News Comments Generation via Mining MicroBlogs

Xuezhi Cao † Kailong Chen† Rui Long† Guoqing Zheng† Yong Yu†

† Dept. of Computer Science
& Engineering
Shanghai Jiao-Tong University
Shanghai, China
{cxz,chenkl,longrui,gqzheng,yyu}
@sjtu.edu.cn

ABSTRACT

Microblogging websites such as Twitter and Chinese Sina Weibo contain large amounts of microblogs posted by users. Many of these microblogs are highly sensitive to the important real-world events and correlated to the news events. Thus, microblogs from these websites can be collect as comments for the news to reveal the opinions and attitude towards the news event among large number of users. In this paper, we present a framework to automatically collect relevant microblogs from microblogging websites to generate comments for popular news on news websites.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval—*Information Filtering*

General Terms

Algorithms, Experimentation

Keywords

Twitter, News Events, Relevance, Content Analysis

1. INTRODUCTION

Microblog has become one of the most popular online social networking, such as Twitter and Chinese Sina Weibo. It allows users to share information with their friends or the public by posting text messages of up to 140 characters, which is called a microblog. Most users update their microblog frequently and over 200 million tweets are generated on Twitter per day. Microblog messages are sensitive sensors of detecting important world events and highly correlated to the news medias [4, 3]. Utilizing these microblog streaming data, our framework automatically collects relevant comments for popular news on news websites.

News comments have become an important component of published news on many news websites such as Yahoo and AOL. People can obtain opinions and attitudes towards the news events and engage in discussion. However, many people are reluctant to discuss about events with strangers publicly. Some news get few comments which is not objective to analyze the attitudes of the public. With the explosion of Microblog, Internet users have more opportunities to express their opinions about the world events than before [2]. From this point of view, microblog is a rich source for collecting varieties of opinions from the public.

Past research has focused on event detection by analyzing microblogs [3] [1] and concluded microblogs is very sensitive to real-world events. Other works also made comparing between the topics from microblogs and from traditional media [5], states that microblogs may have different aspect of view from the news articles. By summarizing all these facts, we conclude that microblogs are good resources for generating comments for news. In this paper, we propose a framework to mining relevant news comments from Microblog data. In the remainder of this paper, we discuss the technical details of the framework and experimental results.

2. CALCULATING RELEVANCE SCORE

First we extract the keywords from both news data and microblog data. Noun phrases , name entities , hashtags are regarded as keyword candidates. Intuitively, if a news event are heatedly discussed in microblogs, the frequency of news-related word has an obvious appreciation in time series. In practise, we use the ratio between maximum frequency and the average frequency of the given keyword as its weight. The weight for a keyword w is calculated as:

$$G_w = \sqrt{\frac{Max(w) + \alpha}{Avg(w) + \beta}} \tag{1}$$

Here $Max(w)$ and $Avg(w)$ indicates the maximum and the average frequency of term w in days during the time period. $\alpha(= 5\ in\ experiments)$ and $\beta(= 5)$ are smoothing parameters. This time series based keyword filtering strategy considers the importance of a word in a global view. To consider the content relevance between news and microblogs, we use the co-occurrence times of keywords as the basic evidence. In order to distinguish title fields and content fields in news, we redefine the content relevance score as below:

$$\begin{aligned} L(w, N) =& \lambda_{content} \times \#w\ appears\ in\ content\ field\ of\ N \\ &+ \lambda_{title} \times \#w\ appears\ in\ title\ field\ of\ N \end{aligned} \tag{2}$$

$\lambda_{content}(= 15)$ and $\lambda_{title}(= 40)$ are parameters which indicates the importance of words appearing in title and content of the news.

Integrating the content relevance and news related measurement of a word, given a microblog S and a news article

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

N, the relevance score is calculated as below:

$$R(S,N) = \sum_{w \in K} T_w \times L(w,N) \times G_w \quad (3)$$

Here T_w is the term frequency of the word w in status S. In our time series keyword filtering strategy, K are keyword candidates with weight G_w larger than a threshold $G_{thr}(= 1.5)$.

