Blog Topic Analysis Using TF Smoothing and LDA

Sungwoo Lee
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Korea
+82 31 290 7987
lsmoney@skku.edu

Jaedong Lee
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Korea
+82 31 290 7987
ultrajaepo@skku.edu

Chang-Yong Park
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Korea
+82 31 290 7987
codep@skku.edu

Jee-Hyong Lee
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Korea
+82 31 290 7154
john@skku.edu

ABSTRACT
In the era of Web 2.0, the number of blogs has explosively increased. With the appearance of social network services, blogs has become the places for sharing professional knowledge and personal branding. So, in order to understand the trends of topics or to analyze the content of blogs, the time sensitive topic extraction and topic change analysis is important and necessary. In the previous studies, most of topic extraction models extracted topic words independently from each time slice and tried to combine those. However, these methods did not show a good performance in analyzing topic trends because the topics extracted from time slices are independent. To cope with this problem, we propose a term frequency smoothing method which weaves time slices so that the more related topics are extracted from each time slice and a better topic trend analysis is generated. In order to extract topics from smoothed term frequencies, LDA, a generative topic model, is adopted. The evaluation of the proposed method on IT blogs shows that it can effectively discover quite meaningful topic patterns and topic words.

Categories and Subject Descriptors
D.3.2 [**Information Management**]: Internet search and Information integration – *selection process.*

General Terms
Algorithms, Measurement, Performance, Experimentation

Keywords
topic trend change, blog text mining, LDA, term frequency smoothing

1. INTRODUCTION
Everyday people put tremendous variety of information on the World Wide Web. The World Wide Web has become a useful knowledge base for anyone who is seeking for an answer or sharing an information. Today one of the most informative and useful tool on the Internet is blog. The definition of blog provided by Wikipedia is, "A blog (short for web log) is a user-generated website where entries are made in journal style and displayed in a reverse chronological order." People can use the blog to publish what is happening in his or her daily life or opinions about anything, such as politics, entertainment, or product reviews. According to the blog search engine Technorati [14], approximately 120,000 blogs are created daily. In the early, a bloggers upload their personal life and useful information. But, over time, a blog service had been under a spotlight as one of the business model. Many bloggers began to consider how can induce a steadily influx of visitors as way of revenue. With these concerns, the purpose of the blog operation was changed to a personal branding and sharing professional knowledge from a diary[5].

In general, because special and latest information can be provided through the blog service, web users search an information in the blog and subscribe to the blog. Recently, social network service like facebook, twitter and me2day gets the explosive popularity. And social network service begins to replace the diary service of the blog. But, due to the nature of the social network service, there is a limit to provide an specialized information. So, still in the future, the blog service still should have played a important role as an window providing good quality information. The next 2-3 years, only the blog service which accumulates specialized contents will be able to survive.

Then, by operating this specialized blog service, what is a benefit that bloggers can obtain? First, they can obtain a fame depending on their personal brand extension. By posting specialized information on their blog service, blogger can show their personal brand related to the field to people. Second, they can raise income through advertising on their blog service. Utilizing advertising platform like Naver's Ad Post, Daum's Ad Clix, Google's AdSense, bloggers can raise income.

When bloggers operate specialized blog service that drags an attention of the people and raises income, they have difficulty in a selection of a topic. So, how does bloggers choose the proper topic for their interest? If bloggers see a topic trend changes, they easily select the topic and write blog posts for the topic. The goolge offers the 'Google Trends' service. In the ‘Google Trends’ service, web users can compare the world’s interest in their favorite topics. If web users enter up to five topics, they can see the topic trend changes on the ‘Google Trends’ over time [8].

ICUIMC(IMCOM)’13, January 17–19, 2013, Kota Kinabalu, Malaysia.

