News Filtering and Summarization System Architecture for Recognition and Summarization of News Pages

Bamber and Micah Jason

Abstract--- *Due to the swift development of text documents, document clustering has turned out to be one of the foremost techniques for accurately organizing large quantity of documents into a small number of significant clusters. On the other hand, there still exist quite a lot of complications for document clustering, like high dimensionality, accuracy, meaningful cluster labels, scalability, overlapping clusters, and extracting semantics from texts.*

Here, semantic relations between phrases are analyzed and lexical chai is used to characterize semantic relation. Key phrases are subsequently extracted and a semantic link graph is built on the lexical chai. This paper presents the recognition and summarization components of the news summarization (NS) system. In order to test this system, Web news pages with core hints (which are the subject keywords presented by the news authors) are selected from the 163 website (www.163.com). Experimental results show that this method can correctly recognize Web news pages with a rate of better than 96 percent. They also show that the keyword-extraction method considerably outperforms methods based on term frequency and lexical chai.

Experiments also conducted on News datasets to evaluate the performance. Results proved that this scheme completely outperforms the influential news document clustering methods with better accuracy. As a result, this approach not only provides more general and meaningful labels for documents, however also efficiently produces overlapping news story clusters.

Keywords--- *Data Mining, Web Mining, Clustering.*

I. INTRODUCTION

RETRIEVING useful Web news involves both filtering and keyword extraction. Because the layouts and styles of Web news pages differ from other webpages, it is especially important to accurately identify Web news for correct filtering.1–3 to do this, we propose an automatic recognition method that uses classification rules for Web news based on a combination of URL, structure, and content attributes. After the automatic recognition and filtering, our system uses a new key-phrase extraction method from Web news content based on semantic relation.

With this approach, we analyze semantic relation between phrases and use lexical chai to represent semantic relation. Key phrases are then extracted and a semantic link graph is built on the lexical chai. This article presents the recognition and summarization components of our news filtering and summarization (NS) system. To test our system, we selected Web news pages with core hints (which are the subject keywords provided by the news writers) from the 163 website (www.163.com) to evaluate our key-phrase extraction method. Our system's experimental results demonstrate that our method can accurately recognize Web news pages with a rate of better than 96 percent. They also show that the keyword-extraction method we propose substantially outperforms methods based on term frequency and lexical chai.

II. SYSTEM ARCHITECTURE

Figure 1 shows our NS system architecture. A user or an application may submit a URL to the NS system as the initial input. The results returned by NS are key phrases and their lexical chai. The NS system consists of two main phases. Phase 1 involves Web news recognition and filtering. This module takes a URL from an end user or an application and then performs an automatic recognition for Web news. It also prepares input data for the summarization phase.

The recognition stage includes two parts: training and recognition. First, we randomly select some news web pages for training; these web pages are represented by vectors of important features such as URL, structure, and content attributes. A classification technique is utilized with those features for the automatic recognition of Web news. Second, according to the given URL, the NS system fetches the webpage content and generates a feature vector by the same features from the training process. Finally, we use the Web news identifier of the classification model obtained from the training process to recognize Web news.

If the URL is considered as a non-news webpage, the system will not do anything further. Otherwise, it performs filtering. We complete the filtering stage using the Web information extractor, which retrieves the news webpage's title and news content by using preconfigured extraction rules. As with W4F4 and XWap, 5 the NS system also adopts extraction rules based on paths of the new webpage's document object model (DOM) tree. The Web information extractor uses extraction rules while it traverses that DOM tree. Because the layouts and styles of Web news pages differ from other web pages, such as a site's homepage, the extraction rules are invalid if the webpage is not a news page.

Bamber, Clausthal University of Technology, Germany.
Micah Jason, University of Toronto, Canada.
DOI:10.9756/BIJDM.8339

For this reason, this phase primarily focuses on accurately identifying a news webpage. Phase 2 involves Web news summarization. During phase 1, we filtered non-Web news and non-news content on a news webpage. The next task is to summarize and extract the key phrases that capture the news webpage's main topic. We first segment the filtered document with a title and a body into words and then remove the stop words.

