# Knowledge Discovery from Twitter

Zhijin Li (zli12@illinois.edu)

*Master Candidate*

*Department of Computer Science*

*University of Illinois at Urbana-Champaign*

## Abstract

Microblogging services like Twitter [1] are more and more popular, forming as a part of social media, social network and communication tool. Currently the number of microblogging entries in Twitter, known as tweets, is quite big and still increasing every day. Information management and organization in microblog are becoming not only a problem, but also an interesting research topic. The huge amount of text data produced by Twitter becomes a very desirable dataset for knowledge mining and discovery. Besides utilizing the text of tweets data, the users in Twitter are connected by "following" relationship (i.e. feeding someone's tweets by "following" him), we can then build a text-associated information network for better modeling and interesting patterns discovery over text data. This paper is basically going to achieve three research tasks over Twitter data: tweets filtering based on a user's interests, community discovery in a large group of people, and tweets classification. Generally, identifying interests helps filtering undesirable information on incoming tweets, community discovery helps find subgroups of certain interests and suggest users who have similar interests to follow, and tweets classification helps user choose his favorite categories of tweets to read. Experiments are designed and performed respectively. The experimental results show the effectiveness of proposed statistical framework and algorithms for these tasks.

## 1 Introduction

### 1.1 Introduction to Microblog

Microblog is a kind of online communication tool by which users update what they are currently thinking and doing, what their opinion about a specific object or phenomenon. A variegated usage of microblog has emerged after the first microblogging service provider, Twitter.com, launched, including daily chatter, conversation, information sharing, news commentary, and political uses.

Microblogging is increasingly popular on the web. It allows users to broadcast brief text updates to the public or to a limited group of contacts. Because of the popularity shown by Twitter, many others companies then follow and provide microblogging service. As an alternative form, microblogging capabilities have been implemented by many social networking websites as status updating functionality, e.g. on Facebook [2] and Myspace [3].

The microblog entry has many characteristics different from traditional text document unit. For example, at Twitter, the length of a microblog post, known as a tweet, cannot exceed 140 characters. Though the messages are very short, it is proven that they have extremely strong ability to effectively express ideas and share information as a communication tool on the web.

## 1.2 Twitter as a Mining Source

The Twitter data are growing extremely fast. According to an unofficial statistics, there are about 155 million tweets posted to the network each day. This is more than 3 times from the company's report of 50 million tweets per day, just one year ago. Besides the huge amount of text data, people using Twitter are interconnected by a special kind of relationship: following. Following a person means you are interested in his or her tweets so you would feed those tweets.

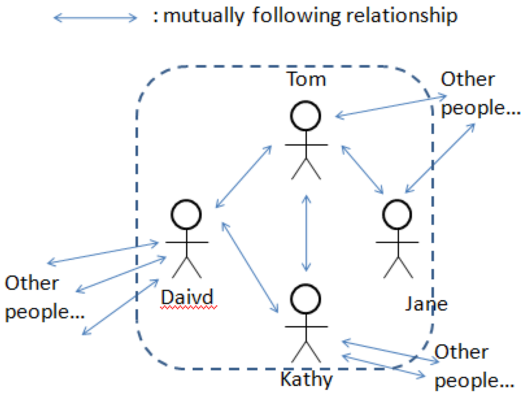

*Figure 1: "following" relationships at Twitter, which can be used to build information network*

Figure 1 gives a illustration of mutual "following" relationships at Twitter. Under these following relationships, Twitter users and their associated tweets are forming an increasing information network. Each user's tweets is a node and each following relationship is a an edge in the information network, respectively. For example, an bidirectional arrow between Tom and David indicates an edge between two nodes made by their tweets. Therefore, Twitter becomes a quite unique and promising dataset to discover interesting patterns. Among various text mining goals, information management and topics finding in microblog are particularly interesting research tasks, as they are closely related to the special features of tweets.

