Expert Systems with Applications 41 (2014) 4330–4336

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Identifying interesting Twitter contents using topical analysis

Min-Chul Yang, Hae-Chang Rim *

Department of Computer & Radio Communications Engineering, Korea University, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Twitter
Interesting content
Topic model
LDA
Social media

ABSTRACT

Social media platforms such as Twitter are becoming increasingly mainstream which provides valuable user-generated information by publishing and sharing contents. Identifying interesting and useful contents from large text-streams is a crucial issue in social media because many users struggle with information overload. Retweeting as a forwarding function plays an important role in information propagation where the retweet counts simply reflect a tweet's popularity. However, the main reason for retweets may be limited to personal interests and satisfactions. In this paper, we use a topic identification as a proxy to understand a large number of tweets and to score the interestingness of an individual tweet based on its latent topics. Our assumption is that fascinating topics generate contents that may be of potential interest to a wide audience. We propose a novel topic model called Trend Sensitive-Latent Dirichlet Allocation (TS-LDA) that can efficiently extract latent topics from contents by modeling temporal trends on Twitter over time. The experimental results on real world data from Twitter demonstrate that our proposed method outperforms several other baseline methods.

1. Introduction

With the rise of the Internet, blogs, and mobile devices, social media has also evolved into an information provider by publishing and sharing user-generated contents. By analyzing textual data which represents the thoughts and communication between users, it is possible to understand the public needs and concerns about what constitutes valuable information from an academic, marketing, and policy-making perspective.

Twitter (http://twitter.com) is one of the social media platforms that enables its users to generate and consume useful information about issues and trends from text streams in real-time. Twitter and its 500 million registered users produce over 340 million tweets, which are text-based messages of up to 140 characters, per day[1]. Also, users subscribe to other users in order to view their followers' relationships and timelines which show tweets in reverse chronological order. Although tweets may contain valuable information, many do not and are not interesting to users. A large number of tweets can overwhelm users when they check their Twitter timeline. Thus, finding and recommending tweets that are of potential interest to users from a large volume of tweets that is accumulated in real-time is a crucial but challenging task.

A simple but effective way to solve these problems is to use the number of retweets. A retweet is a function that allows a user to re-post another user's tweet and other information such as profile credit. Retweeting, similar to forwarding in email, is a key part of information sharing and propagation in Twitter. However, not all retweets are meaningful insights in information sharing. For instance, a tweet such as "blessed and grateful. thank you" from Justin Bieber, the most-followed person on Twitter, is interesting only to his followers and fans; the tweet actually has gotten a lot of retweets, but it contains mundane content and is not informative or useful. According to Boyd, Golder, and Lotan (2010), some users retweet to spread tweets to new audiences, using it as a recommendation or productive communication tool, whereas others retweet a message not because of its content, but only because they are asked to or because they regard retweeting as an act of friendship, loyalty, or homage towards the person who originally tweeted.

Basically, the more social links a user has, the better the chances the user's postings will spread on social media, but the propagation may be limited to the user's social network. On the other hand, the content that attracts large audiences can be easily propagated even if its author is not popular. In other words, to find tweets that are interesting to a large number of users, it is important to consider the content's popularity rather than the author's popularity. Hence, we need to use semantic analysis for short social text messages that usually include noisy and conversational contents. To solve these problems, we adapted the Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), a statistical and popular unsupervised model. This text clustering model has been widely applied to text mining problems, and does not require manually constructed training data. In a training step in LDA, we can collect a set of related words that co-occurred in similar documents, which is

* Corresponding author. Tel.: +82 2 3290 3195; fax: +82 2 929 7914.
E-mail addresses: mcyang@nlp.korea.ac.kr (M.-C. Yang), rim@nlp.korea.ac.kr (H.-C. Rim).

[1] http://en.wikipedia.org/wiki/Twitter

0957-4174/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.12.051

referred to as "topics" from the continuous text streams. Theoretically, the LDA models the two types of probability distributions as latent variables: the probability of words under each topic and the probability of topics under each document. Since these probabilities can indicate each topic's characteristic and quality, we can score the topic based on its generality and specificity. Given a set of test data, we can also obtain the topic distribution of the new tweet from a trained topic model. To analyze a large number of documents, we exploit an alternative analysis such as the latent topic-based analysis, rather than examine the documents closely as previous methods–that used surface textual features–have done.

In this study, we focus on the automatic method to identify tweets that may be of potential interest to a wide audience. We conduct LDA-based topical analysis and infer latent topics to understand the content of an individual tweets and to score their interestingness. Our contributions are summarized as follows:

- We propose a novel and unsupervised approach that identifies interesting contents for a wide audience in Twitter and filters uninteresting contents to prevent users from dealing with an overwhelming number of tweets.
- We model textual contents in Twitter as latent topic structures using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and propose Trend Sensitive-LDA (TS-LDA) that can reflects the trends over time for more effective analysis.
- Based on topic identification, we score the weight of topics as its relative importance. Finally, we select individual tweets that contain more important and interesting topics than other tweets.
- We use the Amazon Mechanical Turk (AMT)[2] platform to collect tweets labeled by various groups of people.
- We conduct extensive experiments on a real dataset crawled from Twitter. The results prove that our model is more effective than the existing models.

