A tracking and summarization system for online Chinese news topics

Hsien-Tsung Chang, Shu-Wei Liu and Nilamadhab Mishra
Department of Computer Science and Information Engineering, Chang Gung University, Taoyuan, Taiwan

Received 31 October 2014
Revised 21 September 2015
Accepted 23 September 2015

Abstract

Purpose – The purpose of this paper is to design and implement new tracking and summarization algorithms for Chinese news content. Based on the proposed methods and algorithms, the authors extract the important sentences that are contained in topic stories and list those sentences according to timestamp order to ensure ease of understanding and to visualize multiple news stories on a single screen.

Design/methodology/approach – This paper encompasses an investigational approach that implements a new Dynamic Centroid Summarization algorithm in addition to a Term Frequency (TF)-Density algorithm to empirically compute three target parameters, i.e., recall, precision, and *F*-measure.

Findings – The proposed TF-Density algorithm is implemented and compared with the well-known algorithms Term Frequency-Inverse Word Frequency (TF-IWF) and Term Frequency-Inverse Document Frequency (TF-IDF). Three test data sets are configured from Chinese news web sites for use during the investigation, and two important findings are obtained that help the authors provide more precision and efficiency when recognizing the important words in the text. First, the authors evaluate three topic tracking algorithms, i.e., TF-Density, TF-IDF, and TF-IWF, with the said target parameters and find that the recall, precision, and *F*-measure of the proposed TF-Density algorithm is better than those of the TF-IWF and TF-IDF algorithms. In the context of the second finding, the authors implement a blind test approach to obtain the results of topic summarizations and find that the proposed Dynamic Centroid Summarization process can more accurately select topic sentences than the LexRank process.

Research limitations/implications – The results show that the tracking and summarization algorithms for news topics can provide more precise and convenient results for users tracking the news. The analysis and implications are limited to Chinese news content from Chinese news web sites such as Apple Library, UDN, and well-known portals like Yahoo and Google.

Originality/value – The research provides an empirical analysis of Chinese news content through the proposed TF-Density and Dynamic Centroid Summarization algorithms. It focusses on improving the means of summarizing a set of news stories to appear for browsing on a single screen and carries implications for innovative word measurements in practice.

Keywords Chinese news summarization, News topic detection, TDT, TF-Density, TF-IDF, TF-IWF

Paper type Research paper

1. Introduction

People want to know what is going on around them on a daily basis. In the past, people would read the newspaper or watch television to obtain new information. When people are interested in specific news stories, they often wish to track a news topic by, for example, learning when the event started, finding information on the event as it progresses, and learning when the event ends. In the past, considerable time would be required for individuals to clip individual news items from the newspaper and paste them into a scrapbook if they wished to keep track of all the articles on a single topic.

Financial support furnished by the Ministry of Science and Technology, Republic of China, through Grant MOST 103-2221-E-182-053 and 104-2221-E-182-069 of Chang Gung University is gratefully acknowledged.

Aslib Journal of Information Management
Vol. 67 No. 6, 2015
pp. 687-699

2050-3806
DOI 10.1108/AJIM-10-2014-0147

More recently, due to the development of new technologies, people can engage in an increasingly wide range of activities on the internet, e.g., making new acquaintances, shopping, e-mailing, and reading news online via dedicated news sources. In a fast-moving consumer goods report on the behavior of netizens, the two most popular online activities are "reading the news" (65 percent) and "music appreciation" (65 percent). In fact, reading news online is so common among adults that it has become a necessary component of many people's daily routines.

Unfortunately, however, there is a glut of news sources on the internet, as shown by the many news stories listed in (Next-Media-Interactive-Limited, 2015; United-Daily-News-Group, 2015; Yahoo!Kimo, 2015). Visiting only one news source no longer suffices for those who desire different perspectives and are eager to participate in discussions on breaking or developing stories. For this reason, internet users spend an increasing amount of time visiting different news sources and performing internet searches on related topics, to the extent that certain web services such as Google News (Google, 2015) have taken the initiative to gather over 350 news sources and classify them into different categories, e.g., politics, sports, and art. Similar news stories 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 providing an umbrella term to describe the various news stories that Google has gathered from different news sources. Such services have vastly reduced the time that users spend searching for different news sources. The scale of different news topics, however, varies greatly. Some news topics may have more than 50 topically related news articles attached to them. Thus, when users wish to keep abreast of every viewpoint on or discussion of a certain news event, they need to spend increasing amounts of time reading multiple stories. In this paper, we describe a summarization technique that captures important information from each sentence to accurately represent a news story. We combine this information to create a new article for users to help them understand the overall situation clearly in less time.

