A Topic Study of Microblog Based on Specific Events

Zhenyu Zhou, Fang Li
Department of Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, China, 200240
{lesbuy,fli}@sjtu.edu.cn

***Abstract*--This paper describes the study of topics on microblog based on specific events. First, we use a famous topic model – LDA to extract topics from microblog about events. Then, we propose three indexes: Attention Factor(AF), Evolution Factor(EF) to see the performance of microblog topics and Diversity Factor(DF) to calculate the divergence of topics from microblog and news reports. Finally, we choose corpuses for four events to study. The experiments show that, on specific events: 1) There are more critical topics, while factual topics less, and both of them get close AFs. 2) Critical topics last long on microblog and have lower EFs, which means their contents vary little, but factual topics last intermittently and their contents vary greatly. 3) To compare with the same events from news reports, critical topics use totally different words, but factual topics use close words.**

Keywords: topic model, microblog, trend

I. INTRODUCTION

The communication media play an important role in the process of getting information by the people. In the past, the unique path for getting information is news or periodical. As Web 2.0 era comes by, people like to receive such events and messages by new types of media, such as weblog, forum, microblog etc. Take microblog as an example, it contains only a little information, as its word count limits to 140, people express their opinions more, interactively and immediately. To automatically analyze the contents of microblog, find out the characteristics of topics, and the trend of content and attention of topics is realistically meaningful.

In this paper, we try to solve three main problems:

1. How topics perform on microblog? What kind of topics do people focus on more?

2. How topics evolve on content and attention as the time goes on microblog?

3. How topics on microblog differ from those on news reports by content?

To solve the three problems, we propose three indexes: Attention Factor(AF), Evolution Factor(EF), Diversity Factor(DF), to analyze the performance of topics on microblog by quantitive data. Our work contains three parts. First, we use topic model – LDA to extract topics of microblog about specific events, then define the three indexes and derive correspondent formulas. Finally, experiment them on four specific events which represent four types. In this paper, section 2 gives some relevant research on this field, and section 3 describes the detailed approach. Section 4 gives the experiments results and some explanations and section 5 is the summary and research outlook.

II. RALATED WORK

So far, researchers often use LDA model[1] and its extensions[2-3]. LDA is an unsupervised method for machine learning without train data. It has been widely used in the topic extraction of news reports or other kind of documents. Hong[4] uses LDA model on twitter corpus for extracting topics, which proves its practicability on microblog. Zhao[5] uses twitter-LDA, which considers the short-message characteristic of microblog, and takes the author information into consideration. He regards topics as a distribution on authors. Also, he assumes the words is generated by topics or the background words, and a variable is set to control them. Some other researchers consider the tags and emoticons, which are soul of microblog, and use unsupervised machine learning like Label-LDA[6].

Researchers like to summary the type of topics to find out the different performances of different types of topics by proposing diverse features. Ramage[7] find that if we divide topics into four types: style, social, status and substance, we can see that in a celebrity's twitter, style and status topics occupy more while in a official organization's, substance topics top. Zhao[5] set three kinds of topics: long-standing, like global warming, entity-oriented, like Michael Jackson, event-oriented, like Haiti Earthquake. On twitter, long-standing topics occupy the most, and event-oriented the least, less than 10%. In aspect of opinion sentences, long-standing topics also performance the best, which means they attract more people to discuss on. Given the proportion of retweets, event-oriented topics do the best, which fits our common sense.

Also, researchers try to analyze the performance of microblog on specific events[8-9]. Qu[8] extracts those topics about Yushu earthquake, and find out opinionated topics occupy the most and the retweet and spread actions of each kind of topics.

Our purpose is like what in [5], but we focus on specific events, and propose three indexes on analysis. The difference between ours and [8] is that we use topic model on extracting topics from microblog.

III. APPROACH

In this paper, we first mine the corpus from microblog about specific events. After some pre-processing, we use LDA model to extract topics from the corpus which is divided by day. Then, we propose Attention Factor to find out how popular the topics are. After defining the Evolution

Published by Atlantis Press, Paris, France.

Factor to identify the same topic of adjacent days, we get an evolution path, which helps to show how topics evolving. Finally, we propose a method to identify the same topic of microblog and new reports to find out the word difference of microblog compared to news reports.

In this paper, we mainly discuss two types of topics:

Critical Topic: People discuss on some phenomenon or entity. E.g. stop inhospitality, scold morality status, joke on dictator etc.