The rest of the paper is organized as follows. In Section 2, we examine previous works related to this paper. In Section 3, we describe how to measure the interestingness of an individual tweet using topical analysis. In Section 4, we present the experimental results and analysis. Finally, Section 6 concludes the paper.

2. Related work

Many studies have provided insights into social media. Kwak, Lee, Park, and Moon (2010) firstly studied Twitter's structure by investigating various Twitter features. Recently, many works have focused mainly on analyzing or obtaining valuable information, such as influential users and posts on Twitter, from a large amount of social data. The most existing approaches (Castillo, Mendoza, & Poblete, 2011; Duan, Jiang, Qin, Zhou, & Shum, 2010; Hong, Dan, & Davison, 2011; Uysal & Croft, 2011) proposed to regard retweet counts as a measure of popularity, influence, and interestingness, and presented classifiers that predicted whether and how often new tweets will be retweeted in the future. They exploited various features of Twitter, such as textual data, author's metadata, and propagation information. Although the overall retweet count indicates a tweet's popularity, this may apply only to the followers of the tweet's author.

Twitter not only has textual data but also has linking data, such as follow and retweet links, which enable us to construct a network structure. The link-based approaches (Romero, Galuba, Asur, & Huberman, 2011; Yang, Lee, Lee, & Rim, 2012) applied a variant of the link analysis algorithm to a designed link structure in order to find interesting messages. However, the link structure requires a large volume of linking data to be analyzed and constructed and cannot be updated effectively when new documents stream in. Alonso, Carson, Gerster, Ji, and Nabar (2010) used crowdsourcing to categorize a set of tweets as interesting or uninteresting and reported that the presence of a URL link is a single, highly effective feature for selecting interesting tweets with more than 80% accuracy. This simple rule, however, may incorrectly categorize an uninteresting tweet (i.e., an uninteresting tweet contains links to meaningless pictures, videos, and advertisements) as interesting. Lauw, Ntoulas, and Kenthapadi (2010) suggested several features to identify interesting tweets but did not experimentally validate them. For user recommendation, Armentano, Godoy, and Amandi (2012) examined the topology of followers/followees network and identified the relevant users using social relation factors. Armentano, Godoy, and Amandi (2013) conducted not only topology-based profiling but also content-based profiling to find semantically similar users.

In social media, semantic analysis and topic modeling are widely used to understand textual data and can facilitate many applications such as user interest modeling (Pennacchiotti & Gurumurthy, 2011), sentiment analysis (Lin & He, 2009), content filtering (Duan & Zeng, 2013; Martinez-Romo & Araujo, 2013), and event tracking (Lee, 2012). Zhao et al. (2011) analyzed the topical differences between Twitter and traditional media using Twitter-LDA for investigating short messages. Wang and McCallum (2006) and Kawamae (2011) conducted topic modeling of temporally-sequenced documents in Twitter and tried to model the topics continuously over time. However, in our approach TS-LDA regards the mixtures of latent topics as a trend of its publishing time and is designed to learn changes in topic distributions, while other works focus on learning topic shifts based on word distributions. Chen, Nairn, Nelson, Bernstein, and Chi (2010) focused on recommending URLs posted in tweets using various combinations of topic relevance and social graph information. The Labeled-LDA (Ramage, Dumais, & Liebling, 2010) modeled a tweet using its labeled information, and then built the probability distribution vector of latent topics to represent the tweet's content. Based on similarity between the topic vectors, the researchers tried to find tweets that are similar to the ones which are already annotated "interesting." In terms of topic inference, our model is based on the model by Ramage et al. (2010) but is an unsupervised learning method with relative importance of latent topics.

3. Identifying interesting contents on Twitter

We treat the task of finding interesting tweets as a ranking problem where the goal is to obtain a scoring function that gives higher scores to interesting tweets in a given set of tweets. Our strategy is to focus on re-ranking the most-retweeted tweets according to their interestingness. We first introduce the two major concepts used in this paper.

Definition 1. Interesting in social media means that the content may be of potential interest to not only the authors and their followers but a wider audience. On the other hand, **uninteresting** means that the content is only interesting to the authors and their friends due to personal interests.

Definition 2. Interestingness indicates the size of the tweet's audience. Specifically, it is measured by the number of users that may be interested in the tweet. Also, an **interesting tweet** refers to a post whose interestingness is larger than a specific threshold.

Note that these definitions are followed by Alonso et al. (2010) and Lauw et al. (2010). Also, the survey on interestingness measures (Geng & Hamilton, 2006) reported that *Generality/Coverage*

[2] http://mturk.com

in data mining represents patterns that cover a relatively large subset of a dataset. That is, we focus on finding the comprehensiveness of an entire dataset rather than the overrepresentation of a specific group, such as the relationship between a celebrity and his/her fans.

First, we describe Trend Sensitive-LDA to learn and infer latent topics in Twitter (Section 3.1). We then measure the weight of latent topics to distinguish interesting topics (Section 3.2). Finally, we score the interestingness of a target tweet using its topic probabilities and corresponding weights (Section 3.3).

