User Profile Extraction from Twitter for Personalized News Recommendation

Won-Jo Lee*, Kyo-Joong Oh*, Chae-Gyun Lim**, and Ho-Jin Choi*

*Department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST)
291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea

**Department of Computer Engineering, Kyung Hee University
1-Seocheon-dong, Gyeonggi-do, Yongin-si 446-701, Republic of Korea

mochagold@kaist.ac.kr, aomaru@kaist.ac.kr, rayote@khu.ac.kr, hojinc@kaist.ac.kr

Abstract— **Extracting personal profiles from various sources such as purchased items, watched movies, mailing records, etc. is important for recommender systems. For personalized news recommendation, in particular, existing methods mostly utilize information obtainable from the news articles read by the users such as titles, texts, and click-through data.**

This paper aims to investigate a different method to build personal profiles using the information obtained from Twitter to provide personalized news recommendation service. For a Twitter user, our method utilizes tweets, re-tweets, and hashtags, from which important keywords are extracted to build the personal profile.

The usefulness of this method is validated by implementing a prototype news recommendation service and by performing a user study. Using a simple cosine similarity measure, we compare the differences among the user profiles, and also among the recommended news lists, in order to check the discriminative power of the proposed method. The prediction accuracy of news recommendation is measured against a small group of users.

Keywords— **Personalized News Recommendation, User Profile, Twitter, tweets/re-tweets, hashtags**

I. INTRODUCTION

Personalized recommender systems make use of previous history of user transactions from various sources such as purchased items, watched movies, mailing records, etc. to predict what the user would like to purchase in the near future [1]. Basically, two approaches have been widely used to build a recommender system: contents-based analysis and collaborative filtering, and in many applications in general, collaborative filtering shows better performance than contents-based approach [2].

For a particular application such as news recommendation, however, contents-based analysis seems to be better suited in the sense that the recommender system does not have time to wait for collecting information about which news articles have been read by other similar users, because the new upcoming news articles are updated and replaced by newer articles every now and then [3], [4]. Moreover, collaborative filtering suffers from the so-called "cold start" problem for newly registered users with little history [5], for which situation contents-based analysis would also become a good viable alternative.

This paper introduces a variant of contents-based analysis for news recommendation. Existing methods of personalized news recommendation mostly utilize information obtainable from the actual news contents read by the users such as titles, texts, and click-through data [3], [6], [7] whereas few approaches attempt to do the job without utilizing the information about the news contents [4]. In this paper, we build personal profiles using the information obtained from the Twitter usage, instead of the news contents. Given a Twitter user, our method utilizes tweets, re-tweets, and hashtags, from which important keywords are extracted to build the personal profile, which is then used for selecting new, upcoming news articles.

We validate the usefulness of this method by implementing a prototype news recommendation service and by performing a user study against a small group of users.

II. RELATED WORK

Many reports suggest the high relevance of utilizing the information about Twitter usage for news recommendation. Kwak et al. [8] reports that over 80% of topics mentioned in tweets have some relationships with news. Lerman and Rumi [9] reports that propagation of Twitter is faster than traditional news, because users having many followers can influence by large on the news propagation. Sakaki et al. [10] reports an approach of analyzing messages passing through Twitter to predict earthquakes and typhoons.

Many approaches exist for building personal profiles for news recommendation. Carreira et al. [7] propose an approach to build user profiles based on user ratings on the news article, in the way similar to the traditional item ratings. Wang et al. [6] propose an adaptive user profiling model to apply collaborative filtering on the news lists read by a similar group of users, while treating news as items in the traditional way. Phelan et al. [3] investigate an approach to utilize Twitter to recommend real-time topical news, where a user profile consists of terms-articles calculated using TF-IDF.

Our approach in this paper can be seen as an extension to Phelan et al. [3] in that we use topic models as well as TF-IDF. Topic modelling [11] has been traditionally known to be a technique for analyzing a large volume of documents, but

more recently, attempts are made to apply the technique for dealing with small-sized documents such as Twitter and microblogs [12], [13]. In this paper, we will use topic modelling when comparing two small sets of news articles selected.