Topic detection and tracking (TDT) tasks are commonly utilized to structure news stories from newswires and broadcasts on specific topics (The National Institute of Standards and Technology, 2011). When a story is received by the system, the system estimates whether the story is breaking news or old news. If the stream has never been seen before, it is likely to be breaking news.

In this age of information explosion, people regularly resort to search engines to determine "what is new" or "what is happening" in the world. With the explosion of information and documents on the internet, however, new tools are needed to better organize this information. In the corporate world, the general public is excited to be the first to know when certain companies release certain information on their products to help customers save on purchases or derive profit quickly.

In this paper, we focus on Chinese articles to build a detection and summarization system for news topics. We use TDT technology and propose a new-term weight algorithm to structure the news stream and maintain a basis in traditional summarization to render the algorithm more suitable for typical Chinese articles. Our system can help people understand topics and information more efficiently and conveniently.

The remainder of the paper is organized as follows. Section 1 provides an introduction on the background of TDT technology. Section 2 focusses on prior literature in the area. Section 3 presents an overview of the system. Section 4 describes a detection and summarization system algorithm. Section 5 provides experimental

results and a discussion, and Section 6 concludes. Before moving ahead, however, we will define our terminology as follows:

Story: a single news document.

Topic: a collection of stories that report the same news.

Corpus: all the stories collected for research.

$DF(w)$: the document frequency of a specific term w in the corpus.

$WF(w)$: the frequency of the term w in the corpus.

$TF(d,w)$: the frequency of term w, which is normalized according to story length in the story d.

WT: the total number of terms in the corpus.

$TermDensity(w)$: the term density of a specific term w in the corpus.

$Similarity_t(d,d')$: similarity value of two stories d and d' at time t.

$S_d(S_i)$: term weight value of sentence S_i in the story d.

$SNorm_d(S_i)$: normalized value of $S_d(S_i)$.

$P_d(Si)$: position value of sentence S_i in the story d.

$O_d(Si)$: overlap value of sentence S_i in the story d.

$Score_d(Si)$: total value of sentence S_i in the story d.

2. Related works

In this section, we examine the pros and cons of standard TDT techniques and summarization algorithms. We also link our research strategy to other, similar studies. Several TDT techniques have been developed over the years (Yang *et al.*, 1998), with the invention of new algorithms aimed at calculating the terms for clustering media streams. The TDT aims to extract useful terms by filtering out the noisy terms to render the overall process more accurate and efficient (Lee and Kim, 2008). The Term Frequency-Inverse Document Frequency (TF-IDF) (Yang *et al.*, 1998; Brants *et al.*, 2003; Zheng and Li, 2009) is a widely used algorithm in TDT and is defined as follows: If a term appears several times in a story, then it is considered important, and if that term appears in multiple sources, it is considered an important word in the story. TF-IDF specifically examines two features, document frequency (DF) and term frequency (TF), but in some situations, the importance of the term can be overestimated or underestimated.

To avoid this problem, a new algorithm has been proposed, namely, Term Frequency-Inverse Word Frequency (TF-IWF) (Wang *et al.*, 2008a, b). The TF-IWF is different from TF-IDF in that it uses word frequency (WF) instead of DF. This algorithm is more efficient and accurate in assigning term weights for structuring media streams.

