World of Computer Science and Information Technology Journal (WCSIT)
ISSN: 2221-0741
Vol. 1, No.8, 333-338, 2011

Improving Business Type Classification from Twitter Posts Based on Topic Model

Chanattha Thongsuk
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand

Choochart Haruechaiyasak
Human Language Technology Laboratory
National Electronics and Computer Technology Center (NECTEC)
Bangkok, Thailand

Somkid Saelee
Faculty of Technical Education
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand

Abstract— Today Twitter, a social networking website, has become a new advertising channel to promote products and services using online social network community. In this study, we propose a solution to recommend Twitter users to follow businesses, which match their interests. Our approach is based on classification algorithms to predict user's interests by analyzing their posts. The challenging issue is the short length characteristic of Twitter posts. With only a few available key terms in each post, classifying Twitter posts is very difficult and challenging. To alleviate this problem, we propose a technique to improve the classification performance by expanding the term features from a topic model to train the classification models. A topic model is constructed from a set of topics based on the Latent Dirichlet Allocation (LDA) algorithm. We propose two feature processing approaches: (1) feature transformation, i.e., using a set of topics as features and (2) feature expansion, i.e., appending a set of topics to a set of terms. Experimental results of multi-classification showed that the highest accuracy of 95.7% is obtained with the feature expansion technique, an improvement of 19.1% over the Bag of Words (BOW) model. In addition, we also compared between *multi-classification* and *binary classification* using feature expansion approach to build the classification models. The performance of feature expansion approach using *binary classification* yielded higher accuracy than the *multi-classification* equal to 2.3%, 3.3% and 0.4%, for airline, food and computer & technology businesses, respectively.

Keywords: Classification; topic model; Latent Dirichlet Allocation (LDA); Twitter.

I. INTRODUCTION

Web 2.0 is a departure from the traditional web to represent the large Internet social networking and its collectively abundant social contents. Web 2.0 allows people to advertise and follow some neighbors based on personal interests. Advertising on social networking websites is growing and interesting because the information can reach a large group of customers with low overhead cost.

Today many businesses are using ***Twitter***[1], a well-known micro-blogging web site, as a new channel to promote their products and services including related activities. Twitter is a fast-growing micro-blogging site and it is becoming a popular choice to advertise interesting business domain based on personal interests and user profiles.

Twitter provides an attractive platform for advertisers to promote the company's products, services including brand. The customer will receive information and promotion from the companies in which they are following in real time. In addition, the customers can reply with their opinions and also complains to the companies. Today many companies have started to advertise, get feedback from the customers and gain more revenue from Twitter.

The key advantages of Twitter are as follows.

1. There are a large number of members. The number of users on Twitter is very fast growing. Nowadays, Twitter has the worldwide users.
2. Twitter has the user profiles and neighbor's network (call "follow" and "follower" relationships). These information can be used for classifying interesting domain and advertise to appropriate users.
3. It is easy to use and free of charge.
4. Customers can receive quick and direct information from the companies.

The challenges of Twitter are as follows.

1. Each post is a micro blog which has fewer than 140 characters.
2. Most posts are often colloquialism and consist of acronyms.
3. There are a lot of the junk posts.

[1]**Twitter,** *http://www.twitter.com*

Figure 1 illustrates a social networking formation under Twitter consisting of users with different roles with their relationships.

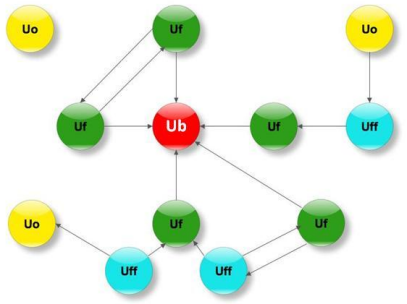


Figure 1. Social networking model on Twitter

U_b is a company in a given business.
U_f is a follower of the company.
U_{ff} is a follower of follower of the company
U_o is another twitter user who does not follow the company or the company's follower.
A → B represents a relation, A is the follower of B
A ←→ B represents a relation, A and B are friends.

Each user can communicate by creating posts. Some examples of Twitter posts are listed in Table 1, 2 and 3

Table 1. Some post examples of airline business

Member Type	Post
Followee (U_b)	JetBlue welcomes up to four small pets onboard each flight with us. Pets are required to remain in their carriers while jetting.
Follower (U_f)	Holiday Travel = BookedJetting to/from Buffalo for the holidays roundtrip for $230 thanks to award travel + voucher. Thanks [@jetblue@]!