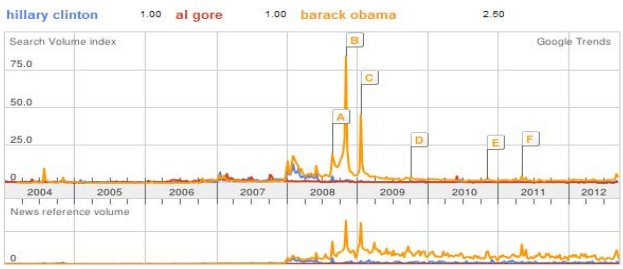


Figure 1. 'Google Trends' service

This trends service is useful for interest in people's favorite topics. But, this service has some weakness. The 'Google Trends' only shows the trend in user's interest. So, when users show trends for a wide range of topic, there is a limit. Services like 'Google Trends' are not suitable for taking a look at the overall trend changes. So far, there have been significant efforts in terms of temporal effects on topic trend changes. A topic trend changes detection researches have almost exclusively focused on analyzing the topic trend changes in each time slice. For example, the Topic Detection and Tracking (TDT) works [1, 2, 3, 9, 10, 11, 16] have focused on detecting new events and tracking known events in each time slice. However, this method~~s~~ revealed limits in the accurate topic trend changes detection. In general, when an event occurs, people say about the event over time. Similarly, bloggers post their contents for an event or interest. When we estimate topic trend changes, we consider this characteristic. In this paper, we propose TF Smoothing method that identifies the topic trend changes and extracts words related with a topic. The basic idea of our approach is to treat documents within a certain period of time into a single document. In other words, we consider that the updated documents in adjacent time zone are similar. Because LDA shows outstanding performance in topic extraction, the term frequency smoothing applies to LDA, generative topic model [4].

The rest of the paper is organized as follows. We first review the related work in the Section 2. Then we define LDA in the Section 3 and present the proposed TF Smoothing method in the Section 4. We report the experimental results in the Section 5 and conclude in the Section 6.

2. RELATED WORKS

So far, to trace topic trend changes over time, many studies were conducted. Studies are divided into two areas: estimating the topic trend changes over time and extracting a main topic or bursty events. First, the works on Topic Detection and Tracking [1, 2, 3, 9, 10, 11, 16] all aim to detect and track events from a stream of news stories. However, all of studies consider documents in each time slice. this researches do not consider similarity of between adjacent documents at certain times. Second, studies were conducted to extract the main topic of the blog. In 2002, Aixin Sun at al.[2] propose a extraction method for a blog's main topic by using tag information. In [11], an infinite automaton was proposed to identify bursty features and their bursty structures; it has been used in [12] to identify the bursty evolution of a blog space. However, the work is restricted in only identifying bursty features one by one and does not consider the time feature to find topic trend changes. Because such method does not consider the characteristics of the document being uploaded over time, such methods have problem that the results are often quite sensitive to the time. Our method is more robust since it identifies the topic trend changes by pooling together words which are close in time. We propose the TF Smoothing model for the exact topic trend changes with an temporal information. In general, a document consists of multiple words. Through term frequency of these words, a single document can be represented by term frequency vector matrix. In previous studies, when researchers consist of the document-word matrix, they only consider term frequency of words in a day. But, because there are several posts in chronological order, this configuration causes inaccurate results when we estimate the topic trend changes. When the blog post is uploaded, it is likely that these blog posts are affected by an event. Also, a topic of blog posts in the adjacent time zone will be similar to that. So, we will consider the temporal continuity of blog posts. And we apply the term frequency smoothing method for words in a specific time zone.

3. LDA (Latent Dirichlet Allocation)

LDA(Latent Dirichlet Allocation) is generative probabilistic model of a corpus[4]. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by an distribution over words. LDA assumes the following generative process for each document w in a corpus D :

1. Choose $N \sim Poisson(\xi)$.
2. Choose $\theta \sim Dir(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n,\beta)$, a multinomial probability conditioned on the topic z_n.

In this model, the value of α and β is determined by the corpus unit. the value of N and θ is determined by the document unit. β is the parameter of the Dirichlet prior on the per-topic word distribution. N is the length of the document. The θ represents the weight of each topic in the document. The z_n is a topic vector of n-th word in a document. the number of topics is fixed k. Therefore the θ and the z_n are the k length of the vector. In other words, if observations are words collected into documents, it determines that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.