Due to the vast amount of news produced every day, TDT techniques cannot satisfy newsreaders' needs. Readers require more convenient methods to understand entire news series in a short time. Summarization algorithms can group the important sentences in a body of stories and deliver the essential meaning of a topic to readers. Numerous summarization algorithms have been studied in which the graph-based summarization algorithms play a vital role. Güneş Erkan purposes a new algorithm called LexRank (Erkan and Radev, 2004), which uses a connectivity matrix that is based on sentence cosine similarity to compute a sentence's relative importance in order to extract the summarized contents. Chen *et al.* (2003) proposes a summarization algorithm that uses not only the named entities to identify the sentence similarities but also verbs and additional nouns to increase performance. Radev *et al.* (2000) proposes a summarization algorithm that attempts to utilize four factors (centroid value, positional value, first-sentence overlap, and a redundancy-based algorithm) to select important sentences from multiple documents. The use of a location feature to quantify the importance of the

sentence is not new to researchers in the field; however, the use of sentence clustering based on the location feature plays a vital role in text processing activities to summarize the associated terms, documents, and words to ensure that articles are easy to find and comprehend (Buitelaar *et al.*, 2005; Gupta and Lehal, 2010; Huang *et al.*, 2014).

The task of word or phrase extraction is an important research topic, especially for Chinese content. For this reason, we plan to use the Sinica Chinese extraction system (CKIP, 2015) to segment an article into words. The system uses an algorithm (Chen and Liu, 1992) to extract the words from the content and identify the unique terms (Chen and Bai, 1998). A language model (Chen and Ma, 2002) and unknown word extraction (Ma and Chen, 2003) are used to analyze the Chinese names or the translations of names from foreign languages. A preliminary version of this work was reported on in (Liu and Chang, 2013).

3. System architecture

Figure 1 shows the architecture of our proposed news tracking and summarization system, in which news stories are collected in Chinese by using an RSS crawler. RSS is a family of web feed formats used to publish frequently updated work (Wikipedia, 2015). We subscribe to the feeds from various Chinese news web sites and store the data on our crawler server; the news data are in XML format. We analyze the news data to extract useful data, including the titles and main bodies of the articles. In the third

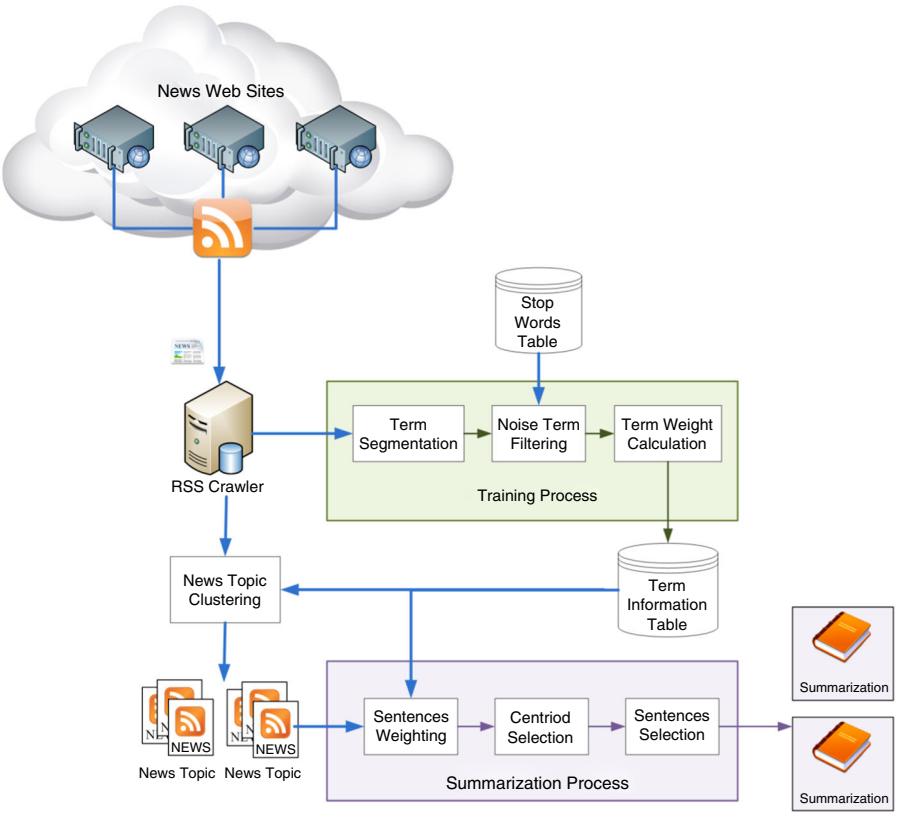

Figure 1.
The architecture of our proposed news tracking and summarization system