Table 2. Some post examples of food business

Member Type	Post
Followee (U_b)	Wifi is now free, one click and unlimited to all US and Canadian stores.
Follower (U_f)	actually not doing horrible.. I even drank a [@starbucks@] Vivianno smoothie -Banana Chocolate. It went down well-soothed the upset stomach.

Table 3. Some post examples of computer & technology business

Member Type	Post
Followee (U_b)	15% off any Dell Outlet Dell Precision? M6500 Laptop! Enter coupon code 2D5KJHB01FQT8 at checkout at . Online only.
Follower (U_f)	[@delloutlet@] what are the models that have at least WSXGA+ res? I don't care about screen size as long as is hi-res

The sample posts which contain relevant terms in each business type would be easy to classify. However most posts do not contain any key terms to help identify the business type. The reason is due to the Twitter policy of allowing only 140 characters per post.

In this paper, we propose the classification framework by using Twitter posts from three business types, i.e., airline, food and computer & technology. We propose two solutions to improve the classification accuracy. The first approach is *feature transformation*, i.e., by using a set of topics as features. The second approach is *feature expansion*, i.e., by appending a set of topics to a set of terms. These feature processing approaches help increase the semantic and expand the key concepts for the feature set used to construct the classification models.

The rest of this paper is organized as follows. In next section, we discuss some related works. In Section III, we present our proposed framework. Section IV presents experiments and results. We conclude the paper in Section V.

II. Related Works

Text Categorization is the task of automatically assigning a set of documents into predefined set of categories. Many related works evaluated different filtering techniques and classification algorithms to improve the accuracy.

Banerjee (2008) proposed a method to improve the classification task by generating the topic model from Wikipedia [1]. Jose Maria Gomez Hidalgo et al. (2006) proposed a method to analyze email spam and block them using extent Bayesian Filtering technique [5]. M. Chau et al. (2008) proposed a method to combine web content analysis and web structure analysis. They also compared web features with two existing web filtering methods [3]. Gabriella Pasi et al. (2007) proposed a new model to filter novel to help users better understand based on multiple criteria defined [11]. N. Churcharoenkrung et al. (2005) suggested URL Filtering using the Multiple Classification Ripple-Down Rules (MCRDR) Knowledge acquisition method [4]. Georgiana Ifrim et al. (2005) proposed a method for mapping every term onto a concept mappings and term sense disambiguation techniques with Naïve Bayes and SVM classifiers [6]. Viet Ha-Thuc et al. (2008) proposed an approach to transforms relevant terms to topics and alleviate the scalability problem and revisit a wide range of well-known [13]. Phan et al. (2008) proposed a method to classify short and sparse texts by some hidden topics [12]. Justin Basilico et al. (2004) proposed the on-line algorithm (JRank) to predict accuracy for different combinations of features on the user and item side and learn a prediction function [2]. Mohammed Nazim et al. (2009) proposed a hybrid recommender system incorporating components from collaborative filtering and content-based filtering by including a diverse-item selection algorithm to select the dissimilar items [14]. Veronica Maidel et al. (2008) proposed a filtering method and examined various parameters by using ontology concepts of user's profiles and items' profiles [10]. Erik Linstead et al. (2007) proposed statistical topic models to extracting concepts form source code [9]. Bernard J. Jansen et al. (2009) investigated micro-blogging containing branding comments, sentiments, opinions and

overall structure of micro-blogs postings [7]. Akshay Java et al. (2007) observed posting of users in Twitter to cluster communities based on the frequency of terms in the user's posts [8].

Our main contribution is to propose two approaches based on (1) *feature transformation* by generating a set of topic probability scores using LDA algorithms and (2) *feature expansion* for selecting terms from the feature selection process and append them into the topic model to improve the classification accuracy.

III. THE PROPOSED FRAMEWORK

The proposed framework of feature transformation and feature expansion to build a short-text classification model for micro-blogging posts is shown in Figure 2.