Factual Topic: The description of objective reality. E.g. Tiangong I launches, interview witnesses, reports on mourning around the world etc.

A. *Topic modeling*

LDA model is a generating-probability model, which is 3-layer Bayesian model with alterable variables. It assumes the words are generated by a multinomial distribution on topics, and topics are generated by a multinomial distribution on documents. Such multinomial distribution is generated by dirichlet distribution on pre-defined variables. When a document comes, it is first to generate the multinomial distribution to get a topic, then use the multinomial distribution on this topics to get a word.

In our research the corpus is dispersed by day. So we use LDA model on the corpus of each day, and get several topics every day about each event.

B. *Attention Factor*

This index reflects the degree of a topic people discuss on. We can simply get a probability of a topic on each document by LDA model. If we simply average those to get a degree of a topic, a problem occurs. Because of the shortness of microblog, there may be only one effective word in a document after pre-processing, so the probability of this topic on the document is 1. But the importance of this kind of document should not be so overestimated, which causes over large of the average. So we set lower weight on such kind of documents to calculate the average probability.

We define the cover strength of topic z on some day is:

$$s(z) = \sum_{d_i \in D} \theta_{d_i,z} \varphi_{d_i} \quad (1)$$

θ means the probability of topic z on document d_i, and φ_{d_i} means the weight of document d_i based on its effective word count. So we define the Attention Factor of topic z is the cover strength normalization by all the topics on some day:

$$AF(z) = \frac{s(z)}{\sum_{z_i \in T} s(z_i)} \quad (2)$$

T means the topic set of that day.

C. *Evolution Factor*

The Evolution Factor of a topic indicates the variation degree of a same topic of adjacent days. The results of LDA describe the distribution of topics on documents, and distribution of words on topics. We can calculate the semantic similarity of a distribution to judge if there exists evolution relationship of topics of adjacent days. The most common approach is Jenson-Shannon Divergence. Imagine that at time t, a topic z of micoblog is expressed as z_t and the distribution of its vocabulary set V_t is p_{z_t}. At time t+1, it is expressed as z_{t+1} and the distribution of its vocabulary set V_{t+1} is $p_{z_{t+1}}$. The vocabulary sets V_t and V_{t+1} are sampled from different days, so we must expand them to a bigger vocabulary, and set the occurrence count of specific words to 0. So we define the EF(evolution factor) as JS divergence of the two distributions:

$$EF(z) = JSdiv(p_{z_t} \parallel p_{z_{t+1}}) = \frac{1}{2}\Big(KLdiv(p_{z_t} \parallel m) + KLdiv(p_{z_{t+1}} \parallel m)\Big) \quad (3)$$

Where $m = \frac{1}{2}(p_{z_t} + p_{z_{t+1}})$, and $KLdiv(P \parallel Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}$.

D. *Diversity Factor*

This factor describes the diversity in content of a same topic on different media. We will take news reports into consideration to find out the characteristics of contents of microblog. We try to compare the topics from the two kinds of media, but it won't perfectly work if we simply use JS divergence to calculate the distribution divergence of topics on the two media. To describe the same topic, people on microblog may use oral vocabularies while news reports official ones. That often occurs on the words those dominate the topic, so the divergence of the topic is exaggerated due to the dominant word diversity. In fact, after LDA modeling, each topic is expressed as a set of words, we define the concepts below:

To a topic z and a word w, if $p_z(w) > \xi$ where ξ means a threshold, we regard the word w a support word of topic z, or $w \prec z$, so the support word set of topic is defined as follow:

$$D(z) = \{w | w \prec z, w \in V\}$$

To a topic z_1 from microblog and a topic z_2 from news reports, we can define their intersection and union.

$$z_1 \cap z_2 = D(z_1) \cap D(z_2)$$
$$z_1 \cup z_2 = D(z_1) \cup D(z_2)$$

We can find from the formula that, the intersection means the set of common words, and the union means the set of all words. The bigger the intersection is, the more similar their semantemes are. We define the word diversity υ_{z_1,z_2} as $\frac{|z_1 \cup z_2| - |z_1 \cap z_2|}{|z_1 \cup z_2|}$, which means the proportion of specific words. λ is a weight that controls the word diversity. So we define DF(diversity factor) as follow:

$$DF(z) = (1-\lambda) JSdiv(p_{z_1} \parallel p_{z_2}) + \lambda \upsilon_{z_1,z_2} \quad (4)$$

IV. Experiments and Analyses

We choose two impactive events, Xiaoyueyue Event and the Death of Kim Jong Il, as the research corpus. We get the microblog corpus by the API of sina weibo. To compare with news reports, we also get news corpus from sina. Before experiment, we preprocess the microblog corpus as follows:

1. Exclude microblog posted by unavailable users.
2. Exclude microblog duplicated more than 20 times.
3. Exclude all the hashtags.

