A Topic Detection and Tracking System with TF-Density

Shu-Wei Liu and Hsien-Tsung Chang

Abstract. In the past, news consumption took place predominantly via newspapers and were hard to track. Nowadays, the rapid growth of the Internet means that news are continually being shared and stored on a previously unimaginable scale. It is now possible to access several news stories on the same topic on a single web page. In this paper, we proposed a topic detection and tracking system with a new word measurement scheme named TF-Density. TF-Density is a new algorithm modified from the well-known TF-IWF and TF-IDF algorithms to provide a more precise and efficient method to recognize the important words in the text. Through our experiments, we demonstrated that our proposed topic detection and tracking system is capable of providing more precise and convenient result for the tracking of news by users.

1 Introduction

With the development of new technologies in recent years, people nowadays can engage in an increasingly wide range of activities on the Internet, e.g. making new acquaintances, shopping, e-mail and reading news online via dedicated news sources. In a FMCG report on the behavior of netizens, the foremost two most popular activities are "news reading" (65%) and "music appreciation" (65%). In fact, reading news online is so common among adults that it has become a necessary component of the daily routine of many.

However, there exists a glut of news sources on the Internet, such as [1,2]. Visiting just one news source no long suffices for many who desire different perspectives and are eager to partake in discussions on breaking or developing stories. As such, Internet users are spending ever more time to visit different news sources and

Shu-Wei Liu · Hsien-Tsung Chang
Department of CSIE, Chang Gung University, Taoyuan, Taiwan
email: ftcloud@gmail.com, smallpig@widelab.org

F.L. Gaol (Ed.): Recent Progress in DEIT, Vol. 1, LNEE 156, pp. 115–120.
springerlink.com

perform internet searches on related topics, to the extent certain web services such as GoogleNews [3] have taken the initiative to gather over 350 news sources and classify them into different categories, e.g. politics, sports or art. Similar news will be clustered under the same news topic, e.g.Republicans Divided on Obama's Proposal to Extend Middle-Class Tax Cuts, with the topic containing the various news stories which Google has gathered from the different news sources. Such services have vastly reduced the time that users spend on searching for different news sources.

Topic Detection and Tracking (TDT) tasks are commonly utilized to structure news stories from newswires and broadcasts into topics [4]. When a story is received by the system, the system will gauge whether the story is breaking news or old news. If the stream has never before been seen, it may very well be breaking news. In this age of information explosion, people regularly resort to the search engines to determine "what is new" or "what is going on" in the world. However, with an explosion of information and documents on the internet, new tools are needed to better organize this information.

Just as in the corporate world, laypeople are also eager to be the first to know when a particular company releases certain information on its products, to save on purchases or to make a quick buck. In this paper, we focused on Chinese articles, whose search focus is different from that of English articles, and proposed a new "term weight algorithm" for structuring news. Our algorithm can help people more efficiently and conveniently locate the news they desire.

2 Related Works

Topic detection and tracking (TDT) techniques have been under development for several years [5], with various algorithms for calculating the terms for clustering the media stream. TFIDF [5, 6, 7] is a widely used algorithm in TDT, and its essence is that if a term appears several times in a story, it is considered important. If the term appears in a few sources, it is considered an important word in the story. TFIDF is specifically on the lookout for two features: DF and TF, but in some situations, the term will be overestimated or underestimated. Therefore, so someone proposed a new algorithm, TFIWF [8, 9, 10], which is different from TFIDF in that it uses a new feature WF (word frequency). The algorithm will be more efficient and accurate in assigning term weight for structuring the stream. In [11] the objective is to extract the useful terms and filter the noise terms. After term filtering, the process will become more accurate and efficient in TDT. And in [12], the focus is on cluster efficiency and cluster accuracy. For the cluster efficiency, an indexing tree has been proposed to speed up the cluster structure and to reweigh certain terms such as the glossary of people's name or certain key points in the article, to bestow greater weight than others and render the cluster more accurate.

3 System Architecture and Algorithms

Our system is running as following steps. The first step was data collection. We collected news stories in Chinese by using the RSS technology. RSS is a family of

web feed formats used to publish frequently updated work [13].We subscribed to feeds from various Chinese news websites and stored these data on our server. Since the news data were in the HTML format, for the second step we analyzed the news data to extract useful data, including the titles and main bodies. In the third step, after extracting the titles and main body contents, we sent them to the CKIP (Chinese Knowledge Information Processing Group) system constructed by Academia Sinica [13] to split the article into terms. After the third step, we were able to determine the word composition of the article .In step four, we processed the filter module which would remove the words that contribute nothing but noises, as well as stop words (such as is, are, you, I). Finally a term list was created for each story. For the fifth step, we created a term information table including term information such as TF (term frequency),DF(word frequency)and WT(total number of terms) that are useful for weighting the article. Next, we applied our proposed TF-Density algorithm to weigh the term list in every story and create a vector list for every story. In the last step, we clustered the stories using cosine similarity.