III. Proposed Method

A news recommendation process is generally assumed to be of two phases: (1) user profiling, and (2) news ranking. In this paper, we build user profiles by extracting information from Twitter, then select and rank news articles based on the user profiles. Figure 1 shows the overall process of our approach with some details. Using API provided by Twitter [14], timelines of the users are collected automatically. User documents (i.e., tweets and re-tweets) collected in this way are processed for identifying important keywords preferred by each user to form a profile. Such a profile is used as the basis for compiling a ranked list of candidate news articles from the set of new, upcoming news. Detailed steps of the two phases will be presented in the following subsections.

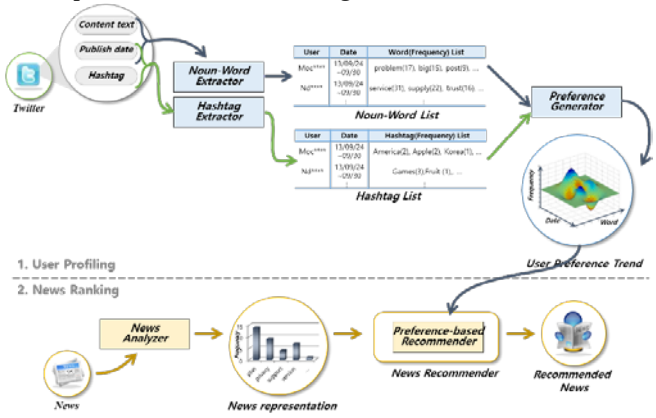

Figure 1. Overall process of news recommendation

A. User Profiling Phase

We take the bag-of-words approach for user profiling. That is, a user profile is compiled from the bag of words extracted from the tweets/re-tweets written by the user. This phase proceeds in the following steps. First, unnecessary information such as website links, emoticons, and twitter ID's are all removed from the original tweets/re-tweets. Second, further "noisy" words (e.g., prepositions, adjectives, adverbs, etc.) are filtered out to obtain the "proper" sentences or phrases. Third, noun phrases are collected from the remaining text.

In addition, we consider hashtags explicitly mentioned by the user. Hashtag in Twitter is a word prefixed by symbol "#", providing a means of grouping to help searching tweets of particular interests. In our approach, we treat hashtags as valuable information to capture a user's profile.

With noun-word and hashtag lists, user profile is compiled (see Figure 2). A user profile P(u) is defined as

$$\mathrm{P(u)} = \{(Word, Frequency) | Word \in \{words, hashtags\}\}$$

where Frequency represents the number of words appeared in user u's tweets. Since hashtags are important keywords selected by the user, we assign higher frequency values to the hashtags than other ordinary words.

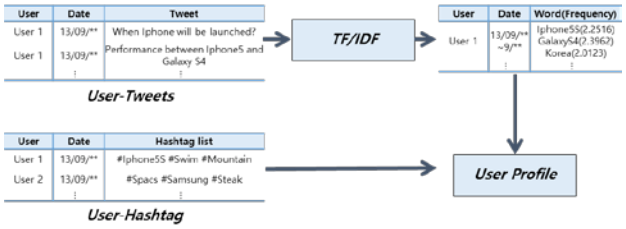

Figure 2. Example of user profile extraction from Twitter

The user profiles obtained from the noun-word and hashtag lists (illustrated by Figure 3) are normalized to allow fair comparison between users and/or news (illustrated by Figure 4).

User 1		User 2		User 3	
Keyword	Frequency	Keyword	Frequency	Keyword	Frequency
Google	3	Twitter	6	Iphone	5
Running	3	Dad	4	Game	5
Apple pie	2	Windows	3	Twitter	4
President	2	Samsung	2	Samsung	3
	…		…		…

Figure 3. Example of user profiles before normalization

Keyword	Apple	Samsung	Google	Twitter	…
User 1	0.01744	0.01161	0.02037	0.01381	…
User 2	0	0.00013	0.00019	0.00034	…
User 3	0.00108	0.00111	0	0.00212	…
User 4	0.00777	0.01576	0.01176	0	…
…	…	…	…	…	…

Figure 4. Example of user profiles after normalization

After the normalization process, a user profile P(u) is re-defined as

$$\mathrm{P(u)} = \{(Word, Weight) | Word \in \{words, hashtags\}, \; 0 < weight \leq 1\}$$

where Weight represents normalized word frequency based on TF-IDF.