Many specialties of Twitter data, such as the short length of tweets, time consideration, hash tag usage and interconnection among users bring many challenges as well as prior knowledge to pattern discovery. Additionally, as Twitter is providing a API to get its tweets for free, it is naturally used as the desirable data source for experiments in this paper.

## 1.3 Research Tasks

### *1.3.1 Tweets Filtering based on User's Interests*

As a definition given by Wikipedia [4], an information filtering system is a system that removes redundant or unwanted information from an information stream using automated or semi-automated methods prior to presentation to a human user. This definition can be naturally applied to information filtering in Twitter: before a tweets stream come to you, ideally there is a system to filter out the uninteresting tweets while retaining interesting ones.

It is a very common scene that when you are following a bunch of people who post tweets actively, a number of tweets posted by them everyday will overwhelm your eyes. Obviously you are not interested in all of them and just want to pick some tweets to read, and in the real case you are likely only interested in a small portion of the flooding tweets.

This paper proposes a threshold based method to filter uninteresting tweets. The interestness of a tweet is represented by a numeric value. If this value is greater than a threshold this tweets would be retained and present to user, otherwise it would be filtered out before forwarding to user. The interestingness score is computed via marginal probability of the tweet given a model that represents user's interests. This model would be estimated by Latent Dirichlet Allocation (LDA) [5, 14, 16], which is a probabilistic topic model that represents a collection of documents as a set of topic distribution for each document. The interests of people are the topics in LDA, and the interests of a user is represented by a interest topic distribution.

### *1.3.2 Community Discovery*

In social networks, like Facebook and Twitter, the task of community discovery means given a number of people, finding subgroups of people who are "similar" to each other within their group. The similarity measure for this task is the interests similarity among people. A particular person may or may not be involved in one or multiple subgroups at the same time, which is depending on the actual application needs.

The functionality of providing similar users recommendation to people is almost like a indispensable part in the whole application, as it benefits in, but not limited to these aspects: first, it helps users find other people they might be interested in following. This also helps the application provider get more users connected together and make the social network growing. Second, it provides a possible way to do accurate online advertising, based on the common interests of subgroups.

The basic research problem here is to find communities (subgroups) of users with similar interests. Besides exploiting text information provided by tweets, there are connections, the "following" relationships, among people. If a person is following another person, it is very likely that they share some interests, this is because the "following" relationships are most likely to be formed by people who share similar interests. As a result, the connection information can be exploited in addition to the text information to do better topic modeling. The idea was inspired by Yizhou's paper introducing iTopicModel [6], which utilizes both the link and text information among documents in the information network to improve traditional topic modeling. Using "following" relationships among Twitter, we can accordingly build such a document network. After the better topic model is trained, by comparing topic distributions of users, we can find subgroups that share interests and thus provide recommendations with groups.

The subgroup is defined as a set of nodes in the original experimenting information network where the number of nodes is greater than a given threshold and the similarity between any pair of nodes is greater than another given threshold. At last, the distance between two nodes, which is the interest similarity of two users in our case, is defined as cosine similarity of the respective two topic distributions.

### *1.3.3 Tweets Classification*

The goal of tweets classification is to divide different tweets into a few categries, depending on the content of each tweet. The purpose of it is that it is often the case that Twitter users just would like to select a few tweets in a few specific categories to read at a certain time.

For example, some professional people may prefer to read some technical tweets while working, watch recent news as relaxation, see some conversation tweets when idle, and care about what their interested people are doing. By providing several categories with related tweets in each of them, the tweets feeding by a Twitter user can be displayed in several columns, and thus the categories allow the user easily pick out some columns to read in different situation.

A simple but very effective algorithm K-Means [9, 10, 17] is applied and modified to achieve this task. In order to classify a number of tweets into a few common reading categories like technical discussion, news, events, sports, daily conversation, which are generally preferred by people's reading behavior, instead of randomly selecting $M$ documents as initial centers, which is what the original K-Means algorithm does, a few topic-specific centers are manually made as supervision. The modified algorithm then classifies tweets and re-computes centers in an iterative way as usual. With the help of the supervision, the classification result can be more desirable and satisfy user's changing information need.