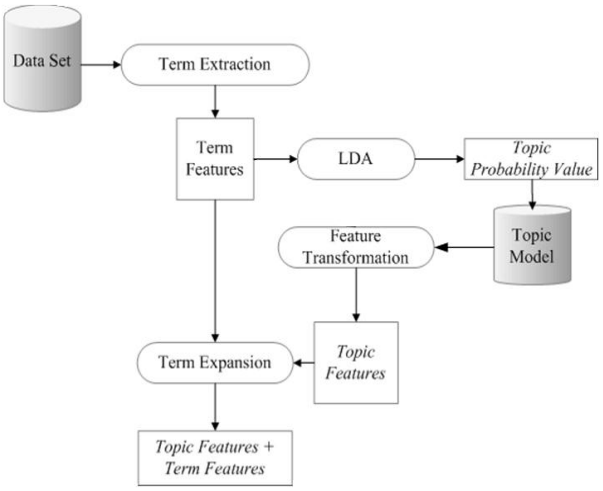


Figure 2. The feature transformation and expansion framework

The first step is to extract term features from posts by applying text processing, i.e., term tokenization, stopword removal and feature selection. We build classification models from different feature sets based on the Bag of Words (BOW) and the topic model. For the topic model, we applied the Latent Dirichlet Allocation (LDA) (Blei et al, 2003) to cluster posts into mixtures of topics.

We explain the topic model based on the Latent Dirichlet Allocation (LDA) as follows.

Given a set of n posts denoted by $P = \{P_1, \ldots, P_n\}$, the LDA algorithm generates a set of k topics denoted by $T = \{T_1, \ldots, T_k\}$. Each topic is a probability distribution over m words denoted by $T_i = \{w^i_1, \ldots, w^i_m\}$, where w^i_j is a probability value of word j assigned to topic i. Each post is represented by $P_i = \{T^i_1, \ldots, T^i_k\}$ where T^i_j is a probability value of topic j and post i.

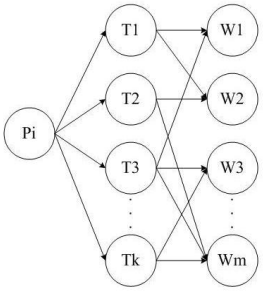


Figure 3. LDA Model

In this paper, we prepare three following feature sets:

(1) Bag of Words (BOW)
(2) feature transformation
(3) feature expansion

In the *Bag of Words (BOW)* approach, we extract terms from Twitter data set by stemming and removing stopwords. Each term is represented with its frequency.

In the *feature transformation* approach, we apply the LDA algorithm to extract words $(w_1,..,w_m)$ from posts of followers and cluster them into the predefined number of topics *T*. We transform term features into topic features using topic probability scores of the posts of each follower and build classification model by those topic features. The efficiency of *topic model* is higher than *Bag of Words (BOW)* depending on the setting of appropriate number of topics.

In the *feature expansion* approach, we extract words $(w_1,..,w_m)$ from the data set and transform them into topic features $(t_1,..,t_k)$ using the LDA algorithm. After that, we transformed the frequency of terms derived from data set and compute the average scores from frequency of each term of the posts and append those scores to topic feature set.

IV. EXPERIMENTS AND RESULTS

We performed experiments using a collection of Twitter posts from three business types: airline, food, and computer & technology. We selected ten companies for each business domain based on the Twitter business directory website "twibs[2]". We collected the follower list of each company using Twitter API and collected posts from follower's blogs using java application. After that, we converted all words to lowercase and removed all punctuation marks, numerical symbols, href tag. The screen name of the receiver is written with "@"symbol and receiver's screen name, i.e., "@Google@". The post which has screen name of the receiver is called "Direct Post".

We categorized posts into four types as follows.

Type A: "Direct Post-Business Domain" The post has "screen name of the receiver" and is the selected companies in our research, e.g., "@JetBlue@ I want to go to Hawaii."

[2]**twibs**, *http://www.twibs.com*

Type B: "Direct Post-Not Business Domain" The post has "screen name of the receiver" but is not the selected companies in our research, e.g., "@Toyota@ Camry is a smart car."

Type C: "Not Direct Post-Relevant Business Domain" The post does not have "screen name of the receiver" but has relevant terms of business domain, e.g., "I want to travel around the world."

Type D: "Not Direct Post-Not Relevant Business Domain" The post does not have "screen name of the receiver" and relevant term of business domain, e.g., "Vios is a smart car too."