3.1. Trend Sensitive-LDA

Algorithm 1 Generative process for Trend Sensitive-LDA

For each topic $t \in T$:
  Draw $\phi_t \sim \mathrm{Dir}(\beta)$
For each time period $p_i \in P$:
  $p_i = \{s_{i-k}, \ldots, s_i, \ldots, s_{i+k}\}$
  Draw $\eta_{p_i} \sim \mathrm{Dir}(\alpha)$
  For each tweet d at timestamp s_i:
    $\alpha_d = \eta_{p_i} \times \alpha$
    $\theta_d \sim \mathrm{Dir}(\alpha_d)$
    For each word position j in tweet d:
      Draw $z_{d,j} \sim \mathrm{Multi}(\theta_d)$
      Draw $w_{d,j} \sim \mathrm{Multi}(\phi_{z_{d,j}})$

We designed the Trend Sensitive-LDA (TS-LDA) model to generate topic distributions of new documents based on the trends at the time the document is written. In terms of textual topics, a trend is regarded as a variation of mixtures of predefined topics. Using textual and temporal features and their combined clustering, TS-LDA can identify users' constantly changing interests and social trends in Twitter. Since some topics are useless in finding interesting contents, we focus on detecting timely and compelling rather than general and mundane, topics.

Our assumption is that when a user wants to write a tweet, he/she first may choose a topic based on the current trends. Then, he/she chooses a bag-of-words one by one based on the chosen topics. The process of generating tweets is described in Algorithm 1 and illustrated in Fig. 1. The multinomial word distribution for each topic t (ϕ_t) is drawn from Dirichlet prior β. Time period p_i represents a series of timestamps $\{s_{i-k}, \ldots, s_i, \ldots, s_{i+k}\}$ with window size k and the timestamp unit is set to days. η_p denotes the topic distribution of tweets (trend) for time period p and the trend is applied to each tweet that is written at timestamp s_i. Like the Labeled-LDA (Ramage, Hall, Nallapati, & Manning, 2009), TS-LDA is similar to traditional LDA except for the constraint: topic prior α_d, is restricted to the set of topics that are popular at the time tweet d is published. Therefore, the multinomial topic distribution of document d (θ_d) is drawn from Dirichlet prior α_d.

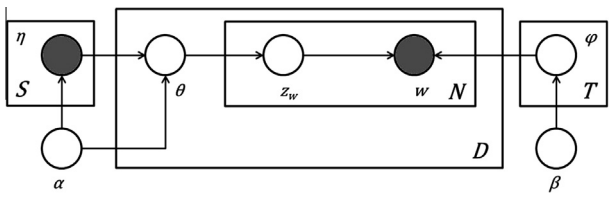

Fig. 1. Plate notation of Trend Sensitive-LDA.

3.2. Weighing the latent topics

The latent topics in LDA have two types of probability distribution: word w probability occurring in topic t ($p(w|t)$) and topic t probability occurring in document d ($p(t|d)$). Thus, we investigate the quality of the latent topics using the above information of the two probabilities. We will discuss the following observations: (1) integrity of topic, (2) spatial variation of topic, and (3) temporal variation of topic.

Integrity of Topic assumes that not all topics may be in appropriate for analyzing textual data, and is measured based on the significant words for each topic. To understand a topic's subject, many studies examine words that have a high probability of representing the topic since each topic has its own word distribution. On the contrary, if a topic consists of improper words (e.g., non-English or unusual words), the topic is not only obscure but utterly incomprehensible. We measure *Integrity of topic* by using a manually constructed dictionary as a deterministic factor.

The dictionary-based scoring method is explained as follows. First, since our target language is English, we collect basic words from an English dictionary (Table 1). Also, to extract the frequent proper nouns, we select the words whose document frequency is higher than 5 K times in the news corpus which contains 13.5 M documents. These extracted words (shown in Table 1) are important entities for representing specific objects, such as celebrities, organizations, brands, and events. Finally, our merged dictionary (*Lexical Dictionary*) is constructed of 36,678 stemmed words. We therefore define the *Integrity of topic (I)* measure as follows:

$$I(t) = \sum_{w \in W} p(w|t) L(w) \tag{1}$$

where $p(w|t)$ denotes the learned probability of word w given topic t, and $L(w)$ is a deterministic function that returns 1 if the *Lexical Dictionary* contains word w and 0 otherwise. If our dictionary includes a higher probability of words under a topic ($p(w|t)$), a higher integrity score of the topic is given. Unlike previous scoring models (Newman, Lau, Grieser, & Baldwin, 2010), this scoring function finds latent topics that are meaningful rather than coherent.

Spatial Entropy represents the distinction of topics based on each document's topic distribution. To discard noisy topics (e.g., topics with incoherent words) and general topics (e.g., topics consisting mainly of common words such as "hello," "people," etc.), we exploit the observation that the most meaningful topics are related to a small number of documents. If a topic is closely related to many documents, it is likely a noisy or general topic. *Spatial Entropy* is given by

$$S(t) = -\sum_{d \in D} p(d|t) \log p(d|t) \tag{2}$$

where document d is a single tweet and $p(d|t)$ denotes the learned probability of document d given topic t. The probability value is measured by $p(d|t) = p(d)p(t|d)/p(t)$ based on the Bayesian inference.