step, after extracting the titles and contents of the main body, we send the data to the Chinese Knowledge Information Processing Group (CKIP) system, which was constructed by Academia Sinica (CKIP, 2015), to break up the article into terms. We determine the word composition of the article and process the filter module that removes the noise words and stop words (such as is, are, you, I). Finally, a term list is created for each story. We create a term information table that includes term information for each story, such as TF, DF, and WT, all of which are useful for weighting the article. The table also includes the value of our proposed TF-Density algorithm to weight the term list for every story. We create a word vector for every story by using the term information table. We cluster the stories using cosine similarity according to the word vector of each story, and after running cosine similarity, we obtain many new topics. The generated news topics process summarization modules to obtain the results after weighting the sentences, centroid selection, and sentences selection steps. The details of the methods are further described in Section 4.

4. Methods

Below, we describe the methods used for our proposed web news tracking and summarization system.

4.1 TF-Density

In the past, TF-IDF and TF-IWF were proposed to evaluate the weight of terms in a document. TF-IDF successfully utilizes the term and document frequencies to calculate the normalized frequency in view of a corpus. TF-IDF, however, ignores the importance of WF. The TF-IWF method utilizes the WF instead of DF. Although WF can represent the partial influence of DF, it is still not as accurate as utilizing the DF and WF directly.

Here, we propose a new TF-Density algorithm to weight the terms in a news story. The algorithm combines features from the TF-IDF and TF-IWF algorithms because there are important features in these two algorithms, namely, DF and WF. We retain these features while providing a more precise and efficient method of recognizing the important words in the text. Our approach is to calculate the number of times that each term appears in all of the documents and divide it by the number of documents in which the term appears. In this way, we can obtain the density of the term w, i.e., the average number of times it appears in all of the documents. Thus, if a specific term w appears more times in a story than the average density, it is likely that the term in the specific story is more important than the others. Equations (1) and (2) are given as follows:

$$TermDensity(w) = \frac{WF(w)}{DF(w)} \tag{1}$$

$$TF{-}Density(d, w) = \frac{TF(d, w)/TermDensiy(w)}{WF(w)/WT} \tag{2}$$

where WF refers to the number of times term w appears in all of the documents and DF refers to the number of documents in which term w appears. Thus, we can calculate *TermDensity*(w), which refers to the average number of times term w appears in one document. *TF-Density*(d, w) denotes the term weight of term w in document d using the concept proposed above.

4.2 Topic tracking

We adopt the concept of cosine similarity to calculate the similarity between two stories d and d' at time t using the terms weight function tw_t (e.g. TF-IDF or TF-Density). The sum of the inner production of the two stories' term weight vectors will indicate their similarity. Equation (3) is given as follows:

$$Similarity_t(d, d') = \sum_{w \in d \cap d'} tw_t(d, w) \times tw_t(d', w) \quad (3)$$

When a news story d_n breaks at time t, it becomes a candidate for a new topic that will be compared with previous topics by means of pair-wise similarities. The similarity value of the story d_n and a topic c is calculated as the average $Similarity_t(d_n, d_c)$, where d_c is any story in topic c. If the similarity value is greater than the threshold ∂, the topic will be considered as having previously appeared before time t. If the value is less than the threshold, the candidate d_n will be deemed a new topic. In this way, we can ascertain whether the topic is old or new.

Every new incoming news story will be put into a new cluster as Cluster, which will contain just one story. We calculate the pair-wise similarities for all of the news topic clusters. We only use the top 30 percent of terms with higher weights in each story as their essential term vector for calculating pair-wise comparisons with one another because in the context of Chinese news (after CKIP), there are many noise words but few stop words such as modal words or adjectives. Therefore, to prevent unnecessary words from affecting the performance, we apply this method to filter those words before calculating the similarity value pair-wise for a new Cluster A, which is a new incoming news story for all existing clusters. Finally, we can identify the largest similarity value with this new Cluster A and a cluster named B in all of existing clusters. If the similarity value is larger than the threshold ∂, then we conclude a new cluster; thus, we combine A and B into a new cluster. If the value is smaller than the threshold, then Cluster A will be a new topic. The following algorithm-1 is the pseudo code for the news topic tracking algorithm.