The Incremental TF-IDF algorithm[5, 6, 7]has been widely used in information retrieval and text mining. The TF-IDF value consists of two components: the TF and IDF values. The TF(term frequency) refers to how many time the term *w* appears in the story *d* while the IDF(inverse document frequency) refers to the number of documents containing the term *w*.

The TF-IWF algorithm was proposed by [8], and it further incorporates the wf (word frequency) to more accurately assign weight to terms of importance in the article. The wf (word frequency) of the term w at time t is calculated as follow:

$$wf_t(w)=wf_{t-1}(w)+wf_{st}(w) \tag{1}$$

Where $wf_t(w)$ refers to the number of times the term *w* appears at the time *t*, $wf_{st}(w)$ refers to the number of times the term *w* appears before the breaking story at the time *t*, $wf_{t-1}(w)$ refers to the number of times the term *w* appears before the time *t*, and w_t refers to the total number of times the term *w* appears before the time *t*.

$$weight_t(d,w) = \frac{tf(d,w)log\ ((w_t+1)/(wf_t(w)+0.5))}{\sqrt{\sum_{w\prime\in d}(tf(d,w\prime)log\ ((w_t+1)/(wf_t(w\prime)+0.5)))^2}} \tag{2}$$

We proposed a new TF-Density algorithm. The algorithm combines features from the TFIDF and TFIWF algorithms, because there are two importance features in these two algorithms, namely *DF*(Document Frequency) and *WF* (word frequency), that we wanted to retain while providing a more precise and efficient method to recognize the important words in the text. Our approach is that we will calculate the number of times each term appears in all the documents, and divide it by number of documents in which the term *w* appears. This way, we can obtain the density of the term *w*, i.e. the average number of times it appears in all the documents. Therefore, if a particular term *w* appears more times in the stories than the average density, it is likely that this particular term in this particular story may be more important than others. The formula is given as follows:

Where WF refers to the number of times the term *w* appears in all documents, and *DF* refers to the number of documents in which the term *w* appears.

Therefore, we can calculate $Density_{term}$, which refers to the average number of times the term w appears in one document.

$$Density_{term} = \frac{WF}{DF} \tag{3}$$

$$weight_t(d,w) = \frac{TF(d,w)/Density_{term}}{WF/Wt} \tag{4}$$

We use cosine similarity to calculate the similarity between two stories d and d'atthe time t with the formula given as follows:

$$similarity = cos(\theta) = \frac{A \times B}{|A||B|} \tag{5}$$

$$similarity_t(d,d') = \sum_{w \in d \cap d'} weight_t(d,w) \times weight_t(d'w) \tag{6}$$

When a news story breaks at the time t, it will become a candidate for a new topic and be compared with previous topics, by means of their pair-wise similarities. If the similarity value is greater than the threshold $d\theta$, the topic will be considered as having previously appeared before the time t. If the value is less than the threshold, the candidate will be deemed a new topic. This way, we can ascertain whether the topic is old or new.

Our algorithm will execute as following. Step 1 every new incoming story will be put into a new cluster, which will contain just one story.Step2 To calculate the pair-wise similarities for each Topic clusters, we choose the top 30% term vector for calculating with one another since, because in the Chinese news context after CKIP, there are many noise words but not stopwords such as Modal or Adjective. Therefore, in order to prevent the unnecessary words from affecting the performance, we apply this method to filter unnecessary words. Step3 finally, we can know the largest similarity θ with new Cluster and Cluster B in all clusters. If θ is larger than the threshold $d\theta$. We conclude that A appears to be a new topic, so we combine A and B into a new cluster. If θ is smaller than the threshold, then cluster B will be considered a new topic.

Each new topic clusters compared with previous topics to check whether the story is new. When a new cluster is compared with Topic cluster B, if the cluster B contains only one story, the calculation is straight or ward. However, if cluster B has more than one story, we will combine all the term vector for the stories in the cluster, and choose the top 30% terms vector according to the term weight for calculation. This is due to the fact that news topic has the Times feature, so that when an event began at the time t, the event's topic may refer to other subjects such as the names of people or places. Therefore, we combine all the term vectors of the stories in the cluster and to Increase the story can assigned in correct cluster.

4 Experiments and Discussions

We collected data from Chinese news web sites such as Apple Library, UNN, etc. and famous portals such as Yahoo!, etc. The data included news stories, discussions and reviews. We used Dataset of different sizes to experiment with our algorithm and for comparison with one another.