Temporal Entropy represents the distinction of topics based on the topic distribution for a specific period. Unlike *Spatial Entropy*, it detects topic changes in Twitter and utilizes a set of tweets that have the same timeline because a single tweet does not contain

Table 1
Example of words in Lexical Dictionary.

Basic dictionary words
People time love money game food life good work give school friend music ...
High-frequency words
obama romney david reuters paul facebook upi santorum phelps bbc cnn ...

Table 2
Top topics of proposed scoring functions.

Function	Topic t	Description
$I(t)$	photo post facebook album nicki minaj status wall wed arriving	Social media
$S(t)$	step completed world domination quest valor level upgrade	Game
$T(t)$	opening ceremony olympics close watch london perform tonight	Olympic ceremony

enough information to find distinct topics. To identify trendy topics (e.g., topics with trendy and bursty words), we exploit the observation that the most meaningful topics are related to a specific time period. If a topic is closely related to many time periods, it is likely a noisy or general topic. *Temporal Entropy* is given by

$$T(t) = -\sum_{s \in S} p(s|t) \log p(s|t) \tag{3}$$

where $p(s|t)$ denotes the learned probability of timestamp s given topic t and is measured by using Eq. (2) $p(s|t) = \sum_{d \in D_s} p(d|t)$. Also, the unit of timestamp s is determined as one day.

Normalizing Components adjust the weight between scoring functions that have different value distributions. Thus, we use *standard score (z-score)* of each weight function.

$$\tilde{x} = z(x) = \frac{x - \mu}{\sigma} \tag{4}$$

where x is a variable, μ is an average of variables, and σ is a standard deviation of variables.

Interesting Topics are identified by the final scoring function based on the above three components. The topic's importance is represented by the weight of latent topics to measure the interestingness of individual tweets. The weight of topics is given by

$$W(t) = \tilde{I}(t) - \tilde{S}(t) - \tilde{T}(t) \tag{5}$$

As stated above, we subtract the two entropy values ($\tilde{S}(t)$ and $\tilde{T}(t)$) from the integrity value ($\tilde{I}(t)$) because a topic that has less entropy is more stable and important. To describe the effectiveness of individual scoring functions, Table 2 reports the words that occupied high-probabilities[3] of the top ranked topics of the functions, and the description field shows the labeled summary with the ordered words. We can observe that these words are clear, interesting, and informative which helps in understanding tweet documents. For example, the topic on London Olympics, the biggest issue in the world, is assigned the highest $T(t)$ score. Table 3 reports the ranked list of topics and their weights; the topics about important issues and trends received high weights. From these interesting topics, we can detect global issues (e.g., the Olympics), and widespread issues (e.g., a social issue about the online abuse on Tom Daley and the accidental shooting in Colorado). On the other hand, the topics that are assigned low weights are meaningless, chatty, and violent.

3.3. Scoring the interestingness of Tweets

By using a topic identification as a proxy to find interesting tweets for wider audiences, the relative importance of each topic is assigned its weight according to Eq. (5). In order to extract a topic distribution of a target tweet, we conduct topic inference on a trained model. Finally, we can obtain the following scoring function for ranking tweet d:

$$Score(d) = \sum_{t \in T} W(t) p(t|d) \tag{6}$$

where $W(t)$ is the weight of topic t and $p(t|d)$ is the result of document d's topic inference. A tweet achieves a higher score if the tweet covers latent topics that have a high weight. For a given dataset, we can rank tweets based on their interestingness score.

4. Experiments

In this section, our proposed method is empirically evaluated over a large crawl of Twitter data. In Section 4.1, we describe our dataset and several preprocessing steps. Section 4.2 introduces the baseline systems for experimental comparisons. As shown in Section 4.4, the experimental result, which is based on manually constructed evaluation data (Section 4.3), demonstrates the effectiveness of our method.

4.1. Dataset and Preprocessing

We used the Twitter dataset collected from English-speaking users for evaluation. We used *Twitter REST API*[4] to facilitate the data collection. The majority of the tweets collected were published in a four-week period from July 24, 2012 through August 23, 2012. To eliminate incomplete and noisy data, document preprocessing is necessary to train the topic model on tweets. Our preprocessing method involved the following tasks: we discarded retweeted tweets, tweets with less than six words, and tweets with non-English words (more than 20%). We also removed meaningless words such as stop-words, URLs, user names, and special characters, and stemmed the remaining words. These steps reduced the number of tweets by about 23%, thus obtaining a final set of 79.6 M tweets. Finally, the LDA model was applied to a set of 1.55 M tweets. To make this set, 50 K randomly extracted tweets were added each day for 31 days.

