LETTER Special Section on Data Engineering and Information Management Analyzing Information Flow and Context for Facebook Fan Pages*

Kwanho KIM[†], Josué OBREGON^{††}, Nonmembers, and Jae-Yoon JUNG^{††a)}, Member

SUMMARY As the recent growth of online social network services such as Facebook and Twitter, people are able to easily share information with each other by writing posts or commenting for another's posts. In this paper, we firstly suggest a method of discovering information flows of posts on Facebook and their underlying contexts by incorporating process mining and text mining techniques. Based on comments collected from Facebook, the experiment results illustrate how the proposed method can be applied to analyze information flows and contexts of posts on social network services. *key words: information flow, context analysis, process mining, text mining, Facebook*

1. Introduction

Recently, social network services such as Facebook and Twitter are increasingly growing as an online platform where individuals and organizations post and share information such as news, opinions and advertisements in a fast and easy manner. The original posts are notified to the subscribers of the sites, and the comments on the post let the information be delivered to the friends of the subscriber again. Since comments on a post reflect how information is actually propagated from and to people, discovering hidden information flows and their underlying contexts from commenting behaviors of people is a significant issue of information diffusion analysis on social networks. Sophisticated understanding about information propagation mechanisms that people actually possess can be achieved by information flows and contexts, which accelerates many social services such as viral marketing [1], social suggestion [2], and influence analysis [3].

Specifically, this research focuses on *Facebook fan pages* (FFPs) which are to connect people with the same interest. A FFP includes many posts each of which is composed of a textual description and comments, as shown in Fig. 1. The description of a post represents its specific *context* while its comments contain the *information flow* of the context based on people's interactions in chronological order. Here, the sequence of ordered comments on a specific

Manuscript revised October 29, 2013.

[†]The author is with Incheon National University, Incheon, 406–772, Republic of Korea.

^{††}The authors are with Kyung Hee University, Yongin, Gyeonggi-do, 446–701, Republic of Korea.

*This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MISP) and the Ministry of Education (Nos. 2013R1A2A2A03014718 and 2012R1A6A3A01038452).

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

DOI: 10.1587/transinf.E97.D.811



Fig. 1 Information flow and context of a post on a Facebook fan page.

post is called the *trace* of comments on the post.

There are many studies on information flows based on social interactions such as calculation of influences [4], analysis on social network structures [5], [6], and discovery of information diffusion processes [7]. However, their major drawback is the lack of ability to suggest the corresponding context with an information flow. Moreover, the previous methods cannot capture the dynamic nature of information flows since they heavily rely on the static links between people, called friendship, rather than focusing on the dynamic information propagation established by people for a context.

In this paper, we present an integrated method of analyzing the context from comments on a FFP, as well as discovering meaningful information flows from the FFP. While the previous method deals with only information flow discovery from re-posting behaviors of people [8], the integrated method proposed in this paper first suggests application of trace clustering to finding several meaningful information flows and then reveals the underlying context of each flow based on frequent keywords. In summary, by using the people's comments on posts, the proposed method can discover information flows that imply the actual relations among people, as well as their underlying contexts simultaneously. Specifically, process mining and text mining techniques are incorporated to achieve the goal of our method.

The experiment results based on comments collected in a FFP demonstrate that our method is able to suggest a yet another viewpoint to understand information propagation among people through revealing hidden information flows and their contexts.

Manuscript received July 4, 2013.

2. Analysis on Information Flows and Contexts

The proposed method is composed of five steps as shown in Fig. 2. Based on the traces of a FFP obtained in the data acquisition step, similar traces are grouped into a trace cluster in the trace clustering step. Then, for each trace cluster, its information flow is discovered by detecting frequently observed sequences of people across the comments related to the information flow, and the discovered information flow is visualized in a graphical manner. In the meantime, the context for each information flow is investigated by means of keywords based on the post descriptions associated with the information flow.

2.1 Information Flow Discovery

Before discovering information flows, similar traces are needed to be clustered since the trace of a single post only reveals a particular case of information propagation rather than the radical flows repeatedly observed across posts. Accordingly, we conduct clustering analysis of traces by employing the agglomerative hierarchical clustering method which heuristically searches for clusters in a hierarchical tree [9]. A trace is represented as a vector of binary values which means the appearance sequence of people in the trace, and the similarity between traces is measured by using Jaccard coefficient. At the initial stage, each trace is set to be a trace cluster itself, and, two trace clusters are iteratively combined into a single trace cluster if they are the most similar pair, until they do not have any trace cluster to be combined.

Subsequently, the information flow of a trace cluster, which successfully explains the interaction patterns among people across the posts in the trace cluster, is discovered. Since the entire relations among people are very complex, information flows are desired to be abstracted by focusing on some important flows. Hence, we adopt the fuzzy mining that discovers meaningful sequences of activities by uti-



Fig. 2 Flow chart of the proposed method.

lizing an adaptive graph simplification especially for less structured flows [10]. The significance between people is investigated based on how much two people are sequentially correlated with each other. According to the significances of people, some are grouped if they are less significant compared to others.