DataSet1: We collected 232 sports related stories with 100 topics from several Chinese news web sites on 2010-08-07, with the topics pre-clustered by users. The number of stories within a topic is between 1 and 15. DataSet2: We collected 523 stories from 2010-08-7 to 2010-08-9, containing 135 topics pre-clustered by users. The number of stories within a topic is between 1 and 40.

In our experiment, we compared our algorithm with TFIWF. However, the WF feature of TFIWF is updated dynamically, and we needed to calculate the initial WF. Therefore, we used data from almost 30 days to train the algorithm to implement TFIWF. We used the Recall, Precision and F-measure to analyze the story has been assigned to the correct Topic cluster. The metric is the traditional evaluation metric, which is widely used in information retrieval and clusters [14]. Recall and Precision usually contradict each other for information retrieval. Therefore, if we cluster all the stories in one cluster, Recall will become 100%. If we cluster every story as a topic, the Precision will become very high. As such, we used the F-measure to average the two value objectives to gauge the performance.

Table 1 Performance of the three algorithms.

	DataSet1 Recall	DataSet1 Precision	DataSet1 F-Measure	DataSet2 Recall	DataSet2 Precision	DataSet2 F-Measure
TF-Density	95.2%	95.2%	93.6%	88.3%	91.4%	87.2%
TF-IDF	95.2%	91.7%	91.3%	89.9%	85.7%	84.4%
TFIWF	95.2%	91.6%	88.7%	87.0%	87.1%	83.6%

Our proposed term weighting model, TF-Dentistry, combines the DF (Document Frequency) and WF (word frequency) features of IDF and IWF respectively. Therefore, our experiments compared its performance with those of the aforementioned two algorithms based on Dataset1 and Dataset2.

Table 1 shows the average Recall, Precision and F-measure for the three algorithms. One can see that the density for our proposed weighting model TF-Density in F-measure was better than those of IDF and IWF. As for Recall and Precision, the Precision of our algorithm was better than that of IDF. However, as we mentioned in the Evaluation Metric section, Recall and Precision usually contradict each other. Therefore, while we were able to tune both Recall and Precision values better than for the other algorithms, the performance in F-measure was not the best for our algorithm.

5 Conclusions

In this paper, we have created detection and tracking system for news topics. In this system, we proposed a new words measurement scheme named TF-Density, derived by modifying the well-known TF-IWF and TF-IDF algorithms to provide

a more precise and efficient method for recognizing important words in the text. Our experiment collected data from various Chinese news web sites and the results indicated that our proposed TF-Density algorithm performed better than TFIDF and TFIWF. The results showed our proposed topic detection and tracking system can provide a more precise and convenient result for users to track news.

References

1. http://udn.com/NEWS/main.html
2. http://news.google.com.tw/
3. http://news.google.com.tw/
4. http://www.nist.gov/speech/tests/tdt/index.htm
5. Yang, Y., Pierce, T., Carbonell, J.: A Study of Retrospective and On-line Event Detection. In: 21th ACM SIGIR Conference, Melbourne,Australia. ACM Press (1998)
6. Brants, T., Chen, F., Farahat, A.: A System for New Event Detection. In: SIGIR 2003, Toronto, Canada (2003)
7. Zheng, D., Li, F.: Hot topic detection on BBS using aging theory. In: Liu, W., Luo, X., Wang, F.L., Lei, J. (eds.) WISM 2009. LNCS, vol. 5854, pp. 129–138. Springer, Heidelberg (2009)
8. Wang, C., Zhang, M., Ma, S., Ru, L.: Automatic Online News Issue Construction in Web Environment WWW 2008, Beijing, China (2008)
9. Wang, C., Zhang, M., Ma, S., Ru, L.: Automatic Online News Topic Ranking Using Media Focus and User Attention Based on Aging Theory. In: CIKM 2008, NapaValley, California, USA (2008)
10. Wang, C., Zhang, M., Ma, S., Ru, L.: An Automatic Online News Topic Key - phrase Extraction System. In: 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Inteligent Agent Technology (2008)
11. Lee, S., Kim, H.: News Keyword Extraction for Topic Tracking. In: Fourth International Conference on Networked Computing and Advanced Information Management
12. Kuo, Z., Zi, L.J., Gang, W.: New Event Detection Based on Indexing-tree and Named Entity. In: SIGIR 2007 (2007)
13. http://ckipsvr.iis.sinica.edu.tw/
14. Salton, G., McGill, M.J.: Introduction to modern information retrieval. McGraw-Hill (1983)