## 2 Modeling

*2.1 Tweets Filtering based on User's Interests*

This section is to give a formal definition and probabilistic model for filtering uninteresting tweets depending on a certain user's interests. As mentioned before, to determine a given tweet is interesting not, or say whether it needs to be retained or filtered out, a interests topic model, constructed from LDA model proposed by David Blei [5], is applied to estimate the model that represents a user's interests, and then use this model to compute marginal probability of the tweet as interestness score.

As a generative model, LDA uses a multinomial distribution over topics controlled by a Dirichlet prior for a document. The probability density function of a Dirichlet distribution over a multinomial distribution p is given by,

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_K^{\alpha_K - 1}$$

Here, $\alpha$ is a collection-level parameter, a length-$K$ vector as the hyperparameter for Dirichlet distribution, where $K$ is the number of topics in the collection. $\Gamma$ is Beta function []. $\theta$ is a document-level parameter, a length-$K$ vector indicating the topic distribution of a document. In our case, the collection of documents is a collection of tweets acquired from people the user is following at Twitter.

The generative process for a tweet $t$ in the model can be described as follows:

(1) Choose the tweet length $N \sim Poisson(\zeta)$

(2) Choose topics distribution $\theta \sim Dir(\alpha)$ for $t$

(3) For each of the $N$ words $w$ in the tweet,

(a) Choose a topic $z \sim Multinomial(\theta)$

(b) Generate a word $w$ under $p(w \mid z, \beta)$, a multinomial probability conditioned on the topic

Under the LDA model, the marginal probability of a newly incoming tweet $t$ is given by,

$$p(t \mid \alpha, \beta) = \int_{\alpha} p(\theta \mid \alpha) \prod_{j=1}^{N} p(w_j \mid t)$$

$$= \int_{\alpha} p(\theta \mid \alpha) (\prod_{j=1}^{N} \sum_{k=1}^{K} p(z = k \mid \theta) p(w_j \mid z = k, \beta)) d\theta$$

The logarithm of this probability is defined as the score of interestness used to compare with a threshold $\lambda$ as filtering condition, i.e.

$$score(t, \alpha, \beta) = \log p(t \mid \alpha, \beta)$$

If $score(t, \alpha, \beta) > \lambda$, this tweet $t$ is defined as "interesting" to the user so as to be retained. Otherwise, it is "uninteresting" to the user so as to be filtered out before presenting to the user.

Therefore, the whole process of tweets filtering for a certain user is as follows:

(1) Collect a number of tweets from the people who are being followed by the user

(2) Treat these tweets as documents to estimate a LDA model as the user's interests model

(3) Given a stream of newly incoming tweets, for each tweet compute the marginal probability under the estimated interests model, and then compare it with the filtering threshold $\lambda$, to decide wether it should be filtered out or not

(4) Present the tweets retained in previous step, which are considered as interesting ones that satisfy long-time information need to the user

*2.2 Community Discovery*

In order to find communities among a large group of people, as mentioned in previous section, one way is to find communities (subgroups share interests) base on people's similar interests. The interests of a person can be modeled as a topic distribution over all topics in the collection, which is comprised of tweets from all the people under experiment. Using topic modeling technique, the interest topic distributions can be estimated for every people. Additionally and importantly, the links among people, which are the "following" relationships, are taken into consideration.