News topic tracking algorithm (Algorithm-1):
C: $\{c_1, c_2, c_3, \ldots, c_n\}$, a set of news topic clusters.
d: Predefined threshold.
s: New incoming news story.
NewsTracking(s){
1: treat s as a new cluster c_{n+1}
2: if(C is not Null)
3: $V_i =$ Similarity (c_i, c_{n+1}) for $i = 1 \ldots n$
4: $V_t =$ MAX (V_1, V_2, … ,V_n)
5: if ($V_t > d$)
6: $c_t = c_t \cup c_{n+1}$
7: return
8: $C = C \cup c_{n+1}$
9: return

4.3 Summarization algorithm

We propose a new summarization algorithm that is based on Erkan and Radev (2004) and is named Dynamic Centroid Summarization. The algorithm depends on three values: sentence value; position value; and overlap value to determine which sentence will be chosen.

The first value is called the sentence value. After term weighting is performed, we know the weight of every term in the story and count the weights for the terms in each sentence, as shown in Equation (4). For example, in "Dangers are real, but deaths are increasingly rare for police officers," the sentence will be assigned a value of 30 (Dangers: 15, Deaths: 5, police officers: 10). Grammar and frequent terms will be ignored.

Typically, in English content, the lengths of different sentences are similar; however, sentences can be of various lengths in Chinese. Some sentences are long, and those lengthy sentences become a problem due to the higher probability of being selected because the total score from the terms is higher. We normalize the score to avoid this situation, as shown in Equation (5), where $S_d(S_i)$ indicates the sentence value of sentence Si in the news story d and $|Si|$ indicates the total number of terms that the sentence contains:

$$S_d(S_i) = \sum_{w \in S_i} TF{-}Density(w, d) \tag{4}$$

$$SNorm_d(S_i) = \frac{S_d(S_i)}{|S_i|} \tag{5}$$

A sentence's position value in an article is an important factor for sentence selection. The first sentence of an English-language news story is typically the most important one. This is the lettering style of English-language news; however, this lettering style does not fully apply to other languages. As shown in Figure 2, especially in Chinese, the most important sentence appears in the middle of the story. The method used in (Erkan and Radev, 2004) describes how existing algorithms always set the highest score for the first sentence, which is not suitable for non-English writing. We propose a new method called Dynamic Centroid Summarization to select the most important sentences in a news story with higher accuracy. We obtain the value of each sentence in a story after the sentence value calculation. Then, we obtain the SMAX of the sentence that has the maximum

Notes: The Max Score reflects the sentence with maximum sentence score $SNorm_d(S_i)$. The news articles are written in Chinese

Figure 2.
The concept of the proposed Dynamic Centroid Summarization

score for its sentence value. The result is the selected dynamic centroid of the story. Figure 2 shows that if a sentence is positioned far from the centroid position, the position value will decrease. We use Equation (6) to calculate the position value for each sentence. The value of n indicates how many sentences are in the story and x indicates the sentence position with the maximum score in the story. $|x\text{-}i|$ indicates the number of sentences between the maximum scored sentence S_x and the current sentence S_i.

Overlap value is a function that indicates if a sentence S_i is similar to another S_j in a story that is reported latter. The overlap value will be less if a similar sentence exists. This value can help us to select different sentences that have different viewpoints or themes. Therefore, if the composition of a pair or group of sentences is quite similar, we must have a penalty mechanism to discard them, shown in Equation (7). The time (S_i) function returns the reported time of S_i:

$$P_d(S_i) = \frac{n-|x-i|}{n} \times S_{MAX} \tag{6}$$

$$O_d(S_i) = SNorm_d(S_i) \times \left(1 - \underset{Time(S_i) < Time(S_j)}{MAX} \left(\frac{|S_i \cap S_j|}{S_i \cup S_j}\right)\right) \tag{7}$$

$$Score_d(S_i) = SNorm_d(S_i) + P_d(S_i) + O_d(S_i) \tag{8}$$

Finally, the score for sentence *Si* is assigned according to Equation (8), which states that if Score (S_i) is higher, the sentence is more important with respect to the topic. We choose the sentences with the highest y percent Score (S_i) in each news story to compose the summarization and the name y percent as compression rate. We adjust the compression rate to change the summarization length; in this paper, we set the compression rate to 20 percent. The following algorithm-2 is the pseudo code for the news topic summarization algorithm.