The experiments are set up into two following steps:

(1) normalized data set

(2) classification and feature processing

In *normalized data set* experiment, we prepared the normalized term list from "Direct Post" of each business domain. We extracted terms from direct posts and prepare normalized term list to transform short terms to complete terms such as *"flgh"* to *"flight"* and *"budgt"* to *"budget"*. The list of normalized terms in this research contains approximately over 100 terms form each business types.

We used the multi-class data set consisting of "airline" "food" and "computer" class. Each class consists of the post type A and C from 1,000 followers. We applied Support Vector Machine (SVM) as the classification algorithm.

Table 4. The accuracy (%) of normalized and un-normalized data method with multi-class data set

Data Set	Approach	
	Un-Normalized	Normalized
Multi-Class	72.0	76.6

From the result of Table 4, the normalized method performs better with higher accuracy than the un-normalized method up to 4.6%.

In *classification and feature processing* experiment, we prepared a multi-class data set to classify using *Bag of Words (BOW)* and *feature processing*. Using the SVM algorithm to evaluate accuracy of classification. The multi-class data set consists of "airline" "food" and "computer" classes. It consists of the normalized post of type A and C from 1,000 followers of each business domain.

We applied the topic model to construct the classification model as explained in Section III. Once the *term features* were obtained, we applied the LDA algorithm to build a topic model by using the linguistic analysis tool called ***LingPipe***[3]. The LingPipe's LDA Model is estimated by using the Gibbs sampling to select the topics, which represents the posts. The number of topics in this research is set to two different values, 50 and 100.

[3] **Lingpipe**, *http://alias-i.com/lingpipe/*

Next, we applied term features from the Bag of Words (BOW) to append the topic model to improve classification accuracy using the feature expansion approach.

The experimental results of *topic model (LDA) and topic model (LDA) + feature expansion* of multi-class data set are presented in Table 5.

Table 5. The accuracy of *feature transformation and feature* expansion using multi-class data set

Number of Topics	Accuracy (%)	
	Feature transformation	Feature expansion
50	94.67	**95.70**
100	93.90	95.00

The result of Table 5 shows that the feature expansion technique could improve performance classification better than the feature transformation technique. The appropriate number of topics for both approaches is 50.

Next, we used two-class data set of each business type with the proposed techniques. The two-class data set consists of individual class of each business type and "other" class. Each class consists of the post type A and C from 1,000 followers. We applied the Support Vector Machine (SVM) as the classification algorithm. The results are presented in Table 6, 7 and 8.

Table 6. The accuracy of feature transformation and feature expansion using two-class data set in airline business

Number of Topics	Accuracy (%)	
	Feature transformation	Feature expansion
50	97.9	**98.0**
100	97.4	97.8

Table 7. The accuracy of feature transformation and feature expansion using two-class data set in food business

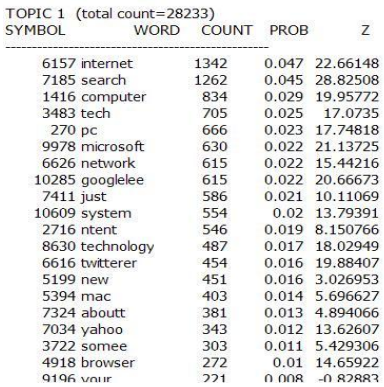
Number of Topics	Accuracy (%)	
	Feature transformation	Feature expansion
50	98.1	98.3
100	98.6	**99.0**

Table 8. The accuracy of feature transformation and feature expansion using two-class data set in computer & technology business

Number of Topics	Accuracy (%)	
	Feature transformation	Feature expansion
50	95.9	**96.1**
100	95.4	95.9

From the results in Table 5 through 8, the optimum number of topics is equal to 50, except for the domain of food business. We used the topic's probability scores of each follower to build the classification model.

The sample topics derived from LDA algorithm are as follows.



TOPIC 1 (total count=28233)

SYMBOL	WORD	COUNT	PROB	Z
6157	internet	1342	0.047	22.66148
7185	search	1262	0.045	28.82508
1416	computer	834	0.029	19.95772
3483	tech	705	0.025	17.0735
270	pc	666	0.023	17.74818
9978	microsoft	630	0.022	21.13725
6626	network	615	0.022	15.44216
10285	googlelee	615	0.022	20.66673
7411	just	586	0.021	10.11069
10609	system	554	0.02	13.79391
2716	ntent	546	0.019	8.150766
8630	technology	487	0.017	18.02949
6616	twitterer	454	0.016	19.88407
5199	new	451	0.016	3.026953
5394	mac	403	0.014	5.696627
7324	aboutt	381	0.013	4.894066
7034	yahoo	343	0.012	13.62607
3722	somee	303	0.011	5.429306
4918	browser	272	0.01	14.65922
9196	your	221	0.008	-0.82883

Figure 4. Word's probability score in each topic from LDA Algorithm.