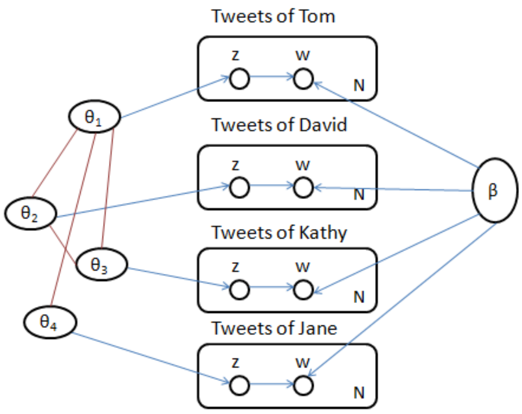

*Figure 2: information network involved topic model that explores both link and text*

Figure 2 is the new graphic model representation for Twitter information network, followed by iTopicModel [6]. Each $\theta_i$ is a topics distribution of the tweets of a user. Each rectangle is the generation process of the respective tweets. The model is different from traditional topic model in the left part in the figure: no dependencies among those topic distributions in traditional topic modeling. The dependencies are directly related and modeled by the mutually "following" relationships in the Twitter information network. The belief is that at Twitter mutually followed people likely share similar interests.

For our task, we can group 1000 tweets of a Twitter user as a document, and if two users are following each other at Twitter, there would be a link between their topic distributions and their distributions will eventually affect each other. The links will reinforce its neighbours by its knowledge, and thus propagated information may generate better models.

These links in the graphic model are modeled by Markov Random Fields [11], which gives structural probability, i.e. $p(\theta \mid G)$ . Specifically,

$$p(\theta \mid G)=\frac{1}{Z}\exp\{-\sum_{c\in C}V_c(\theta)\}$$

Where $C$ is a set of cliques in the graph $G$ , $Z=\sum_{\theta}\exp\{-\sum_{c\in C}V_c(\theta)\}$ is a partition function, and $V_i(\theta)$ is a potential function defined as,

$$V_i(\theta_i)=-(\alpha_i^0-\vec{1})^T\log(\theta_i)$$

$$V_{i\to j}(\theta_i,\theta_j)=-(w_{ij}\theta_j)^T\log(\theta_i)\text{ , if }<t_i,t_j>\in E\text{ ;}$$

otherwise it equals to 0. If $i$ -th user is following $j$ -th user, there is an edge in the graph and $w_{ij}=1$ indicates the weight of the edge.

Then the structural probability is derived by,

$$p(\theta \mid G)=\frac{1}{Z}\exp\{\sum_{i}[(\alpha_i^0+\sum_{j\in N(i)}w_{ij}\theta_j-\vec{1})^T\log(\theta_i)]\}$$

At the same time, the traditional topic model gives the probability of tweets, i.e. $p(t_i \mid \theta,\beta)$ for each tweet. Specifically, each grouped tweets of a user, denoted as $t_i$ is deemed as a $N$ -word document in topic modeling and modeled by a mixture model over $K$ topics, where each document is assumed independent to others. Therefore, the probability of the whole collection $T$ for $M$ -user's tweets is given by,

$$\begin{aligned}p(T\mid\theta,\beta)&=\prod_{i=1}^{M}p(t_i\mid\theta,\beta)=\prod_{i=1}^{M}\prod_{j=1}^{N}p(w_j\mid t_i,\theta,\beta)\\&=\prod_{i=1}^{M}\prod_{j=1}^{N}\sum_{k=1}^{K}p(z=k\mid x_i)p(w_j\mid z=k)\\&=\prod_{i=1}^{M}\prod_{j=1}^{N}\sum_{k=1}^{K}\theta_{ik}\beta_{kj}\end{aligned}$$

Given the structural and text probabilities we can take the joint probability as objective function, then use EM algorithm to estimate model parameters, $\theta$ , which is a set of topics distributions for each grouped tweets for each user, and β, which records the probability of each word under each topic.

$$p(T,\theta\mid G,\beta)=p(\theta\mid G)p(T\mid\theta,\beta)=p(\theta\mid G)\prod_{i=1}^{M}p(t_i\mid\theta_i,\beta)$$