News topic summarization algorithm (Algorithm-2):

C: $\{d_1, d_2, d_3, \ldots, d_n\}$ a documents set of one topic cluster

d_i: $\{s_{i1}, s_{i2}, s_{i3}, \ldots, s_{im}\}$ a sentences set of one document

TopicSummarization(C)

1: for i = 1…n, $SMAX_i = 0$
2: for j = 1…$|d_i|$
3: $VS1_{ij} = SNormd_i(s_{ij})$
4: if($VS1_{ij} > SMAX_i$)
5: $SMAX_i = VS1_{ij}$, Pos i = j;
6: for i = 1…n
7: for j = 1…$|d_i|$
8: $VS2_{ij} = ((|d_i|\text{-}(|Pos\ i\text{–}j|)/|d_i|) \times SMAX_i$
9: $k = SMAX_i + (j/2)((j\%2) \times 2\text{-}1)$
10: $VS3_{ik} = O(sik)$ // based on equation (8)
11: for i = 1…n
12: for j = 1…$|d_i|$
13: $VS_{ij} = VS1_{ij} + VS2_{ij} + VS3_{ij}$
14: Select top 20% sentences in di according to VS_{ij}

4.4 Evaluation methods

The performance evaluation of topic tracking and summarization seeks to measure the effectiveness within different algorithms. In the performance evaluation of topic tracking, we can first manually tag the stories with different topics to be the correct results. We can also utilize the common evaluation methods, with recall defined in Equation (9), precision in Equation (10), and *F*-measure as Equation (11), to evaluate the algorithm's performance:

$$Recall = \frac{|\{Algorith\ mresults\} \cap \{Correct\ results\}|}{|\{Correct\ results\}|} \tag{9}$$

$$Precision = \frac{|\{Algorithm\ results\} \cap \{Correct\ results\}|}{|\{Algorithm\ results\}|} \tag{10}$$

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{11}$$

We know that the quality of a summarization result depends on the individuals making the assessment, and it is difficult to set correct results for the evaluation of topic summarization. Therefore, we randomly select several topics and change the position of the results from two algorithms to be compared within the topic area when we present the material to people for assessment. People selected the result on a new topic that they found easier to understand. If the user tags one algorithm as better, the algorithm will gain 1 point. If the user thinks the results from the two algorithms are similar, both algorithms will gain 0.5 points. The algorithms' scores will be tallied after several rounds to indicate which algorithm performs better. We call this evaluation method a blind test.

5. Experiments

5.1 Experimental data set

We collect data from several Chinese news web sites, namely, Apple Library, UDN, and Yahoo!Kimo News. Table I provides basic information regarding the test sets. We use three data sets of different sizes to experiment with our algorithm and compare different algorithms to one another. We first manually clustered and tagged news stories into several topics for DataSet1 and DataSet2. The clusters of DataSet3 are based on Google News results.

DataSet1: news collected from three Chinese news web sites on August 7, 2010, with the topics pre-clustered by the users.

DataSet2: a total of 523 stories collected from August 7, 2010 to August 9, 2010, containing 135 topics that were pre-clustered by users.

DataSet3: a total of 354 stories collected from Google News (Google, 2015) containing 37 topics.

	DataSet1	DataSet2	DataSet3
Number of stories	232	523	354
Number of topics	100	40	37
Average number of sentences in a story	10.3	9.8	9.3
Max number of stories in a topic	15	40	22
Min number of stories in a topic	1	1	1

Table I. Basic information regarding the test data sets

5.2 Results of topic tracking

We first evaluate the performance of news story classification for tracking. We used the recall, precision, and *F*-measure to analyze whether stories were assigned to the correct topic cluster. The metrics are the traditional evaluation metrics that are widely used in information retrieval and clusters (Salton and McGill, 1986).

Our proposed term weighting model, TF-Density, combines the DF and WF features of IDF and IWF, respectively. Therefore, in our experiments, we compared its performance with the performance of the aforementioned two algorithms based on DataSet1, DataSet2 and DataSet3.