Our experiments include the three indexes above.

A. *Experiment data*

We get 208601 posts about Xiaoyueyue Event from Oct. 17th to 28th , 2011, and 122 news reports for comparison. We get 339589 posts about the Death of Kim Jong Il from Dec. 19th to 31st , 2011, and 623 news reports for comparison. All we use are whole texts. We use Gibbs Sampling[10] for LDA modeling. We set 6 topics for each event.

B. *Analysis of AF*

Table I shows the AF results of Xiaoyueyue Event. It shows the 6 topics extracted and their AFs on the fourth day.

From the result we can find that, all the 6 topics in Table I are critical topics, but none of them are factual topics. E.g. The topic with the highest AF is about people discussing stopping to be cold and distant, and the 4th high one is about people's mercy to the poor girl. Some factual topics on news reports, such as the process of rescue, the interview at site of incident, etc are excluded in microblog. It is also interesting that the AFs of the topics are close to each other, which indicates that, people pay similar attention on those topics, with little disparity.

Table I AF result of Xiaoyueyue Event on 4th day

	Topic (top 4 words)	AF
1	冷漠,良心,停止,政府	0.1777
2	生命,路人,谴责,司机	0.1724
3	社会,爱心,良知,温暖	0.1671
4	社会,孩子,一路走好,可怜	0.1648
5	见义勇为,保护,立法,法律	0.1637
6	法律,社会,老人,援助	0.1543

Table II The variation of topic words (Critical topic Morality)

道德 老人 冷漠 冷血	→ 社会 道德 问题 法律	→ 社会 道德 中国 问题	→ 道德 法律 问题 帮助	→ 道德 中国 社会 中国人	→ 社会 中国 道德 良心	→ 社会 道德 悲剧 问题	→ 社会 道德 法律 良知	→ 社会 道德 中国 中国人	→ 道德 中国 中国人 良知	→ 道德 问题 世界 悲剧	→ 冷漠 道德 社会 中国

Table III The variation of topic words (Factual topic Mourning)

逝世 领导人 最高 同志	→ 去世 韩国 逝世 哀悼	→ 逝世 悼念 金日成 国内	→ 去世 哀悼 平壤 表示	→ 默哀 联合国 大会 联大	→ 默哀 联合国 拒绝 安理会	→ 默哀 去世 韩国 联合国	→ 喜鹊 默哀 哀悼 逝世	→ 逝世 喜鹊 去世 哀悼	→ 平壤 灵车 民众 送别	→ 联合国 葬礼 逝世 降半旗	→

Table IV The performance of formula (4) compared to Jenson-Shannon Divergence

Event	Formula	Precision	Recall	F1
Death of Kim Jong Il	Baseline(JS-Div)	0.6473	0.6855	0.6659
	Formula (4)	0.6875	0.7021	0.6947
Xiaoyueyue	Baseline(JS-Div)	0.4286	0.5526	0.4828
	Formula (4)	0.5454	0.6316	0.5854

C. *Analysis of EF*

EF shows the variation trend of a topic with the time passes. Table II shows the evolution path of critical topic Morality of Xiaoyueyue Event, and Table III shows the evolution path of factual topic Mourning of the Death of Kim Jong Il.

From the results, we can find that critical topics vary little as the time passes. The main words those dominate the topic are samely, most are "morality", "society" and so on. But to factual topics, the dominated words vary a lot. E.g. In Table III, the cores on the topic vary from Korea to UN, from magpie to Pyongyang. They are all discussing mourning but the leading role changes.

D. *Analysis of DF*

First of all, we propose formula (4) to calculate the distance of distribution on topics of different media. The experiments show that it performs better than simply using Jenson-Shannon Divergence. Table IV shows the results.

We can see that our formula performs better in precision, recall and F value, especially of the Death of Kim Jong Il. That means our method of decreasing the distance of topics whose core words are expressed totally different in different media is available.