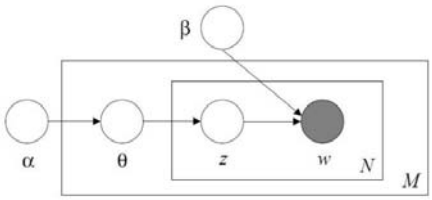


Figure 2. Latent Dirichlet Allocation model

The figure2 summarizes this process. If we successfully model the document in this way, it may be possible to find the parameter θ of the already known document. Even if the topic of document d_1 and d_2 is similar, the type and frequency of words is different in each document. Because of this, there is limit to calculate the similarity or topic identification by a simple keyword-based model. However, if we know α, β and calculate θ, similarity calculation and classification works are much more easy and accurate.

3.1 PARAMETER ESTIMATION

The process described above can be expressed by the following formula.

$$p(z_1,...,z_n) = \int p(\theta)\left(\prod_{n=1}^{N} p(z_n \mid \theta)\right)d\theta \quad (1)$$

$$p(w,z) = \int p(\theta)\left(\prod_{n=1}^{N} p(z_n \mid \theta)p(w_n \mid z_n)\right)d\theta \quad (2)$$

The formula 1 is the topic generation in the document. The formula 2 is the generation of a topic and a word in the document. z representing the document topic is the conditional probability.

$$\underbrace{p(\theta \mid data)}_{posterior} \propto \underbrace{\ell(data \mid \theta)}_{likelihood}\underbrace{\tilde{p}(\theta)}_{prior} \quad (3)$$

So, what we want to do estimates the parameter θ from documents. If we know a distribution form of a posterior, this work is much easier. And then, a conjugate prior necessity comes from right here. In formula 3, this conjugate prior for a multinomial distribution is a Dirichlet distribution. In other words, just by regarding prior as the Dirichlet distribution, we can calculate posterior easily.

3.2 MODEL SIMPLIFICATION

Given the parameter α, β and w, the conditional probability can be written as follows:

$$p(\theta, z \mid w, \alpha, \beta) = \frac{p(\theta, z, w \mid \alpha, \beta)}{p(w \mid \alpha, \beta)} \quad (4)$$

But, it is intractable to compute this probability. So, as shown in figure3, this model should be simplified.

$$q(\theta, z \mid \gamma, \phi) = q(\theta \mid \gamma)\prod_{n=1}^{N} q(z_n \mid \phi_n) \quad (5)$$

A variational parameter γ and φ should seek to minimize the KL divergence between the variational distribution and the original distribution. the parameter α and β is estimated to use this values and observed document. The figure4 is the LDA analysis for AP corpus. Through this process, we identify an topic for words.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Figure 3. LDA analysis for AP corpus

4. TERM FREQUENCY SMOOTHING METHOD

We propose the term frequency smoothing method. This method assumes that uploaded blog posts at the adjacent time zone are mentioned for a similar topic. First, we explain the tf-idf (term frequency-inverse document frequency) method. Tf–idf, term frequency–inverse document frequency, is a numerical statistic which reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in the information retrieval and text mining [15]. The term frequency *tf(t,d)* is the number of times that term t occurs in document *d*.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (6)$$

The inverse document frequency is a measure of whether the term is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

$$idf_i = \log\frac{|D|}{|\{d_j : t_i \in d_j\}|} \quad (7)$$

$$tf - idf_{i,j} = tf_{i,j} \times idf_i \quad (8)$$

In the term frequency, the higher the term frequency value, the more important word in the document. Through this value, we can extract topic words in the document or classify topics of documents. So far, the previous studies adjust term frequency values to document classification models. But, in analysis problem of documents with the time information, when the term frequency value was just applied, inaccurate topic words were extracted. For example, when the iPhone4s had been released, in IT professional blog site, blog posts related to the iPhone4s were uploaded over time. However, in this period, besides the iPhone4s, a variety of information related to IT issues will also be uploaded. This noisy information will make it hard to figure out the main topic, in a certain time zone. Therefore, we apply more weight for words related to the topic and less weight for words unrelated to the topic. And then, we can make more the accurate document-word matrix. The method is similar bag-of-words model. In this model, a text such as a sentence or a document is represented as an unordered collection of words.