4.2. Methods for comparison

We describe the baseline methods mentioned in Section 2 and compare them with our proposed method. The first baseline method, *#RT*, is obviously based on retweet counts; tweets with a higher retweet count are ranked higher than those with a lower retweet count. The second baseline method, *#URL+#RT*, favors tweets that contain URL links (Alonso et al., 2010). Since it is less likely for a tweet to contain more than one link, we additionally use *#RT* to break ties in tweet ranking. The third baseline method is based on various link analysis algorithms and is run on the retweet-based graph that is constructed by the user and tweet subgraphs. More precisely, *User-Hits* ranks tweets by their publisher's authority score based on the HITS algorithm (Kleinberg, 1999), while *PageRank* ranks tweets by their PageRank (Brin & Page, 1998) score. Also, Yang et al. (2012) scores individual tweets by using a variant of the link analysis algorithm that runs on the user and tweet subgraphs. For the last baseline method, we choose recent works (Castillo et al., 2011; Duan et al., 2010; Hong et al., 2011; Uysal & Croft, 2011) that address a problem related to ours and aim to predict the popularity of a given tweet. Although interestingness and popularity are two distinct concepts, these works present a wide range of features that may be applied to assess the interestingness of tweets using machine learning. For re-implementation, we trained the RankSVM classifier (Joachims, 2002) using 45 features that include 17 text-based features and 28 Twitter-specific features such as users' metadata, retweet links, follower links, and favorite user lists (Castillo et al., 2011; Duan et al., 2010; Hong et al., 2011; Uysal & Croft, 2011). We used four types of methods, each with its own scope: $ML_{message}$, ML_{user}, ML_{prop}, and ML_{all}. The first

[3] The words that are ranked by the LDA model were stemmed, but we transformed them to their original form for readability.

[4] https://dev.twitter.com/docs/api/1.1

Table 3
Predicted weights and their corresponding high probability terms that describe the particular topic.

$W(t)$	Topic t	Description
14.560	opening ceremony olympics close watch london perform tonight	Olympic ceremony
6.598	gold medal olympics win team london silver won proud teamgb	Olympics
6.348	tom daley diving peter proud riley board dad diver olympics	Social issue
5.853	shoot gun video omg shot aurora victim christian colorado	Accidental shooting
5.554	lost day found product pound lb magic week lucky told glad	Weight-loss diet
−4.484	im dont gonna lol bore wont glad put ill didnt kinda mood yea	Chatty words
−4.534	im gonna mad scare cuz bout piss kill tho turn glad freak hurry	Violent words
−5.952	lol tweet time remember hahaha guess lmao idk didnt xd aha	Chatty words

three variants rely only on message-based, user-based, and propagation-based features of tweets, respectively. The ML_{all} method exploits all kinds of features. We utilized the probability estimates of the learned classifier to rank a set of tweets, and we used leave-one-out cross validation to evaluate these supervised learning methods. For our method, *Proposed*, we set the number of topics to 500 (based on preliminary experiments) and ran 2,000 iterations of Gibbs sampling using the MALLET toolkit.[5] The window size on TS-LDA was 2.

4.3. Gold standard generation

Since there does not exist a test collection that can measure a tweet's interestingness, we manually constructed our own. The goal of this evaluation is to demonstrate that our approach is able to produce better ranked lists of tweets by giving interesting tweets high scores. For the evaluation dataset, we selected the top 50 retweeted tweets every morning and afternoon from July 24, 2012 to August 17, 2012 (25 days), but the tweets of duplicated authors were removed. For each tweet, seven annotators[6] that were randomly selected from the Amazon Mechanical Turk (AMT) platform were instructed to categorize the tweet as "interesting" or "uninteresting" after inspecting its content as described in Definition 1. The interestingness was based on the total number of annotators that labeled it as "interesting" according to Definition 2; thus, the final score ranges from 0 to 7. As Table 4 shows, the annotators labeled most of the tweets as "uninteresting" even if they were retweeted the most. The average rating was just 1.34, with the majority of tweets (1,063) labeled only as "uninteresting." We regarded each tweet as an "interesting tweet" if its final score was more than three points, according to Definition 2.

4.4. Experimental Results

Tweet interestingness prediction task models content-driven information recommendation scenario: given a set of tweets, interesting tweets are recommended to users. To evaluate this task, we split the set of judgments, organized as 50 separate sets, by morning and afternoon for each day. Like other recommender systems or search engines with multi-grade relevance judgments, we used *normalized Discounted Cumulative Gain (nDCG)* (Jrvelin & Keklinen, 2002) as the evaluation metric to compare the ranking performance of each method. The results are shown in Table 5. We observed that *#RT* alone is not a sufficient measure for ranking interesting tweets. Additionally, leveraging *#URL* is helpful but the improvements are only marginal. After manually inspecting the tweets with high retweet counts and URL links, it was found that many of the tweets were from celebrities. Their followers are most likely the only ones who are interested in these tweets which contain links to pictures of the celebrities. *User-Hits* and *PageRank*, which are retweet graph-based, fail to perform well, but their modified model (Yang et al., 2012) performs better than the previous methods. $ML_{message}$ always outperforms not only the first five models but also other single class based learning methods (ML_{user} and ML_{prop}). We observed that the length in characters and the number of words in a tweet are the two most effective message-based features for estimating its interestingness. The result of ML_{all} demonstrates that greater reasonable performance can be achieved when different class features are combined. Our proposed method significantly outperforms all the other baseline methods. The performance improvement of the proposed method over the best baseline (ML_{all}) on each run has been revealed to be statistically significant using the paired Student's t-test ($p < 0.05$). This is a significant result in that our method is an unsupervised approach that relies on only a few number of tweet features and does not require complex training with labeled data.

