, pp.713–732, July 2006.