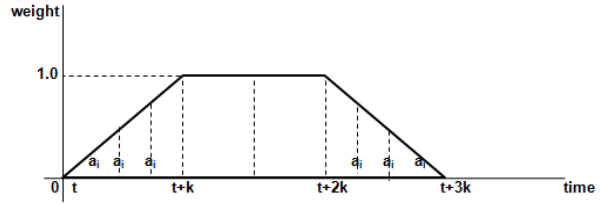


Figure 4. Graphical representation for tf smoothing method

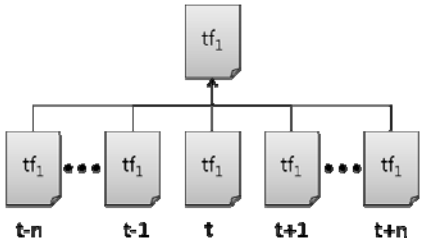


Figure 5. Term frequency smoothing method

The figure 4 and 5 show main principle for the term frequency smoothing method. Documents mentioned a few days are to be treated as a single document. Given a weighted value to each word before and after a specific time zone, important topic words are extracted. Then, the term frequency for words of the uploaded documents is smoothed. Through this process, the document-term matrix reflecting the time continuity is composed. Using this smoothed term frequency and the LDA, we can estimate a topic distribution of words and documents. The formula of the proposed method is as follows:

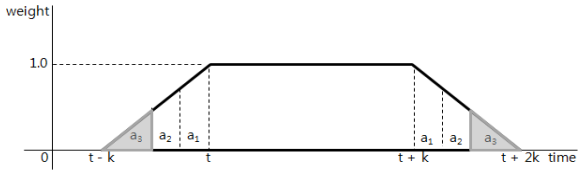
$$tf_{i,t\sim t+k} = \sum_{i=t}^{t+k-1} tf_i \times \alpha_i + \sum_{i=t+k}^{t+2k} tf_i + \sum_{i=t+2k+1}^{t+3k} tf_i \times \alpha_i \quad (9)$$

In above the formula, a parameter α is the smoothing weight on each word. An importance for mentioned each word before and after a specific time zone are different from each other. In case upload time for mentioned each word further away from baseline time zone, an importance for each word is lower. The

term frequency of each word in the document is readjusted. And then, we more accurately estimate the blog topic trend changes and extract topic words throughout the readjusted term frequency for each word.

4.1 AUTOMATIC DECISION FOR WEIGHT VALUE

We illustrated the tf smoothing method in the Section 4. This method needs weight value which is applied to the documents.



We decide weight value by the expert knowledge. But, to exploit the expert knowledge can cause uncertain results in topic trend changes. So, it is necessary to automatically calculate a weight value. Before, we mentioned that an importance for mentioned each word before and after a specific time zone is different from each other. We assume that weight value have linear function characteristic before and after a specific time zone. Because mentioned each word before and after a specific time zone has less relationship for an event, it is reasonable assumption. So we propose automatic decision method for weight value. When we calculate weight value, we regard it as a problem for the width of the shape surrounded by the graph of the linear function. In the figure 10, we regard it as a full width range of from t to t-k. And, to determine the value of each weight, we regard weight value as the ratio of the area occupied each width by each from the entire width.

5. EXPERIMENTAL RESULTS

To evaluate our methods of identifying blog topic trend changes, we conduct experiments on an IT blog data set. In general, a topic patterns in a blog are highly related to major real world events. In this section, we show that our proposed method can identify meaningful topic patterns from the IT blog site to reveal the major real world events with appropriate time line.