Table 4
Distribution of interestingness scores.

Score	# of Tweets	% of Tweets	Class
7/7	28	1.12	Interesting (20.28%)
6/7	57	2.28	
5/7	84	3.36	
4/7	122	4.88	
3/7	216	8.64	
2/7	336	13.44	Uninteresting (79.72%)
1/7	594	23.76	
0/7	1063	42.52	
	2500		

Table 5
Tweet ranking performance of methods that predict interestingness. The values in bold represent the highest scores in the tables.

Method	*nDCG*			
	@1	*@5*	*@10*	*all*
#RT	.283	.319	.366	.673
#URL+#RT	.368	.423	.456	.721
User-Hits	.153	.197	.220	.605
PageRank	.283	.319	.367	.675
Yang et al. (2012)	.421	.459	.483	.733
$ML_{message}$	.470	.492	.537	.768
ML_{user}	.396	.461	.483	.740
ML_{prop}	.256	.299	.352	.667
ML_{all}	.509	.559	.594	.794
Proposed	**.667**	**.633**	**.636**	**.822**

Interesting tweet classification task models content-driven filtering for information seekers: given a set of tweets, uninteresting or meaningless tweets are filtered. For each judgment set, the interesting tweets (20.28%) were used as positive examples, and the remaining were used as negative examples (79.72%) for supervised learning methods. Unlike the previous experiment, we used *precision at N (P@N)*, *r-precision (R-Prec)*, and *mean average precision*

[5] http://mallet.cs.umass.edu/

[6] 105 human annotators labeled several tweets in the entire dataset. In other words, the seven annotators are chosen among 105 people.

Table 6
Tweet ranking performance of methods that classify tweets. The values in bold represent the highest scores in the tables.

Method	*P@1*	*P@5*	*P@10*	*R-Prec*	*MAP*
#RT	.300	.284	.278	.261	.302
#URL+#RT	.380	.356	.324	.322	.366
User-Hits	.060	.156	.136	.150	.210
PageRank	.300	.284	.274	.264	.302
Yang et al. (2012)	.480	.408	.350	.362	.384
$ML_{message}$	.400	.444	.404	.403	.444
ML_{user}	.460	.392	.338	.349	.391
ML_{prop}	.200	.252	.252	.254	.282
ML_{all}	.600	.532	.458	.468	.516
Proposed	**.780**	**.572**	**.472**	**.493**	**.556**

Table 7
Ablation of the weighting factors. The values in bold represent the highest scores in the tables.

Method	*nDCG*			*P@1*	*P@5*	*MAP*
	@1	*@5*	*all*			
Proposed	**.667**	**.633**	**.822**	**.780**	**.572**	**.556**
w/o I (t)	.666	.623	.814	.760	.556	.527
w/o S (t)	.647	.584	.796	.740	.524	.497
w/o T (t)	.619	.601	.813	.680	.544	.527

(MAP) to evaluate the ranking problem with binary class judgments. Table 6 reports the performance across the same methods that the previous experiment used. This result also demonstrates that our proposed method outperforms all the other baseline methods in terms of the overall evaluation metrics. The difference in effectiveness between the proposed and the best baseline method (ML_{all}) for each run is also revealed to be statistically significant ($p < 0.05$, paired Student's t-test). In both tasks, our method achieves significant performance when the higher ranked tweets are evaluated (e.g., *nDCG@1* and *P@1*).

Table 7 shows the contributions of individual weighting factors for Eq. (5). For each row, we removed the weight components from each of the types described in Section 3.2. We can see the two entropy functions are complementary: *Temporal Entropy* function is better in finding the top-1 ranked interesting tweet (*nDCG@1*, *P@1*), whereas *Spatial Entropy* function plays a crucial role in the other metrics. We believe that the most interesting tweet is more likely to rely on the temporal issue.

To further analyze interestingness, we compared the top ranked sample tweets of our method with those of the typical methods (Table 8). From the *#RT* rows we can see that the most retweeted tweets contain only mundane or conversational content of the celebrities (e.g., Niall Horan and Ryan Lochte). For example, Ryan Lochte's tweet is not interesting; however, it contains many retweet counts via retweet requests to simply evoke attention. The URL links of the *#URL + #RT* rows do not provide useful information. For example, both links are for music videos of the famous singer Justin Bieber. The retweet graph-based link analysis method, Yang et al. (2012), cannot find the most interesting content in the English Twitter dataset as they can do in the Korean Twitter dataset because the users of the two datasets have different retweet behaviors. We observed that most Korean users participate in Twitter activities so that their real data can also show dynamic retweet acts. As content curators (Bhargava, 2009), many Korean-speaking users tend to continuously spot and retweet interesting tweets to share information; however, many English-speaking Twitter users do not do the same, and only opinion leaders have influence on social media such as CNN or BBC (Choi & Han, 2013). For this reason, our dataset that consists of tweets in English does not enable each user to play an important role as a hub-node which can indicate interesting or important tweets using the HITS algorithm. As described in the ML_{all} rows, the supervised learning method notably depends on the length features of a tweet. This method is similar to the typical high-quality document classification method which is based on the assumption that long documents include high-quality content. Although the examples of ML_{all} are long, they show trivial content such as an advertisement. The *Proposed* rows show tweets that contain a single, distinct, and meaningful theme; therefore, our proposed method can help find interesting topics and trends from the current buzz textual clusters through our topic model. For instance, James Harden's tweet is related to the opening ceremony of the London Olympics, and Mitt Romney's post discusses an important political issue of the 2012 U.S. presidential election.