Once $\theta$ is estimated, a set of communities can be collected, given the definition of community below,

$$Communities = \{c \mid \forall \theta_1, \theta_2 \in c, c \subset \theta, sim(\theta_1, \theta_2) > \phi, |c| > \mu\}$$

where $\mu$ is the minimum number of users in a community and $\phi$ is the minimal similarity between any pair of elements within this community. The cosine similarity [12, 13] is selected as the similarity measure,

$$sim(\theta_1, \theta_2) = \frac{\theta_1 \bullet \theta_2}{\|\theta_1\| \|\theta_2\|} = \frac{\sum_{i=1}^{K} \theta_{1j} \times \theta_{2j}}{\sqrt{\sum_{j=1}^{K} \theta_{1j}^2} \times \sqrt{\sum_{j=1}^{K} \theta_{2j}^2}}$$

where $\theta_{ij}$ is the probability of $j$ -th topic in $i$ -th user's interest topic distribution and $K$ is the total number of topics, also the length of each $\theta_i$ .

*2.3 Tweets Classification*

The tweets classification task is achieved using a simple but a little modified K-Means algorithm: a small number of dummy tweets are manually built used as initial classification centers rather than randomly selection in the original algorithm. The dummy tweets are made by a few commonly interested reading category, like professional discussion, news, events, sports, and daily conversation. Of course these categories can be changed and predetermined by a certain user's preference.

The tweets would be classified into its currently nearest center in a iterative way. The distance measure between two tweets $p$ and $q$ used is Euclidean distance, given by,

$$\begin{aligned} d(p,q) &= \sqrt{(c_{11} - c_{21})^2 + (c_{12} - c_{22})^2 + \ldots + (c_{1n} - c_{2n})^2} \\ &= \sqrt{\sum_{i=1}^{|V|} (c_{1i} - c_{2i})^2} \end{aligned}$$

Here, each tweet is represented by a word count vector over a underlying vocabulary $V$ that records every word. $c_{1i}$ and $c_{2i}$ are the counts of the word in $p$ and $q$ respectively, where the word is in $i$ -th position of the vocabulary.

The algorithm performs in an iterative way: in each iteration tweets are classified into their nearest centers (also tweets), and then new centers would be re-computed by tweets in each category. The whole process is outlined below:

(1) Transform every tweets to be classified into word count vectors over the vocabulary

(2) Set initial centers using manually built dummy tweets

(3) Iterate over every tweet, find its closest center by comparing distance between it and each current center, and then classify it into that category respectively

(4) Re-compute centers by averaging all the tweets in each category

(5) Repeat (3) and (4) until there is minor change in classification or a given maximum iteration number is reached

After step (5) is complete, we can show the classification result to the user by different columns of tweets representing different reading categories, possibly with a few highly-ranked words in each class as category indicators.

## 3 Experiments

### 3.1 Data Collection

Twitter is providing a set of application programming interfaces (API) that allows anyone to retrieve tweets, profiles, activities or events of people, as well as the connections among people (e.g. the "following" relationship). Twitter is currently providing a streaming API and two discrete REST APIs. The API is HTTP based, and GET, POST requests can access the data. The API uses basic HTTP authentication and requires a valid Twitter account. Data can be retrieved as XML or succinct JSON format.

Three APIs are used in the experiments, one that retrieves IDs of people followed by a given user, one that retrieves recent 200 tweets of a given user, and one that judge whether a "following" relationship exists between two users.

Like many experiments on text data, in order to generate more reasonable result, all the experimental data retrieved from Twitter APIs are preprocessed. The preprocessing includes two parts: first, stop words removal, which removes some extremely commonly used words (e.g. "a", "the", "that") and prevents those words from dominating among topics, and second, words stemming, which normalize a word in different forms into a uniform form.