B. News Ranking Phase

User profiles constructed as above are used as the basis for compiling a ranked list of news articles from the set of new, upcoming news. For each user, every news article in the set is evaluated in its similarity to the user profile. This procedure consists of the following steps. First, "noisy" words are filtered out from the article to obtain the "proper" sentences. From each news article, the title and the text content remain, but author information, date, and E-mail address are removed. Second, noun phrases are collected from the remaining text. Third, TF-IDF scores for the keywords are computed to form a "profile" of the article. Last, the similarity score between this "article profile" and the user profile is computed. In our implementation, we used cosine similarity for this purpose.

Given a user, all the news articles in the set of upcoming news are ranked based on such similarity scores (illustrated by Figure 5). Finally, top-k news articles are recommended to the user.

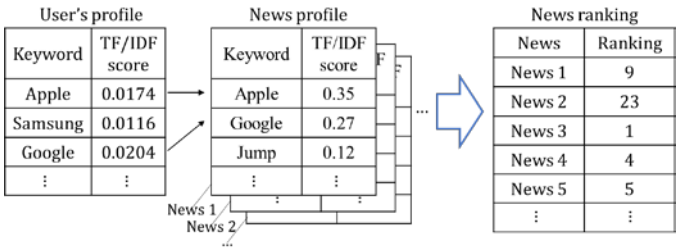

Figure 5. News ranking with user profiles

IV. EXPERIMENT

To validate the usefulness of the proposed method, we have implemented a prototype news recommendation service and performed a simple user study against a small group of users. Using the Twitter API provided in the Twitter website [14], we have collected the timelines of eight users who agreed the user study and gave us permission for collecting their data. As for the new news articles to service to the users (we call them "upcoming news list"), we collect them from the Daum media [15]. About 3,000 new news articles were collected each day for this purpose.

In order to check the discriminative power of our method, we have compared the differences among the user profiles, and also among the recommended news lists, using a simple cosine similarity measure. The prediction accuracy of news recommendation is roughly measured by asking the users directly about their preference of choosing which news to read. Details of the experimental results are presented in this section.

A. Data Collection

Table 1 summarizes the basic statistics of the Twitter usage (i.e., sending tweets/re-tweets) of the eight users during the three month period of our experiment. On average, a user in our study writes a Twitter message about 46 times per month. For the sake of extracting important keywords for profiling, it is important for each user to write substantial amount of texts in the Twitter messages. According to the table, however, user u1 writes tweet/re-tweet only about 11 times per month, seemingly not enough for user profiling. Though this usage is far less than average, this user tends to write relatively longer tweets, allowing us to obtain a user profile comparable to others.

TABLE 1. NUMBER OF TWITTER USAGE (I.E., SENDING TWEETS/RE-TWEETS)

User	Month1 (Sep 2013)	Month2 (Oct 2013)	Month3 (Nov 2013)	Average
u1	29	5	0	11.3
u2	26	81	48	51.7
u3	51	37	34	40.7
u4	19	27	14	20
u5	14	33	20	22.3
u6	189	70	46	101.7
u7	51	23	60	44.7
u8	0	34	40	24.7
Average	55.8	44.0	38.9	46.2

B. Difference among the User Profiles

For the eight users, we have constructed the user profiles, each represented in a vector of normalized weighted keywords as explained in Section III. To assess the discriminative power of the profiles, we compare their pairwise difference using the inverse of their cosine similarity. Table 2 shows the result of this comparison, indicating that the user profiles are basically distinctive enough to each other.

TABLE 2. DIFFERENCE BETWEEN USER PROFILES (%)

User Profile	P(u1)	P(u2)	P(u3)	P(u4)	P(u5)	P(u6)	P(u7)	P(u8)
P(u1)	-	99	98	99	91	98	97	99
P(u2)	99	-	96	98	95	98	97	95
P(u3)	98	96	-	98	93	94	89	94
P(u4)	99	98	98	-	94	95	97	98
P(u5)	91	95	93	94	-	94	94	96
P(u6)	98	98	94	95	94	-	96	92
P(u7)	97	97	89	97	94	96	-	96
P(u8)	99	95	94	98	96	92	96	-
Average	97.28	96.85	94.57	97	93.85	95.28	95.14	95.71

C. Difference among the Recommended News Lists

Given a user and the set of 3,000 upcoming news articles daily, our implemented news ranking algorithm evaluates the similarity of each article to the user's profile and selects top 10 articles (out of 3,000) for final recommendation.