5. Discussion

Topics of user-generated content are dynamic and uncertain. This implies for the users to constantly generate content in various points of view. Therefore, like in the Online Reputation Monitoring scenario (Amigó et al., 2013), specific topics are mainly fine-grained, on the other hand, general topics are coarse-grained. We believe that our method can handle arbitrary topic distributions

Table 8
Top ranked tweets of individual ranking methods.

Methods	Author	Text
Morning, July 27, 2012		
#RT	*Niall Horan*	Dont know if its just me, but the excitement in london is unbelievable! Its gona be a great games
#URL+#RT	*Alfredo Flores*	#AsLongAsYouLoveMeVIDEO Preview! http://t.co/yOPgMuYJRT to the WORLD! LET'S GOOO!
Yang et al. (2012)	*Debenhams*	It's time for #FaithFriday! Follow us & RT for a chance to win a beautiful pair of pale pink heels from Faith. http://t.co/Q3iSxcc9
ML_{all}	*Sainbury's*	To celebrate British Food Fortnight we're giving away 3 × 100 vouchers. RT & follow to enter. T&Cs: http://t.co/ihlVRuho
Proposed	*James Harden*	Today is Opening Ceremony for the 2012 Olympics!!!!! #TeamUSA
Morning, August 11, 2012		
#RT	*Ryan Lochte*	LOCHTE NATION! If I hit 1 million followers b4 I land on US soil tomorrow I'll pick a random follower to fly out for a photo & lunch RETWEET
#URL+#RT	*Justin Bieber*	people trying out from ALL AROUND THE WORLD. #BelieveTourAuditions - watch what he does with the basketball. #swaggy - http://t.co/a3kBV1Kg
Yang et al. (2012)	*Football Manager*	It's 12PM, 11th of August and I want YOU to win a copy of Football Manager 13 delivered to your door on release day!! Just RETWEET & FOLLOW!
ML_{all}	*Justin Bieber*	people trying out from ALL AROUND THE WORLD. #BelieveTourAuditions - watch what he does with the basketball. #swaggy - http://t.co/a3kBV1Kg
Proposed	*Mitt Romney*	I am proud to announce @PaulRyanVP as my VP. Stand with us today. http://t.co/0lH4WbTW #RomneyRyan2012

because unimportant topics tend to show a consistent pattern; general and mundane topics appear any time spans. In terms of the coverage of our method, its capabilities are as follow: (1) Our model is based on probability distributions of the trained topic model and can find interesting topics if the volume of input tweets is enough to distinguish each latent topic. (2) Our model is general that it is applicable to a variety of time spans of the collection if the appropriate topic number is determined. The longer time spans of the collection, the more latent topics should be captured to cover the whole issues at the time spans. (3) Our time-related variables on the model can be configured according to the target system. For example, if setting the unit of timestamp s to seconds, we can apply our model to real-time tweet recommendation.

6. Conclusion

In this paper, we proposed a novel method to discover interesting contents on Twitter using topic identification. To analyze textual data more effectively, we developed a new TS-LDA model to understand the trends over time. To measure the interestingness of an individual tweet, we first scored extracted latent topics based on our proposed functions. From the latent variables of TS-LDA, we scored the integrity of a topic by utilizing its representative words. Also, for each topic, we calculated two types of entropy values by examining the spatial and temporal variations of topic probabilities. Given our observations, each latent topic with a high weight value covers a specific topical theme. We then investigated how important topics spread with a target tweet. In a series of experiments, we demonstrated the ways in which our model can be naturally applied to recommend, filter, and understand textual posts in social media.

In terms of understanding a large number of documents, weighing and analyzing latent topics using the LDA model can reduce cost and complexity because the clusters from each topic can be viewed as a set of significant contents. In our future work, we plan to apply our proposed method to the personalized tweet recommender system based on the tweet history and social relations of users. For a time-varying dictionary, we also plan to facilitate automatic entity detection from news articles.

Acknowledgements

This research was partially supported by the Ministry of Knowledge Economy(MKE), Korea and Microsoft Research, under IT/SW Creative research program supervised by the National IT Industry Promotion Agency(NIPA) (NIPA-2012-H0503-12-1012) and by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Plannig (NRF-2012M3C4A7033344).