## 3.2 Experiments Design and Results

### *3.2.1 Tweets Filtering based on User's Interests*

The person selected for this experiment is Hilary Mason (http://twitter.com/hmason), who is the chief data scientist at bit.ly [7], interested in machine learning and data science. Using Twitter's API, 200 tweets from each of 150 randomly selected people she is following are retrieved. A LDA model representing interests is trained based on these tweets. 30 newly incoming tweets are pushed to the filter, with filtering threshold $\lambda = -100$. This threshold can change according to different degrees of information filtering need. More interestingly, priors to some particular topics that the user would like to retain can be added for high-quality filtering.

Limited to space, table 1 shows only 10 out of the 30 newly incoming tweets, along with the interestness scores, which is the logarithm of marginal probability of each tweet given Hilary Mason's interests model. Tweets whose interestness score are above -100 are retained to show, whereas the others are considered as "uninteresting" to the Hilary Mason and thus filtered out before presenting to the her.

When looking into table 1, not surprisingly, the topics of the 4 retained tweets are prone to talk about data science and web stuff, which are the topics she concerns. The 6 filtered tweets seem talk about little about data science and machine learning, which are Hilary Mason's interests .

| **Newly Incoming Tweets** | $\log(t \mid \alpha, \beta)$ | **Status** |
|---|---|---|
| New blog: What should I cut from Team Time Management?: I am rewriting my class Advanced Time Management. | -138.08359 | filtered |
| RT @Algebra: The condition number of a matrix A relative to the euclidean norm is the ratio of its smallest igenvalues. | -107.43564 | filtered |
| @gruber how exactly are Mobile Safari exploits worse with Nitro? Remote address book hijack possible pre-Nitro. | -99.891827 | retained |
| Pretty interesting RT @OpenHQR: San Francisco Rainwater: Radiation 181 Times Above US Drinking Water Standard. | -146.24928 | filtered |
| Harley to slow for city traffic. This is a modified R1200C, also a massive machine but more torque for city driving. | -100.24776 | filtered |
| reading #OSCON Data proposals. have coined a new acronym. YAWNS - Yet Another Wanking NoSQL Solution. | -90.669528 | retained |
| Your infographic has one design flaw - I impulsively want to hover over the points in the scatterplot and see the couples. | -83.292602 | retained |
| You are working on exciting stuff that will revolutionize fashion Fashionistas and entrepreneurs stay tuned | -109.35694 | filtered |
| Having a lot of fun with the Beads Processing library. Easy, full-features sound synthesis, analysis, and playback. | -111.00092 | filtered |
| #followfriday @aghose NYU Stern professor, new Tweep, and one of the winners of the WWW2011 best paper award. | -94.841583 | retained |

*Table 1: A stream of 10 newly incoming tweets, their marginal probabilities and the filtering result*

### *3.2.2 Community Discovery*

To find communities in a information network, the first step is to build a Twitter information network with text as its major attribute for nodes. Each node in the information network is created by grouping 1000 tweets of a user (some of them have posted less than 1000 tweets till now), and an edge in between two nodes is formed if the two users associated with the two nodes are following each other at Twitter.

In order to provide a way to evaluate the final result, 30 Twitter users who have relatively "clear interests" are selected as seed nodes for breadth first searching for incorporating more users as nodes in the information network. The "clear interests" is determined by manual looking at his or her feeding tweets, seeing if those tweets are great indicators for some certain interests. In the experiment, the "clear interests" selected include "cooking", "Internet and web", "multimedia", "social networking", "current affairs", "digital library", "travel", "game industry" and so on. As an evaluation, we would like to compare the interests of communities found by the information network integrated topic model with the previously selected ones (e.g. to see if there is a community where people within it share common interest in "Internet").