2.2 Context Analysis

To extract keywords according to information flows, the textual descriptions of the posts associated with an information flow are aggregated into a single text document. Then, the importance of a word in a text document is examined in terms of term frequency (TF) and inverse document frequency (IDF) [11]. Particularly, as the importance metric of a word, we use a TF-IDF weighting scheme which is widely applied in many text mining applications such as information retrieval and text summarization.

In this research, a word is considered a keyword of an information flow, if the word frequently appears in the descriptions associated with the information flow and rarely appears in the descriptions associated with the other information flows. After calculating the importance values of all the words for each information flow, top ranked words are regarded as the keywords of the information flow.

3. Experiment Results

For the experiments, we used a FFP of CNN (https://www. facebook.com/cnn) which publishes news articles, and more than 0.2 million people share opinions by making comments on the posts in the FFP. Posts published between May 17, 2013 and May 31, 2013 were gathered by utilizing the developed software shown in Fig. 3. To reduce noises, we removed comments of the users who appeared less than eight times, and a total of 297 posts and 22,297 comments were finally involved in the experiments. To cluster traces and discover information flows, a process mining tool ProM was used [12].

Data Ana	lysis Al	bout.											
								1.2.22					
in Page Id	cnn						Loas	Acces Token	Cuesconiniano	262920	SG2CHOE	JUESCMU	WHNAZUZU
Name	CNN					c	ategory	Medialnews/publis	hing		Likes	491745	17
Gathering data					Process	Process mining analysis			Novelty analysis				
Choose one way to analize the Fan Page									And the second se				
Distant	Dente		tange of date						Select the pa	ameres	5 10 4545	ze me nov	eny of the ran .
C Laicol	Puoto		building Barton						tiscale		III MIN	IUTE	-
sumber of F	osts		Install Date	2013.05.17		Like thresh	old	20	-				
Type	P	-	Final Date	2013 05 31		Comment t	hreshold	[20]	HIStorical use	HS AL	Posts		
			÷										
Cat Data				Constants MYMI file				Get and Analyze Doste					
			001000			- OCTAC	cic noone.			OCC DIN	Pontagen	0.00	-
No Cro	nto d li mu				00	etxat			Tune	1 and	Common	0.4	
1 Dritter	Fo May 310E 0 A shud-finds that more and more manied mothers						s are out-earning their husbands. Peggy Drexter.			402	1	00	
2 Fri Mar	En May 31 03.5 Authorities intercented two speedboots - sinking one -									2290	4	- 8-	100
3 Fri May 31 03 0 Glearning high-tech cities are being planned across Africa.							rica. Some say they are the continent's futur .			2277	1	2	-
4 Fri May 31 02 3 BREAKING NEWS: Arvind Mahankali, a 13-year-old						from Bauside Hills. New York, won the Scripps			STATUS	6572	1	100	
5 FnMa	"One I	evito empowers	nent and an is	sue that's	a personal priority for me - is letting women dec.			K. UNK	913	1	121		
6 Fri May 31 01 3. Are you looking forward to the Emmys? The 40th Annual Davlime Emm								ts will air five on HL	N. UNK	170	1	1	
7 Fri May 31 01 0 . A Tudor warship sunk of the English coast more than 400 years apo will go on display in a new m									n., LINK	348	1	R	
8 Fri Ma	3100:0	The pl	ight of a 10-year	old girl who cou	d die withi	n weeks without	a lung tran:	plant has sparked	a LINK	678	1	2	
9 Thu M	ay 30 23. Winning the National Spelling Bee can lead to great experiences as well as disappointing letdown.									195	1	2	
10 Thu M	ay 30 22	Aeros	Aerosmith and Jason Aldean are among several artists slated to perform at tonight's Boston Stron STATUS 2704 1 😰								8		
11 Thu M	ay 30 22	_ JCPer	mey is setting a	new teapot that a	oparently.	looks a lot like Ad	tolf Hiller. T	he billboard in Calif	or. UNK	823	1	R	
12 Thu M	ay 30 21.	Have 3	rou visited Wyon	ing's expansive	Grand Tetr	on National Park	See the ra	nger-recommende-	d., LINK	282	1	R	
13 Thu M	ay 30 21.	What i	to you think abo	ut standardized t	ests for ad	missions? See v	vhy a group	of 15 New York City	UNK	314	1	X	
14 Thu M	ay 30 20	Defen	ders say its an i	a. JUNK	160	1							
15 Thu M	ay 30 19	BREA	BREAKING NEWS: Parts of central Oldahoma are under a tornado warning, with forecasters urging.								1	R	
16 Thu M	ay 30 19	Was s	Was she framed? An Arizona mom accused of drug smuggling says she is innocent http://on.cnn?								1	R	
17 Thu M	ay 30 19.	- When	When the head of the World Health Organization says a new virus is her "greatest concern right no [LINK								1	R	
18 Thu M	ay 30 18	Can yi	ou imagine a wo	nt. LINK	375	1	×						
	ALC: 100 100	OThe of sheet	dillors in this my	they of a foundry	A incase with	to many marchine and many	ferrers clin.com	for the main manen	te it it.it.	1580	4		1.00

Figure 4 shows the dendrogram of trace clustering for

Fig. 3 Screenshot of the software developed for gathering posts and comments on FFPs.