Table II shows the average recall, precision, and *F*-measure for the three algorithms. It is clear that the density for our proposed weighting model TF-Density performed better in the *F*-measure than TF-IDF and TF-IWF in DataSet1 and DataSet2. As for recall and precision, the precision of our algorithm was better than that of the other algorithms, although all showed similar recall. As we mentioned in the Evaluation Metric section, however, recall and precision usually work against each other. Thus, the *F*-measure result is more significant than that of recall and precision. The evaluation of Data set3 is based on the results from Google News, and the proposed TF-Density algorithm once again showed superior performance.

5.3 Results of topic summarization

Table III shows the result of the summarization system using the blind test. We randomly selected five news topics from our proposed news topic detection system and set the compression rate to 20 percent. For example, if there are 30 sentences regarding a news topic, then six sentences will output from our summarization system. We used the blind test described above to evaluate our summarization system compared to LexRank (Erkan and Radev, 2004). In total, 43 computer science and information engineering

Table II.
Evaluation of topic tracking for TF-Density, TF-IDF, TF-IWF with recall, precision, and the *F*-measure

	TF-Density (%)	TF-IDF (%)	TF-IWF (%)
DataSet1-recall	95.2	95.9	90.9
DataSet1-precision	95.2	91.7	91.6
DataSet1-*F*-measure	95.2	93.8	91.3
DataSet2-recall	88.3	89.9	87.0
DataSet2-precision	91.4	85.7	87.1
DataSet2-*F*-measure	89.8	87.8	87.1
DataSet3-recall	74.9	72.9	70.1
DataSet3-precision	86.7	76.1	76.9
DataSet3-*F*-measure	80.4	74.5	73.3

Table III.
Blind test results of the summarization algorithm

	Dynamic centroid	LexRank (Erkan and Radev, 2004)
Topic 1	30	12
Topic 2	29.5	12.5
Topic 3	26.5	15.5
Topic 4	32.5	9.5
Topic 5	24.5	17.5
Average	28.6	13.4

students in the "Web Service System" class completed the blind test. The students were aged from 20 to 21 years old. Before the blind test, the students were introduced to the blind test method, and the steps were demonstrated. The results show that our proposed Dynamic Centroid Summarization algorithm can more accurately select the sentences that are more representative of the news topic, as shown in Table III.

6. Conclusions

This paper describes a new tracking and summarization system for web news topics. In this system, we proposed a new measurement scheme for words, TF-Density, and a new summarization algorithm. This system can track the news according to different topics and deliver summarizations of the topic to help readers understand topics more quickly. It also provides the information regarding cause and effect in a news topic to readers.

TF-Density is derived by modifying the well-known TF-IWF and TF-IDF algorithms. Our experimental results indicated that our proposed TF-Density algorithm performed better than TF-IDF and TF-IWF. The precise evaluation of terms within a document will also improve the precision of relative document mining algorithms.

Our new summarization algorithm is called Dynamic Centroid Summarization. The algorithm is based on Erkan and Radev (2004) and is more suitable for Chinese articles. The results show that our tracking and summarization system for news topics can provide a more precise and convenient result for users to track Chinese-language news articles.

References

Brants, T., Chen, F. and Farahat, A. (2003), "A system for new event detection", *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM*, pp. 330-337.

Buitelaar, P., Cimiano, P. and Magnini, B. (2005), *Ontology Learning from Text: Methods, Evaluation and Applications*, IOS Press, Amsterdam, Dutch.

Chen, H.H., Kuo, J.J., Huang, S.J., Lin, C.J. and Wung, H.C. (2003), "A summarization system for Chinese news from multiple sources", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 13, pp. 1224-1236.

Chen, K.-J. and Bai, M.-H. (1998), "Unknown word detection for Chinese by a corpus-based learning method", *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 3 No. 1, pp. 27-44.

Chen, K.-J. and Liu, S.-H. (1992), "Word identification for Mandarin Chinese sentences", *Proceedings of the 14th Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics*, pp. 101-107.

Chen, K.-J. and Ma, W.-Y. (2002), "Unknown word extraction for Chinese documents", *Proceedings of the 19th International Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics*, pp. 1-7.

CKIP, S. (2015), "National digital archives program Taiwan", available at: http://ckipsvr.iis.sinica.edu.tw/ (accessed September 15, 2015).

Erkan, G. and Radev, D.R. (2004), "LexRank: graph-based lexical centrality as salience in text summarization", *Journal of Artificial Intelligence Research*, Vol. 22 No. 1, pp. 457-479.