The threshold of formula (4) is set to 0.64, which is the experiment result. Figure I shows the performance of F value of different thresholds. In Figure I, the square plots means the Death of Kim Jong Il, while the rhombus ones means Xiaoyueyue Event. We can see the curves reaching its top at about 0.64. So we choose the value as the threshold of formula (4).

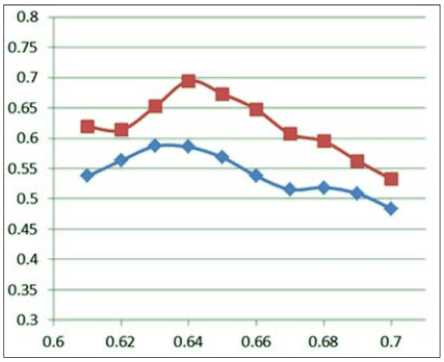

Figure I The Threshold-F curve of formula (4)

To find out the word characteristics of topics of microblog, we make a comparison with news event. We use the formula of DF to calculate the distance of topics of different media, and choose some topics whose distance is less than the threshold. We can try to find out the performance of word differences by the two kinds of topics. Table V shows the result.

Table V Topic words from different media

topic	Factual Topic Mourning	Critical Topic Dictatorship
Micro blog	去世,韩国,逝世,哀悼,表示,美国,政府,时代,日本,关注	金正恩,全世界,日头,领袖,生活,国度,独裁者,金日成,媒体,历史
News report	韩国,表示,逝世,半岛,美国,总统,日本,稳定,政府,消息	金正恩,接班人,父亲,军事,委员长,媒体,国际,工作,成为,电影

From the results, we can find that factual topics often have a lower DF, which means the topic of microblog has only a few difference to news report, while critical topic performs reversely. Though the topics of different media express the same meaning, they may use totally different core words. That means there are often some arbitrary and informal words on microblog, instead of official and formal ones on news reports.

V. SUMMARY

This paper uses LDA to model the topics, and propose three indexes to see how critical topics and factual topics perform on the microblog. We may have some conclusions as follows:

1. Critical topics occupy more on microblog, while factual topics not.
2. Critical topics vary little with time, but factual topics not. Critical topics often last during the whole span of the event, but sometimes, factual topics may not.
3. The words of critical topics are totally different to those on news reports, but factual topics similar. It reflects that netizens often express their feeling or opinions of a topic in optional ways instead of official and formal words like the news reports.

In the future, we may take more into consideration on our research. For example, to discover a more available way to relate topics and to analyze the differences of topics from more aspects. Specifically, we may try more types of topics, e.g. topic of natural disaster, topic of people's livelihood, topic of politics, topic of economy etc. Among them, some are originated from microblog itself, but some are originated from news reports. They may perform totally different in some other ways.

ACKNOWLEDGMENT

National Natural Science Foundation of China (60873134)

REFERENCES

[1] D.M.Blei, A.Y.Ng, and M.I.Jordan. Latent Dirichlet Allocation. The Journal of Machine Learning Research, 2003, vol.3, pp.993-1022.
[2] D.M.Blei, J.D.Lafferty. A Correlated Topic Model of Science. The Annals of Applied Statistics 2007, Vol.1, No.1, pp.17-35.
[3] D.M.Blei and J.D.Lafferty. Dynamic Topic Model. In International conference on Machine Learning, 2006, pp.113-120.
[4] LiangjieHong,and B.D.Davison.Empirical study of topic modeling in Twitter. Proceedings of the SIGKDD Workshop on SMA,2008
[5] Xin Zhao, Jing Jiang, JianshuWeng, et al. Comparing Twitter and traditional media using topic models. In Proceedings of the European Conference on Information Retrieval, 2011
[6] D.Ramage, S.Dumais, and D.Liebling. Characterizing Microblogs with Topic Models. Proceedings of AAAI on Weblogs and Social Media, 2010
[7] D.Ramage, D.Hall, R.Nallapati, et al. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2009
[8] Yan Qu, Chen Huang, Pengyi Zhang, et al. Microblogging after a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake. Proceedings of the ACM 2011 conference on Computer supported cooperative work, 2011, pp.25-34.
[9] S.Vieweg, A.L.Hughes, K.Starbird, et al. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness.Proceedings of the 28th International Conference on Human factors in computing systems, 2010, pp. 1079-1088.
[10] S.Geman, D.Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, Vol.6, No.6, pp.721–741.

Tips:
Due to the complexity of Chinese, it is not convenient to translate specific Chinese words to English, so all the topic words in this paper are remained Chinese.