Fig. 4 Dendrogram of traces extracted by trace clustering.



Fig. 5 Information flow of cluster C discovered from CNN's FFP.

the gathered posts and comments. The graph visualizes similar posts in terms of information flows of their comments, and the lines on the graph represent the results of iterative clustering steps. A clustering threshold that determines the number of trace clusters was empirically chosen. After removing trace clusters with less than three traces, four trace clusters were remained. The clusters A, B, C and D contained 179, 10, 104, and 4 traces, respectively.

For each trace cluster, we discovered an information flow which representing how information was propagated among people in the FFP. Figure 5 depicts the information flow discovered from cluster C. A square box shows a person and his/her significance in the information flow, and an octagon box is a group of people whose significances were marginal. Three types of roles can be found in the view point of information flow such as: (i) information source

 Table 1
 Top 5 keywords of four clusters and their TF-IDF values.

clusters	# of traces	top 5 keywords
A	179	awards (13.86), found (12.80), bangladesh (9.70), kelly (9.70), millions (9.70)
В	10	polaroid (4.16) , hospitals (2.77) , education (2.39) , storm (2.39) , technology (2.39)
С	104	tsarnaev (13.86), amanda (10.40), tamerlan (10.40), berry (8.72), ramsey (8.72)
D	4	collins (2.77), jason (2.77), player (2.77), NBA (1.39), after (1.15)



Fig. 6 Visualization of the underlying context of cluster C.

which initiates information flows, (ii) information mediator which connects between people or groups, and (iii) information sink which mainly consumes information and hardly introduces subsequent reactions of the others.

At last, the keywords of each cluster, which represent its underlying context, were analyzed based on the aggregated descriptions of the posts associated with its information flow. We removed stop-words to eliminate noises. Top 5 keywords for four clusters are described in Table 1, in which their TF-IDF values are presented in the parentheses. In addition, the discovered context can be visualized in the word cloud based on the TF-IDF values of the corresponding keywords.

For instance, the keywords of cluster C in the table indicate that its information flow shown in Fig. 5 was mainly about the victims of Boston bombing. It is expected that the information related with Boston bombing is propagated among people. First, in Fig. 5 it was found that people such as "Andrew carter" and "Elizabeth Sugg" mainly initiated information flows about Boston bombing. Second, people such as "Lanora McNeil" and "Molly Diaz" diffused the related information to others by connecting between groups of people. Finally, a person, "Glenda Dean Paul", was observed to frequently finalize the information flow about Boston bombing, indicating that people rarely reacted to her comments about the event.

4. Conclusion and Discussion

In this research, we developed a method that aims not only to discover information flows based on trace clusters by combining similar traces but also to analyze their underlying contexts. The experiment results illustrated how the proposed method could be applied to the understanding of the information propagation of posts on FFPs. It is believed that the method allows Web analysts to figure out who are interested in the posts on the site and how the information of the post are propagated on the social networks based on the comments on FFPs.

The proposed method can also be widely adopted for most social network services just as illustrated for Facebook. It is because the method depends only on information related to comments such as user sequences and textual descriptions which are provided by the social network services.

For the future work, we plan to further extend the proposed method to accommodate additional information of social networks such as friend relations and users' profiles.

References

- [1] P. Domingos, "Mining social networks for viral marketing," IEEE Intell. Syst., vol.20, no.1, pp.80–82, Jan. 2005.
- [2] Q. Li, J. Wang, Y.P. Chen, and Z. Lin, "User comments for news recommendation in forum-based social media," Inf. Sci., vol.180, no.24, pp.4929–4939, Dec. 2010.
- [3] 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,

May 2010.

- [4] X. Jin, C. Wang, J. Luo, X. Yu, and J. Han, "LikeMiner: A system for mining the power of 'Like' in social media networks," Proc. 17th Int. Conf. on Knowledge Discovery and Data Mining, pp.753–756, San Diego, CA, Aug. 2011.
- [5] 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, Sept. 2004.
- [6] 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.
- [7] 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.
- [8] K. Kim, J.-Y. Jung, and J. Park, "Discovery of information diffusion process in social networks," IEICE Trans. Inf. & Syst., vol.E95-D, no.5, pp.1539–1542, May 2012.
- [9] J.-Y. Jung, "PROCL: A process log clustering system," J. Society for e-Business Studies, vol.13, no.2, pp.181–194, May 2008.
- [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] N. Zhong, L. Yuefeng, and W. Sheng-Tang, "Effective pattern discovery for text mining," IEEE Trans. Knowl. Data Eng., vol.24, no.1, pp.30–44, Jan. 2012.
- [12] W.M.P. van der Aalst, T. Weijters, A.J. Weijters, B.F. van Dongen, A.K. Medeiros, M. Song, and H.M.W. Verbeek, "Business process mining: an industrial application," Inf. Syst., vol.32, no.5, pp.713– 732, Sept. 2006.