Google (2015), "Google news", available at: http://news.google.com.tw/ (accessed September 21, 2015).

Gupta, V. and Lehal, G.S. (2010), "A survey of text summarization extractive techniques", *Journal of Emerging Technologies in Web Intelligence*, Vol. 2 No. 3, pp. 258-268.

Huang, X., Wan, X. and Xiao, J. (2014), "Comparative news summarization using concept-based optimization", *Knowledge and information systems*, Vol. 38 No. 3, pp. 691-716.

Lee, S. and Kim, H.-J. (2008), "News keyword extraction for topic tracking", *Fourth International Conference on Networked Computing and Advanced Information Management, 2008. NCM'08, IEEE*, pp. 554-559.

Liu, S.-W. and Chang, H.-T. (2013), "A topic detection and tracking system with TF-Density", in Gaol, F.L. (Ed.), *Recent Progress in Data Engineering and Internet Technology*, Springer, Berlin, pp. 115-120.

Ma, W.-Y. and Chen, K.-J. (2003), "A bottom-up merging algorithm for Chinese unknown word extraction", *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing-Volume 17, Association for Computational Linguistics*, pp. 31-38.

(The) National Institute of Standards and Technology, I.T.L (2011), "Topic detection and tracking (TDT)", available at: www.nist.gov/speech/tests/tdt/index.htm (accessed September 21, 2015).

Next-Media-Interactive-Limited (2015), "Apple daily news", available at: http://tw.nextmedia.com/ (accessed September 21, 2015).

Radev, D.R., Jing, H. and Budzikowska, M. (2000), "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies", *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization, Association for Computational Linguistics*, pp. 21-30.

Salton, G. and McGill, M.J. (1986), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.

United-Daily-News-Group (2015), "United Daily News site", available at: http://udn.com/NEWS/main.html (accessed September 21, 2015).

Wang, C., Zhang, M., Ma, S. and Ru, L. (2008a), "Automatic online news issue construction in web environment", *Proceedings of the 17th International Conference on World Wide Web, ACM*, pp. 457-466.

Wang, C., Zhang, M., Ru, L. and Ma, S. (2008b), "Automatic online news topic ranking using media focus and user attention based on aging theory", *Proceedings of the 17th ACM Conference on Information and Knowledge Management, ACM*, pp. 1033-1042.

Wikipedia (2015), "RSS", available at: http://en.wikipedia.org/wiki/RSS (accessed September 21, 2015).

Yahoo!Kimo (2015), "Yahoo!Kimo News Taiwan", available at: http://tw.news.yahoo.com/ (accessed September 21, 2015).

Yang, Y., Pierce, T. and Carbonell, J. (1998), "A study of retrospective and on-line event detection", *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM*, pp. 28-36.

Zheng, D. and Li, F. (2009), "Hot topic detection on BBS using aging theory", in Liu, W., Luo, X., Wang, F.L. and Lei, J. (Eds), *Web Information Systems and Mining*, Springer, Berlin, pp. 129-138.

About the authors

Hsien-Tsung Chang got his MS and PhD Degrees in the Department of Computer Science and Information (CSIE) from the National Chung Cheng University in July 2000 and July 2007, respectively. He joined the faculty of Computer Science and Information Engineering Department at Chang Gung University in August, 2007. He is also a Member of the High-Speed Intelligent Communication Center at Chang Gung University. Professor Chang's research areas focus on

search engine, data engineering, information retrieval, web services, web 2.0, social network, and cloud computing. Professor Chang is the Director of the Web Information & Data Engineering Laboratory (WIDELab). Hsien-Tsung Chang is the corresponding author and can be contacted at: smallpig@widelab.org

Shu-Wei Liu got his MS Degrees in Computer Science and Information Engineering (CSIE) from Chang Gung University in June 2011. He is now an Engineer at Quanta Inc. in Taiwan.

Nilamadhab Mishra got his MCA, MPHIL, and MTech Degrees in Computer Science and Engineering from Indian Universities. He is now a PhD scholar in WIDELab under the department of Computer Science and Information Engineering, Chang Gung University, Taiwan. He focusses his research on network centric data science and analytics, IoT Big-data system, and cognitive learning apps design and exploration.