After collecting 2313276 tweets in total from 400 Twitter users as nodes, edges between two nodes are then added if the users are mutually following at Twitter, which can be tested a Twitter API. An iTopicModel runs over the built information network, with initial setting number of topics $K = 50$, Dirichlet prior $\alpha_i = 50$ and word distributions under each topic $\beta_{ij} = 0.01$. After topic distribution $\theta_i$ for each user (represented by 1000 grouped tweets) is estimated via EM algorithm, $\mu = 30$ is set as the minimum number of users and $\phi = 0.65$ is set as minimum similarity within a community. The table 2 below shows the communities found. Because it seems meaningless to show the users' names in each community, instead, some representative topic words, number of users, and minimum pairwise similarity in 4 largest communities are given. The communities result resembles the "clear interests" provided by seed nodes: the major topics for community 1-4 is like "game and entertainment", "travel and events", "Internet and web" and "research and study", accordingly.

| | **Community 1** | **Community 2** | **Community 3** | **Community 4** |
|---|---|---|---|---|
| **Number of Users** | 48 | 67 | 91 | 42 |
| **Min Pairwise Sim** | 0.76 | 0.82 | 0.90 | 0.71 |
| **Community Label** | Game, Media, Entertainment | Travel, Events, World | Internet, Web, Online | Research, Study |
| **Top-ranked Words** | game, games, startup, microtask, samplereality, gameloft, cpuo, ps3, love, crowd, play, flash, ea, starcraft2, ibogost, xbox, mobile, amazing, tv, free, experience, video, zynga | tonight, pm, airport, waiting, honolulu, hotel, international, boston, checked, car, center, blog, hawaii, world, interesting, article, photos, event, libya, story, egypt, piece, east, missing, chinese | http, mobile, live, space, tech, elearning, action, marketing, twitter, facebook, tweet, fb, follow, page, users, link, google, book, free, ipad, email, read, search, phone, cool, books, code, site, apps | digital, library, university, research, culture, job, pdf, humanities, projects, conference, public, tech, studies, harvard, talk, year, congrats, listening, talking, times, paper, dr, interesting |

*Table 2: found communities and their related information*

### *3.2.3 Tweets Classification*

The tweets used in this classification experiment are from people followed by Gemma Petrie (http://twitter.com/GemmaPetrie), who is generally interested in information and media, social events and various food. Therefore she is following many people who usually keep broadcasting these topics. However, obviously people followed by Gemma may post many tweets that may not be of her interests. In order to make the result more accurate, 7 dummy tweets containing some representative words in these topics (e.g. social, media, event, activity, food, gourmet) are made as the initial centers for the K-Means algorithm.

In total 3400 tweets are collected from her following people. From the classification result, about 2000 tweets out of them fall into the 3 below categories, which interest her most. For space limitation, the table 3 below shows only the top-3 interesting tweet categories of her and 7 tweets within each categories.