However, some microblogs related to the news are not retrieved to the top because they have few words occurred in the news article. For example, informal language such as nickname or abbreviation used in microblogs will cause this issue. We solve this problem by applying pseudo relevance feedback. We refine the keyword set and re-calculate the term weights by taking the top k ranked microblogs into consideration. These microblogs are used to enlarge keyword set which is not appear in the original news article. The pseudo relevance feedback is embodied by modifying the value of occurrence time $L(w,N)$:

$$L'(w,N) = L(w,N) + \gamma_1 \sum_{i=1}^{k} (k+1-i) R_{w,i} + \gamma_2 M_w^2 \quad (4)$$

Here $R_{w,i}$ is the term frequency of w in the i^{th} microblogs retrieved, $\gamma_1(= 1)$ and $\gamma_2(= 1)$ are parameters, and M_w is the number of microblogs which contains w in the top k microblogs.

A microblog related to a news event is more likely to appear in a short period before or after the news article is published. We modify the relevance value between microblog and news according to the length of time period between their appearance. The final relevance value $R'(S,N)$ is calculated by multiplying the relevance value by a decay coefficient function:

$$R'(S,N) = R(S,N) \times (1 - \frac{D^2}{\lambda_T}) \quad (5)$$

Where D is the the length of time period between their appearance, $\lambda_T(= 1000)$ is the parameters to control the decaying speed.

3. EXPERIMENTS

We use Sina Weibo, the largest Chinese microblogging website as experimental data source. We crawled Sina Microblog from January 1th, 2011 to March 8th, 2011 through streaming API and finally collected 20 million microblog messages published by 6.8 million users. In news data set, 71 popular news on Sina.com are picked up and classified into two categories: political news and entertainment news. For each news, microblogs are retrieved according to relevance score. Each resulting microblog is manually labeled as relevant or non-relevant to the given news article and the performance are measured by using mean average precision.

We begin by evaluating the performance of our method for keyword detection. For comparison, we use all noun phrases, name entities,hashtags from news and microblogs to form the keywords set as a baseline. The experiments show that selecting and weighting the keywords with frequency analysis highly improve the baseline (Table 1). To evaluate the effects of pseudo relevance feedback and decaying with time method, we compare the experimental results within and without applying them. Figure 1 shows that they all have positive effect on the ranking performance. Also we observe

MAP@100	Political	Entertainment
Baseline	58.88%	51.85%
Keyword Filtering	65.96%	72.86%
MAP@50	Political	Entertainment
Baseline	63.34%	57.62%
Keyword Filtering	68.18%	76.67%

Table 1: Effect of Keyword Filtering Strategy.

that they perform better in the political category comparing to the entertainment. For pseudo relevance feedback, the size of the keyword set after feedback is 1.3 time as large as the origin. But the pseudo relevance feedback method may bring noisy keywords and reduce the precision. People often gossip about a singer and use many informal words which are not relevant to the news article. That is why relevance feedback gets more improvement on the political category because tweets about politics are often written officially and formally. In the future, we plan to use the collected comments from microblogs to do opinion classification and sentiment analysis for comments summarization.

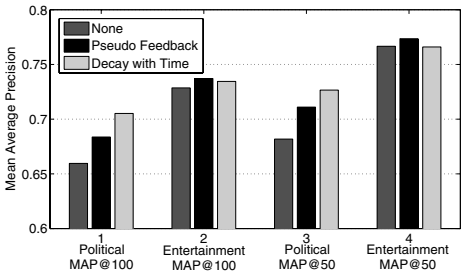

Figure 1: Comparison of Feedback and Decay methods

4. ACKNOWLEDGEMENTS

The team is supported by grants from NSFC-RGC joint research project 60931160445.

5. REFERENCES

[1] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. In *ICWSM*, 2011.

[2] L. Gandy, N. Nichols, and K. Hammond. Shout out: integrating news and reader comments. In *WWW*, pages 1095–1096. ACM, 2010.

[3] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860. ACM, 2010.

[4] J. Sankaranarayanan, H. Samet, B. Teitler, M. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *GIS*, pages 42–51. ACM, 2009.

[5] W. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. *Advances in Information Retrieval*, pages 338–349, 2011.