5.1 DATA SET AND PARAMETER SETTING

Our IT blog data consists of 10 years' blog posts of GigaOM IT professional blog site from 2001 to 2011. There are altogether 2,704 blog posts. The main body of each blog post is extracted through Goose html parser, the html extraction java api[7]. By applying the proposed term frequency smoothing technique, we make the document-term matrix. Using this document-term matrix and the LDA, we trace blog topic trend changes and identify how much accurate is words for each topic trend through the survey. We compare results from the LDA with the term frequency smoothing and LDA without term frequency smoothing. In the LDA model, there are several user input parameters which provide flexibility for the topic trend changes analysis. These parameters are set empirically, as in principle, it is impossible to optimize these parameters without relying on domain knowledge. We set the number of topic = 10, iterations = 200, α and β = 0.5 in the following experiments.

5.2 RESULTS

The experiment is performed in three cases:

(1) Documents in a day merged into a single document with the general term frequency.

(2) Documents in consecutive three days merged into a single document with the term frequency smoothing.

(3) Documents in consecutive six days merged into a single document with the term frequency smoothing.

In each cases, we observe the blog topic trend changes and extract top10 topic words.

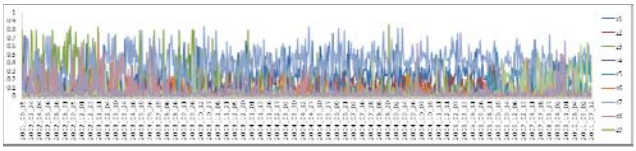


Figure 6. GigaOM blog site trend with a windows size of one day

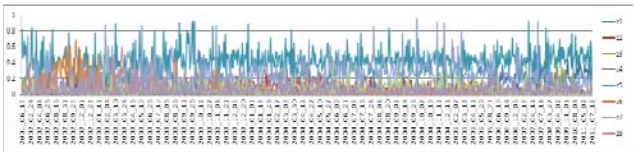


Figure 7. GigaOM blog site trend with a windows size of three day

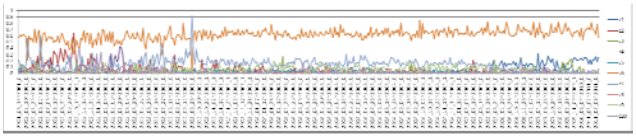


Figure 8. GigaOM blog site trend with a windows size of six day

In the figure 6, 7, 8, we confirm that each topic trend changes become increasingly evident from one day to six day. In the figure5, we do not figure out the topic trend changes. But, in the figure7, 8, we can see more the accurate topic trend changes. From the figure7, we can see that the probability of the z6 topic is more than 50%. Because the z6 is the background topic in the blog, this results appear. The background topic within the document is commonly mentioned word set. subject, article, pronoun,..etc. In this graph, the observation is conducted to the selected two topics. Next, we explain the google-social trend changes. In the figure 9, 10, 11, we can see the google-social trend changes. In this graph, the topic trend changes are more smoothed from 2005 to 2007. Also, we identify that blog posts related to the google-social have been steadily uploaded during this period. The topic trend changes from the October 2009 to the July 2011 are similar between the proposed and original method. During this period, because the proportion of blog posts related to the google-social is high.