| Tweets Category | Tweets within Category |
|---|---|
| **Social, Media, News**<br>Top ranked words: social, twitter, media, interesting, people, digital, nytimes, post, information, looking, online, public | Back online after a fantastic weekend with @PeregrineKiwi ! Looking fwds to partying again in a month, next time in LA :D<br>@karenwickett ...but I wish I could have more face2face conversations with you.<br>tweetvalue.com calculates your value for twitter, type in username. Mine's worth $45 , much less than the avg $136 for FB<br>If you're at #asist2010 today, check out my colleague Dave' talk on social media emergency #KM during the Haiti earthquake<br>Tried to reduce how many people I follow on Twitter and ended up adding 2 more #informationoverloadfail<br>finished analysis of tweets, status messages, and blogs regarding the type of information provided by user generated content<br>He also said that Zuckerberg rarely posts anything on Facebook |
| **Event, Celebration, Activity, Award**<br>Top ranked words: congrats, watching, photo, show, family, nice, afternoon, birthday, life, fun | RT @nmtechcouncil: Reminder: #OpenCoffee this Thursday AM at the Santa Fe Business Incubator -hope to see you there!<br>Wow! Congratulations @BAVC for receiving the 2010 MacArthur Award for Creative & Effective Institutions<br>thanks for sharing and again congrats! \| @fstutzman dissertation - Networked Information Behavior in Life Transition<br>A really funny Daily Show with Ricky Gervais earlier this week.<br>@janedavis @veruka2 Ha! That sounds like quite an evening.<br>@clhw1 Great weather, lots of family, and a reasonable number of fish.<br>woke up after a great wine night with a friend... how could I say goodbye to all of them and still go to work in the morning?!? |
| **Food, Drink, Meal, Gourmet**<br>Top ranked words: coffee, food, birthday, beverage, delicious, drink, amazing, favourite, home, beer | Intelligentsia Goes Back to Basics for Brewed Coffee - “If you’re an Intelligentsia regular and drink brewed...<br>@midcenturysal I am unsupervised, eating brisket, drinking drinks named for inventors, and watching bats<br>A Hot Dog for Everyone at The Slaw Dogs - Good Food on the Road on KCRW: http://bit.ly/aTQ6cv via @addthis<br>Ending a long day with warm cookies and cold beer. http://tumblr.com/xf6b8cg9m<br>@barbermatt wheeze the juice!<br>Macaroni & Cheese with Blue Cheese, Figs, and Rosemary: Sure to comfort the winter blues.<br>I'm serious about not needing to cook dinner tomorrow - Serious Ragù Bolognese http://t.co/5SbhBmL |

*Table 3: A stream of newly incoming 10 tweets, their marginal probabilities and the filtering result*

As a potential extension, once the classification model is relatively stable after times of iteration using a large set of training tweets, the centers can be set as fixed so that when new tweets come we can quickly find its nearest center and then put it into the respective category. This might be able to allow the system to go online as an application running in a real-time manner.

## 4 Conclusions

This paper proposes three research tasks all about Twitter data analysis: tweets filtering based on user's interests, community discovery and tweets classification. For tweets filtering, a number of tweets from people who are followed by a certain user is collected, used to train a LDA model as his interests model, and then use this model to compute the interestness score for newly incoming tweets to decide filtering. For community discovery, a information network is built on Twitter users and their "following" relationship, used to estimate interest topic distribution of users by iTopicModel and find communities by comparing them. For tweets classification, a modified K-Means algorithm is applied to classify tweets into different common reading categories. All the tasks proposed are aimed to enhance user's reading experience and better satisfy his or her changing information need from Twitter.

## References

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

[2] Facebook, http://www.facebook.com

[3] MySpace, http://www.myspace.com

[4] Information Filtering System, http://en.wikipedia.org/wiki/Information_filtering_system

[5] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

[6] Yizhou Sun, Jiawei Han, Jing Gao, and Yintao Yu, iTopicModel: Information Network-Integrated Topic Modeling, Proc. 2009 Int. Conf. on Data Mining (ICDM'09 ), Miami, FL, Dec. 2009.

[7] Bit.ly, http://www.bit.ly

[8] Beta Function, http://en.wikipedia.org/wiki/Beta_Function

[9] K-Means, http://en.wikipedia.org/wiki/K-means_clustering

[10] K Wagstaff, C Cardie, S Rogers. Constrained k-means clustering with background knowledge. ICML 2001.

[11] R Kindermann, Markov random fields and their applications.

[12] Cosine Similarity, http://en.wikipedia.org/wiki/Cosine_similarity

[13] Michael Steinbach George Karypis Vipin Kumar. A Comparison of Document Clustering Techniques. KDD 2000.

[14] T.Hofmann. Probabilistic latent semantic analysis. In Proceedings of UAI 1999, pp. 289–296, 1999.

[15] Rui Wang, Rongshen Jin. An Empirical Study on the Relationship between the Followers' Number and Influence of Microblogging. 2010 International Conference on E-Business and E-Government

[16] Statistical Language Models for Information Retrieval, ChengXiang Zhai

[17] 2001Paul B., Usama M. Refining initial points for k-means clustering. ICML 1998.