The examples of word's probability values of each topic from the LDA algorithm from each business type are shown in Table 9 through 11.

Table 9. Example of word's probability values of a topic related to airline business

Business : airline			
Topic#5		**Topic#31**	
Word	**Prob.**	**Word**	**Prob.**
bag	0.051	travel	0.056
upgrade	0.026	pilot	0.055
airline	0.025	bag	0.055
flight	0.025	flight	0.054
…	…	…	…

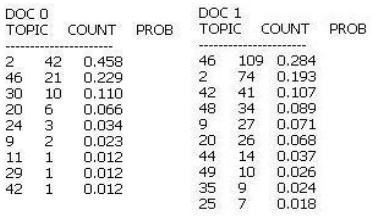
Table 10. Example of word's probability values of a topic related to food business

Business : food			
Topic#26		**Topic#37**	
Word	**Prob.**	**Word**	**Prob.**
cream	0.043	food	0.043
cake	0.040	eat	0.040
brownie	0.002	lunch	0.038
caramel	0.002	sweet	0.032
…	…	…	…

Table 11. Example of word's probability values of a topic related to computer & technology business

Business : com & technology			
Topic#0		**Topic#10**	
Word	**Prob.**	**Word**	**Prob.**
internet	0.047	computer	0.063
search	0.045	laptop	0.033
computer	0.029	hp	0.033
technology	0.017	dell	0.028
…	…	…	…

In Table 9, 10 and 11, each word may belong to many topics with the same or different probability values. We applied the topic probability values of each follower, e.g. as shown in Figure 5 to build the classification model.



DOC 0

TOPIC	COUNT	PROB
2	42	0.458
46	21	0.229
30	10	0.110
20	6	0.066
24	3	0.034
9	2	0.023
11	1	0.012
29	1	0.012
42	1	0.012

DOC 1

TOPIC	COUNT	PROB
46	109	0.284
2	74	0.193
42	41	0.107
48	34	0.089
9	27	0.071
20	26	0.068
44	14	0.037
49	10	0.026
35	9	0.024
25	7	0.018

Figure 5. Topic's probability score of the posts of each follower from LDA Algorithm.

In Figure 5, "DOC 0" refers to the topic representation of posts from a follower. For example, the posts of this follower has a topic probability value of Topic#2 equal to 0.458 and of Topic#46 equal to 0.229.

Next, we applied term features from the *Bag of Words* (BOW) to append into the *topic model* to improve classification accuracy using the *feature expansion*. We selected term features with frequency above 20 and transformed all term into extra topics. We then computed the average of each term (extra topic), which is in the posts of each follower to append the topic model features set. This method refers to *"feature expansion"*.

We compared three different feature sets: *Bag of Words (BOW)*, *feature transformation*, and *feature expansion.* We apply SVM algorithm by using **Weka** [4] to build the classification model. The experimental results are summarized in Table 12.

Table 12. Comparison of the accuracy among three different feature sets using multi-class data set.

Feature Sets	**Accuracy (%)**
Bag of Words (BOW)	76.60
Feature transformation (50 Topics)	94.67
Feature expansion (50 Topics)	95.70

Table 12 shows the accuracy comparisons of the three approaches using the SVM algorithm on the multi-class data set. The accuracy of the *feature expansion* approach is higher than the *feature transformation* approach up to 1.03% and higher than the *Bag of Words (BOW)* method up to 19.1%.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed and compared several approaches for improving the performance of classification models using the Twitter posts. We focus on two different approaches including text normalization and feature expansion technique. We applied the term normalization to improve the quality of data. For multi-classification models, the normalization process yielded the improved accuracy up to 4.6%. For the feature processing, we applied three approaches, i.e., *Bag of Words (BOW)*, *feature transformation* and *feature*

[4] **Weka**, http://www.cs.waikato.ac.nz/ml/

expansion to build multi-classification models. From the experimental results, the performance of *feature expansion* approach yielded the accuracy higher than the *BOW* and *feature transformation* approach up to 18.07% and 19.1%, respectively.