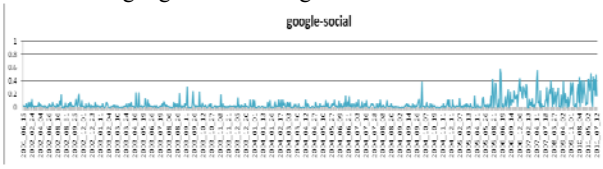


Figure 9. "Google Social" trend with a windows size of one day

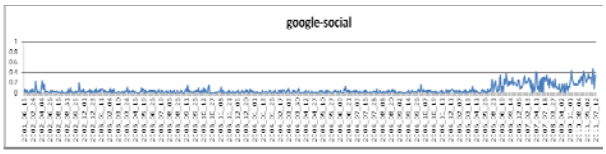


Figure 10. "Google Social" trend with a windows size of three day

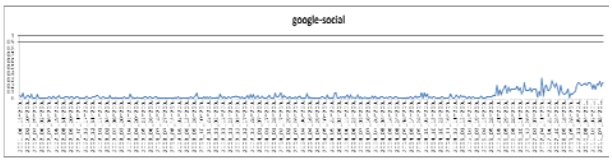


Figure 11. "Google Social" trend with a windows size of six day

The table1 shows top10 topic words for the google-social topic. In case of the one day, 's', 'google', 'skype', 'search', 'social', 'we', 'web', 'network' and 'media' are extracted. In case of the three day, 'google', 'web', 's', 'search', 'users', 'social', 'we', 'just', 'skype' and 'facebook' are extracted. In case of the six day, 'google', 'social', 'web', 'user', 'search', 'still', 'users', 'better' and 'most' are observed. In the one day, 'google', 'social' and 'skype' are noticeable. The 'skype' word is interesting. In the one day, this word is top3 ranked. But, in three day, this word is top10 ranked. And then, in six day, this word is disappeared. Also, words 'google' and 'social' are top1,2 ranked. When we applied the proposed method, we confirm that words related to the topic are higher ranked and words unrelated to the topic are lower ranked. We perform survey to evaluate how accurate is extraction words.

Table 1. Top10 words for google-social topic

1 day		3 day		6 day	
s	0.025856	google	0.01817	google	0.013903
google	0.025621	web	0.008742	social	0.006356
skype	0.01483	s	0.007014	web	0.006144
search	0.012672	search	0.006916	user	0.005509
social	0.008262	users	0.006175	search	0.004874
we	0.007652	social	0.005682	still	0.004169
web	0.007464	we	0.005632	users	0.003957
network	0.007089	just	0.005583	better	0.003605
media	0.006432	skype	0.005336	most	0.003322
its	0.006198	facebook	0.004003	site	0.003322

The experiment participants are asked to answer the accuracy score point for extraction words in each day. The score is from 1 to 10. The participants are 5 people.

Table 2. Accuracy of extraction words in each day

Topic	1day	3day	6day
Google-social	7.2	8.2	8.7
Qwest-bankruptcy	4.3	7.6	8.5

In above the table2, when the proposed method is applied, we confirm that extraction words is more accurate from one day to six day. The figure 12, 13 and 14 show the topic related to the Qwest-communications company. Like the google-social topic graph, topic trend changes are increasingly clear from one day to six day. Also, we observe interesting results for extraction words in the table3. In the one day, 'new', 'offering', 'company', 'million', 'market' and 'telecom' are extracted. Through these words, we figure out this topic is just related to an communication company. But, we can see that an words appear increasingly from one day to six day. In six day, 'qwest' and 'bankruptcy' are ranked high. Looking for words that are extracted, we confirm that words which clarify the topic appear from one day to six day. We can see that blog posts related to the 'qwest-bankruptcy' have been mentioned a lot between June 2002 and October 2002. During this period, a company called KPN-Qwest Dutch was bankruptcy.

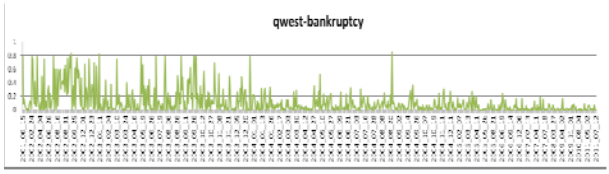


Figure 12. "Qwest bankruptcy" trend with a windows size of one day

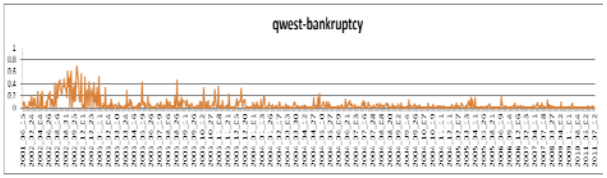


Figure 13. "Qwest bankruptcy" trend with a windows size of three day

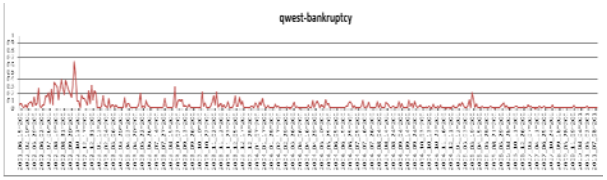


Figure 14. "Qwest bankruptcy" trend with a windows size of six day

Table 3. Top 10 words for Qwest-bankruptcy topic

1 day		3 day		6 day	
s	0.019839	telecom	0.010771	telecom	0.008057
new	0.017635	former	0.00716	worldcom	0.007051
offerings	0.013334	worldcom	0.006604	qwest	0.005313
company	0.010312	according	0.005771	york	0.005222
million	0.010277	wall	0.005771	former	0.005131
market	0.00889	chief	0.005632	mci	0.004673
telecom	0.008357	york	0.005563	inc	0.004125
revenue	0.007824	billion	0.005354	firm	0.004033
billion	0.00775	bankruptc	0.00507	mr	0.00403