References

Alonso, O., Carson, C., Gerster, D., Ji, X., & Nabar, S. U. (2010). Detecting uninteresting content in text streams. In Proceedings of the SIGIR 2010 workshop on crowdsourcing for search evaluation CSE '10 (pp. 39–42).

Amigó, E., deAlbornoz, J. C., Chugur, I., Corujo, A., Gonzalo, J., Martín-Wanton, T., et al. (2013). Overview of replab 2013: Evaluating online reputation monitoring systems. In *CLEF*. In P. Forner, H. Müller, R. Paredes, P. Rosso, & B. Stein (Eds.). *Lecture notes in computer science* (Vol. 8138, pp. 333–352). Springer.

Armentano, M., Godoy, D., & Amandi, A. (2012). Topology-based recommendation of users in micro-blogging communities. *Journal of Computer Science and Technology, 27*, 624–634.

Armentano, M., Godoy, D., & Amandi, A. (2013). Followee recommendation in twitter based on text analysis of micro-blogging activity. *Information systems, 38*, 1116–1127.

Bhargava, R. (2009). Manifesto for the content curator: The next big social media job of the future? <http://rohitbhargava.typepad.com/>.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In Proceedings of the 2010 43rd Hawaii international conference on system sciences HICSS '10 (pp. 1–10).

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*, 107–117.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In Proceedings of the 20th international conference on world wide web WWW '11 (pp. 675–684).

Chen, J., Nairn, R., Nelson, L., Bernstein, M., & Chi, E. (2010). Short and tweet: Experiments on recommending content from information streams. In *Proceedings of the SIGCHI conference on human factors in computing systems CHI '10* (pp. 1185–1194). New York, NY, USA: ACM.

Choi, S.-M., & Han, Y.-S. (2013). Representative reviewers for internet social media. *Expert Systems with Applications, 40*, 1274–1282.

Duan, J., & Zeng, J. (2013). Web objectionable text content detection using topic modeling technique. *Expert Systems with Applications, 40*, 6094–6104.

Duan, Y., Jiang, L., Qin, T., Zhou, M., & Shum, H.-Y. (2010). An empirical study on learning to rank of tweets. In C.-R. Huang & D. Jurafsky (Eds.), *COLING* (pp. 295–303). Tsinghua University Press.

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys*, 38.

Hong, L., Dan, O., & Davison, B. D. (2011). Predicting popular messages in twitter. In Proceedings of the 20th international conference companion on world wide web WWW '11 (pp. 57–58).

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining KDD '02* (pp. 133–142). New York, NY, USA: ACM.

Jrvelin, K., & Keklinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems, 20*, 2002.

Kawamae, N. (2011). Trend analysis model: Trend consists of temporal words, topics, and timestamps. In *Proceedings of the fourth ACM international conference on web search and data mining WSDM '11* (pp. 317–326). New York, NY, USA: ACM.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46*, 604–632.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on world wide web WWW '10* (pp. 591–600). New York, NY, USA: ACM.

Lauw, H. W., Ntoulas, A., & Kenthapadi, K. (2010). Estimating the quality of postings in the real-time web. In Proceedings of the WSDM 2010 workshop on search in social media SSM '10.

Lee, C.-H. (2012). Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams. *Expert Systems with Applications, 39*, 13338–13356.

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management CIKM '09* (pp. 375–384). New York, NY, USA: ACM.

Martinez-Romo, J., & Araujo, L. (2013). Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications, 40*, 2992–3000.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics HLT '10 (pp. 100–108). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pennacchiotti, M., & Gurumurthy, S. (2011). Investigating topic models for social media user recommendation. In *Proceedings of the 20th international conference companion on world wide web WWW '11* (pp. 101–102). New York, NY, USA: ACM.

Ramage, D., Dumais, S., & Liebling, D. (2010). Characterizing microblogs with topic models. In Proceedings of the fourth international AAAI conference on weblogs and social media. AAAI.

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the 2009 conference on empirical methods in natural language processing: Vol. 1 – Vol. 1 EMNLP '09 (pp. 248–256). Stroudsburg, PA, USA: Association for Computational Linguistics.

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011). Influence and passivity in social media. In Proceedings of the 2011 European conference on machine learning and principles and practice of knowledge discovery in databases ECML-PKDD '11 (pp. 18–33).

Uysal, I., & Croft, W. B. (2011). User oriented tweet ranking: A filtering approach to microblogs. In C. Macdonald, I. Ounis, & I. Ruthven (Eds.), *CIKM* (pp. 2261–2264). ACM.

Wang, X., & McCallum, A. (2006). Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining KDD '06* (pp. 424–433). New York, NY, USA: ACM.

Yang, M.-C., Lee, J.-T., Lee, S.-W., & Rim, H.-C. (2012). Finding interesting posts in twitter based on retweet graph analysis. In W. R. Hersh, J. Callan, Y. Maarek, & M. Sanderson (Eds.), *SIGIR* (pp. 1073–1074). ACM.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., et al. (2011). Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European conference on advances in information retrieval ECIR'11* (pp. 338–349). Berlin, Heidelberg: Springer-Verlag.