In order to see the diversity of recommended news articles from user to user, we assess the pairwise difference among the recommended lists (again using the cosine similarity). For the fairness and effective measuring of the similarity among the sets of news articles, we first obtain topic distribution of each news set using LDA [11], then measure the similarity between two sets by their topic distribution for all possible pairs.

Table 3 shows the result of this comparison, indicating that the recommended news articles are quite distinctive from user to user. Since the performance of the LDA topic modeling may vary by the initial setting of the number of topic clusters, we have repeated the same experiments using different settings of 10, 20, and 30 topic clusters. Table 4 shows the results of averaged difference among all pairs of users for five days, indicating that the diversity of news recommendation over different users remains similar for different settings of topic clusters.

TABLE 3. DIFFERENCE BETWEEN RECOMMENDED NEWS ARTICLES (%) – FOR A PARTICULAR DAY

News Articles	N(u1)	N(u2)	N(u3)	N(u4)	N(u5)	N(u6)	N(u7)	N(u8)
N(u1)	-	98	98	99	99	2	99	98
N(u2)	98	-	93	93	98	90	96	0
N(u3)	98	93	-	97	96	91	97	92
N(u4)	98	93	97	-	99	93	98	98

N(u5)	98	98	96	99	-	96	98	94
N(u6)	98	90	91	93	96	-	95	89
N(u7)	98	96	97	98	98	95	-	98
N(u8)	98	0	92	98	94	89	98	-
Average	98	78.3	94.3	96.3	96.8	92.3	97	78.5

TABLE 4. AVERAGED DIFFERENCE BETWEEN RECOMMENDED NEWS ARTICLES (%) – FOR FIVE DAYS

Number of topic clusters	Day 1	Day 2	Day 3	Day 4	Day 5
10	96	90	98	96	93
20	97	94	98	97	96
30	97	94	98	98	98

D. Prediction Accuracy of News Recommendation

For assessing the qualitative performance of the proposed method, we have measured the prediction accuracy of news recommendation by a user study against a small group of users. Since there is no standard guideline for this kind of qualitative assessment, we adopt a comparative study between our method and random selection as follows. Basically, each user will receive two sets of 10 news articles – one set generated by random selection, and one set recommended by our news ranking algorithm based on the user's profile. When given to the user, these two sets are mixed together into a bag of 20 news articles, so that the user does not know which news article comes from which set. Then, for each news article, the user checks whether he/she would like to read this article or not. This study was repeated for five consecutive days, with about 3,000 upcoming news articles renewed daily, against a small group of eight users participated.

Tables 5 and 6 summarize the result of this study. Table 5 shows the daily hit ratios for the random selection (averaged over the eight users), and Table 6 the daily hit ratios for the recommended list (averaged over the eight users). The results show that the proposed method achieves better performance than random selection in terms of the hit ratios.

TABLE 5. HIT RATIO (%) – RECOMMENDED AT RANDOM

User	Day 1	Day 2	Day 3	Day 4	Day 5
u1	20	10	10	30	30
u2	40	60	30	30	50
u3	30	20	10	40	40
u4	30	40	60	50	60
u5	30	30	30	70	60
u6	40	50	40	40	40
u7	30	0	10	0	30
u8	60	70	40	50	50
Average	35	35	28	39	45

TABLE 6. HIT RATIO (%) – RECOMMENDED USING USER PROFILES

User	Day 1	Day 2	Day 3	Day 4	Day 5
u1	70	70	20	60	40
u2	60	70	70	10	40
u3	80	10	50	50	20
u4	90	80	40	70	70
u5	60	60	40	70	70
u6	70	70	70	50	70
u7	50	20	20	0	10
u8	60	40	60	40	50
Average	67.5	52.5	46.3	43.8	46.3

V. CONCLUSION

This paper has proposed a method to build personal profiles using information obtainable from Twitter for personalized news recommendation. The method utilizes tweets, re-tweets, and hashtags, from which important keywords are extracted to build the personal profile.

The usefulness of the method has been validated by a user study experimented over a prototype news recommendation service. The discriminative power of the method has been shown by examining the differences among the user profiles, and also among the recommended news lists. The prediction accuracy has been measured in terms of hit ratios against a small group of users.

ACKNOWLEDGMENT

This work was supported by Samsung Electronics Co. Ltd.