	3	y	7		3
technology	0.007646	stock	0.004868	bankruptcy	0.003942

6. CONCLUSION

In this paper, we proposed the term frequency smoothing method which estimates blog's topic trend changes accurately. Throughout this method, we can extract words in each topic. To interpret each topic was quite useful. As shown in the experiment, if this term frequency smoothing method is applied to topic classification model, it will be able to track the topic trend changes. And it will be helpful to blog topic selection. Also, We evaluated our method on IT professional blog site. The effectiveness of proposed method is proved. Since the proposed method is general, it can be applied to any topic model to discover topic trend changes. In the future, we will further study the problem of identifying topic trend changes. And, we will study an summarization method of the topic trend. We also plan to apply our method to other topic classification model.

7. ACKNOWLEDGMENTS

This work was supported by the IT R&D program of MKE/KEIT. [KI001810041244 , SmartTV 2.0 Software Platform]

8. REFERENCES

[1] Aixin Sun, Maggy Anastasia Suryanto, and Ying Liu. 2007. Blog classification using tags: an empirical study. In *Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers* (ICADL'07). Springer-Verlag, Berlin, Heidelberg, 307-316.

[2] Aixin Sun, Ee-Peng Lim, and Wee-Keong Ng. 2002. Web classification using support vector machine. In *Proceedings of the 4th international workshop on Web information and data management* (WIDM '02). ACM, New York, NY, USA, 96-99. DOI=http://doi.acm.org/10.1145/584931.584952.

[3] ChengXiang Zhai, Atulya Velivelli, and Bei Yu. 2004. A cross-collection mixture model for comparative text mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '04). ACM, New York, NY, USA, 743-748. DOI=http://doi.acm.org/10.1145/1014052.1014150 .

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993-1022.

[5] Darren Rowse, Chris Garret. 2008. PROBLOGGER:Selects for Blogging Your Way to a Six-Figure Income.

[6] George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3 (March 2003), 1289-1305.

[7] Goose html parser, https://github.com/jiminoc/goose/wiki.

[8] Google Trends Service, http://www.google.com/trends.

[9] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. 1998. Topic detection and tracking pilot study: Final report. In Proceedings of the DARPA.

[10] James Allan, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '98). ACM, New York, NY, USA, 37-45. DOI=http://doi.acm.org/10.1145/290941.290954

[11] Jon Kleinberg. 2002. Bursty and hierarchical structure in streams. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '02). ACM, New York, NY, USA, 91-101. DOI=http://doi.acm.org/10.1145/775047.775061.

[12] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. 2003. On the bursty evolution of blogspace. In *Proceedings of the 12th international conference on World Wide Web*(WWW '03). ACM, New York, NY, USA, 568-576. DOI=http://doi.acm.org/10.1145/775152.775233

[13] Salton G. and McGill, M. J. 1983. Introduction to modern information retrieval.

[14] Technorati, http://www.technorati.com/.

[15] Wikipedia, http://en.wikipedia.org/wiki/Tf%E2%80%93idf.

[16] Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat. 2007. Mining correlated burstytopic patterns from coordinated text streams. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '07). ACM, New York, NY, USA, 784-793. DOI=http://doi.acm.org/10.1145/1281192.1281276.

[17] Yiming Yang, Tom Ault, Thomas Pierce, and Charles W. Lattimer. 2000. Improving text categorization methods for event tracking. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '00). ACM, New York, NY, USA, 65-72. DOI=http://doi.acm.org/10.1145/345508.345550.