For two classification models, we applied two approaches, i.e., *feature transformation* and *feature expansion* to build two classification models. From the experimental results, the performance of feature expansion approach using *two class* data set of airline, food and computer & technology business yielded the accuracy higher than feature expansion approach using *multi-class* data set up to 2.3%, 3.3% and 0.4%.

REFERENCES

[1] S. Banerjee, "Improving text classification accuracy using topic modeling over an additional corpus," Proceedings of the 31 st annual international ACM SIGIR conference on research and development in information retrieval, 2008, pp. 867-868.

[2] J. Basilico, and T. Hofmann, "Unify Collaborative and Content-Based Filtering". ACM International Conference Proceeding Series. Vol.69. Proceedings of the twenty-first international conference on Machine learning, 2004.

[3] M. Chau, and H. Chen, "A machine learning approach to web page filtering using content and structure analysis," ScienceDirect, Decision Support System, Volume 44, Issue 2 (January 2008), pp. 482-494

[4] N. Churcharoenkrung, Y. S. Kim, and B. H. Kang, "Dynamic Web Content Filtering based on User's Knowledge," Proceedings of the International Conference on Information Technology: Coding and Computing, 2005.

[5] J. M. G. Hidalgo, G. C. Bringas, and E. P. Sanz, "Content Based SMS Spam Filtering," Proceedings of the 2006 ACM symposium on Document engineering, Amsterdam, The Netherlands, 2006, pp. 107-114.

[6] G. Ifrim, M. Theobald, and G. Weikum, "Learning Word-to-Concept Mappings for Automatic Text Classification," International Conference on Machine Learning, Bonn, Germany, 2005.

[7] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Micro-blogging as Online Word of Mouth Branding," Proceedings of the 27th international conference extended abstracts on Human factors in computing systems, 2009, pp. 3859-3864

[8] A. Java, X.Song, T. Finin, and B. Tseng, "Why we Twitter: Understanding microblogging usage and communities," Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, 2007.

[9] E. Linstead, P. Rigor, S. Bajracharya, C. Lopes, and P. Baldi, "Mining Concepts from Code with Probabilistic Topic Models," ASE'07 November 5-9, Atlanta, Georgia, USA, 2007.

[10] V. Maidel, P.Shoval, B. Shapira, and M. Taieb-Maimon, "Evalution of an Ontology-Content Based Filtering Method for a Personalized Newspaper," Proceedings of the ACM conference on Recommender Systems, 2008, pp. 91-98.

[11] G. Pasi, G. Bordogna, and R. Villa, "A multi-criteria content-based filtering system," Proceedings on the 30th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 2007, pp. 775-776.

[12] X. H. Phan, L. M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," Proceeding of the 17th international conference on World Wide Web, 2008, pp. 91-100.

[13] V. Ha-Thuc, and P. Srinivasan, "Topic Models and a Revisit of Text-related Applications," PIKM'08, October 30, 2008, Napa Valley, California, USA, 2008.

[14] M. N. Uddin, J. Shrestha, and G. S. Jo, "Enhanced Content-based Filtering using Diverse Collaborative Prediction for Movie Recommendation," First Asian Conference on Intelligent Information and Database Systems. ACIIDS 2009. pp. 132 -137.

AUTHORS PROFILE

Chanattha Thongsuk: Chanattha Thonsguk was born in 1979 and is studying in a Ph.D. program of Faculty of Information Technology, King Mongkut's University of Technology North Bangkok, Thailand. She received her Bachelor degree from Siam University and Master degree from Mahidol University. She is interested in many research topics including text mining, social network and recommender system.

Choochart Haruechaiyasak: Choochart Haruechaiyasak received a B.S. from the University of Rochester, an M.S. from the University of Southern California and a Ph.D. degree in Computer Engineering from the University of Miami. His current research interests are search technology, data/text/web mining, information filtering and recommender system. Currently, he is a senior researcher in the Intelligent Information Infrastructure Section under the Human Language Technology Laboratory (HLT) at the National Electronics and Computer Technology Center (NECTEC), Thailand.

Somkid Saelee: S. Somkid is a teacher at Department of Computer Education, King Mongkut's University of Technology North Bangkok, Thailand, He is M.CS in Information Technology from Kaetsart University (2000) and Ph.D. in Computer Education from King Mongkut's University of Technology North Bangkok (2008).