REFERENCES

[1] Y. J. Park and K. N. Chang, "Individual and group behavior-based customer profile model for personalized product recommendation," *Expert Systems with Applications*, vol. 36, no. 2, pp. 1932-1939, 2009.

[2] P Drineas, K Iordanis, and R Prabhakar, "Competitive recommendation systems," *in Proc. ACM STOC'14*, pp. 82-90, 2002.

[3] O Phelan, K McCarthy, and B Smyth, "Using twitter to recommend real-time topical news," *in Proc. ACM Recsys'09*, pp. 385-388, 2009.

[4] F Abel, Q Gao, GJ Houben, K Tao, "Analyzing user modeling on twitter for personalized news recommendations," *User Modeling, Adaption and Personalization*. Springer Berlin Heidelberg, pp. 1-12, 2011.

[5] AI Schein, A Popescul, LH Ungar, and DM Pennock, "Methods and metrics for cold-start recommendations." *in Proc. ACM SIGIR'02*, pp. 253-260, 2002.

[6] J Wang, Z Li, J Yao, Z Sun, M Li, and W Ma, "Adaptive user profile model and collaborative filtering for personalized news," *Frontiers of WWW Research and Development-APWeb'06*. Springer Berlin Heidelberg, pp. 474-485, 2006.

[7] R Carreira, JM Crato, D Gonçalves JA Jorge, "Evaluating adaptive user profiles for news classification," *in Proc. ACM IUI'04*, pp. 206-212, 2004.

[8] H Kwak, C Lee, H Park, and S Moon, "What is Twitter, a social network or a news media?," in *Proc. WWW'10*, pp. 591-600, 2010.

[9] K Lerman, G Rumi, "Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks," *in Proc. ICWSM'10*, pp. 90-97, 2010.

[10] T Sakaki, O Makoto, and M Yutaka, "Earthquake shakes Twitter users: real-time event detection by social sensors," in *Proc. WWW'10*, pp. 851-860, 2010.

[11] DM Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no.4, pp. 77-84, 2012.

[12] L Hong, and D Brian, "Empirical study of topic modeling in twitter," in *Proc. SOMA'10*. ACM, pp. 80-88, 2010.

[13] D Ramage, ST Dumais, and DJ Liebling, "Characterizing Microblogs with Topic Models," *ICWSM'10*, 2010.

[14] (2013) Twitter API. [Online]. Available: http://dev.twitter.com

[15] (2013) Daum Media, [Online]. Available: http://media.daum.net

Won-Jo Lee
He received a bachelor's degree of computer science in 2012 from Hanyang University. He is currently a MS candidate student in the department of computer science at Korea Advanced Institute of Science and Technology.

Kyo-Joong Oh
He received a bachelor's degree of computer science in 2011 from Korea Advanced Institute of Science and Technology. He is currently a MS/Ph. D. candidate student in the department of computer science at Korea Advanced Institute of Science and Technology.

Chae-Gyun Lim
He received a bachelor's degree of medical computer science in 2011 from Eulji University. MS candidate student in the department of computer engineering at KyungHee University.

Ho-Jin Choi is currently an associate professor in the Dept. of Computer Science at KAIST. In 1982, he received a BS in Computer Engineering from Seoul National University, Korea, in 1985, an MSc in Computing Software and Systems Design from Newcastle University, UK, and in 1995, a PhD in Artificial Intelligence from Imperial College, London, UK. From 1982 to 1989, he worked for DACOM, Korea, and between 1995 and 1996, worked as a post-doctoral researcher at Imperial College. From 1997 to 2002, he served as a faculty member at Korea Aerospace University, Korea, then from 2002 to 2009 at Information and Communications University (ICU), Korea, and since 2009 he has been with the Dept. of Computer Science at KAIST. Between 2002 and 2003, he visited Carnegie Mellon University (CMU), Pittsburgh, USA, and has been serving as an adjunct professor of CMU for the program of Master of Software Engineering (MSE). Between 2006 and 2008, he served as the Director of Institute for IT Gifted Youth at ICU. Since 2010, he has been participating in the Systems Biomedical Informatics National Core Research Center at the Medical School of Seoul National University. Currently, he serves as a member of the boards of directors for the Software Engineering Society of Korea, for the Computational Intelligence Society of Korea, and for Korean Society of Medical Informatics. His current research interests include artificial intelligence, data mining, software engineering, and biomedical informatics.