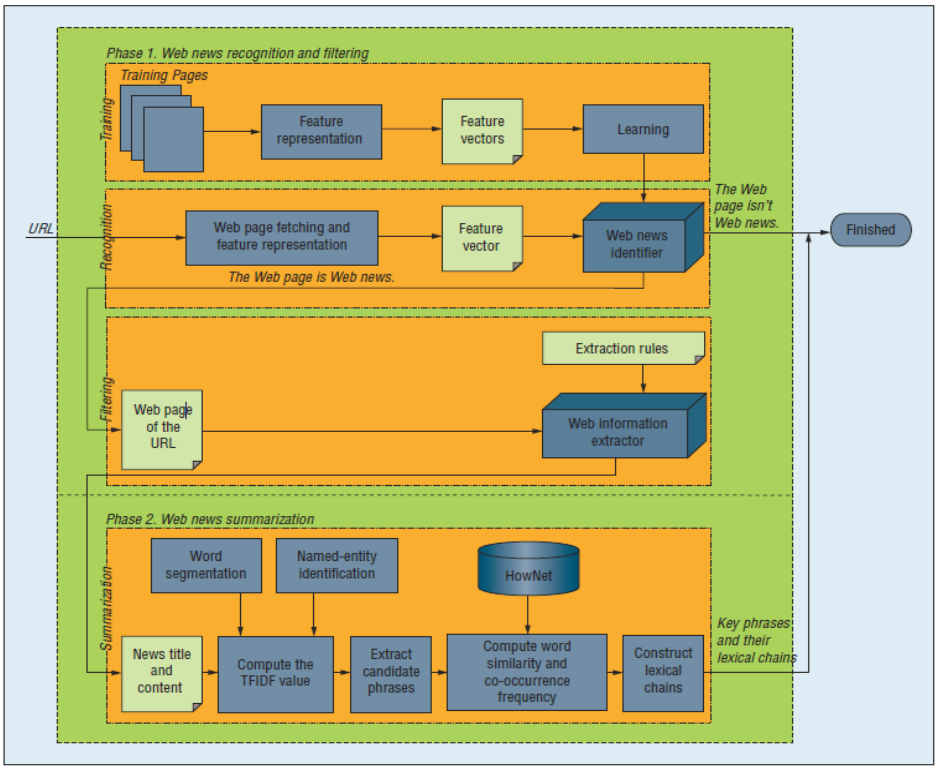

Figure 1: The News Filtering and Summarization (Ns) System Architecture

Word frequencies are counted and the TFIDF (term frequency–inverse document frequency) values are computed according to the corpus.6 Candidate phrases are identified by the TFIDF values. For the candidate phrases that occur in How Net, we compute word similarities based on HowNet.7,8 We also compute word co-occurrence frequencies and construct lexical chai with word similarities and word co-occurrence frequencies. Then key phrases are extracted from the candidate phrases according to the TFIDF values and the semantic information in the lexical chai.

A. *Interrelated Attributes in Web News Recognition*

Our method for automatic recognition of Web news pages is based on the selection of important attributes and then the utilization of a classification algorithm to identify Web news. Existing classification algorithms are well developed, and the key to accurately recognizing a news webpage is knowing which attributes represent the features of the given news webpage. We can automatically recognize Web news by applying a decision-tree learning technique—the C4.5 algorithm. If the webpage is classified as news, we can perform filtering and summarization of the Web news.

B. *URL Attributes*

The URL of a news Web page has the following characteristics. First, in most cases, the URL contains time related attributes. Second, news websites often use the same structure for their URLs, so we can use the URL attributes as a kind of important identification features for Web news. URL attributes can be both positive and negative. Positive attributes can help us recognize Web news, while negative attributes cannot. As such, positive attributes include time attributes, second-level domain attributes, and first-level catalog attributes:

- Time attributes. We randomly selected 478 URL addresses of Web news through Baidu news search (http://news.baidu.com/), and the test results demotrate that 97.4 percent of these Web news pages contain time attributes—for example, /2007-08-06/, /20070806/, and /2007-08/06/. This indicates that the time attributes are crucial to the recognition of Web news.
- Second-level domain attributes. We discovered that the same departments of different webpages share similar structure attributes. For example, in addition to the time attributes, the URLs of sports and entertainment Web news of the 163 website also contain second level domain attributes such as "sports" and "end."

First-level catalog attributes. We discovered that only 2.6 percent of the 478 Web news pages we randomly selected had URLs that did not contain time attributes. There are also several first-level catalog attributes of Web news such as "news," "news html," and "news center." When the webpages do not contain time attributes, the first-level catalog attributes provide an important basis for the recognition of Web news.

We selected Web news URL addresses from 10 popular news websites in China, including 163, Sohu, Sina, Tom, QQ, CCTV, the China news net, the Chinese news net, the New China net, and the Chinese youth, to create a training corpus, and our training obtained 59 attributes of the second-level domains. On the other hand, we can use negative attributes to eliminate webpages that are not Web news. For example, URLs that end with the term index and a "/" usually indicate that a webpage is an index page, not a news Web page. "Blog," "blogs," or "video" in the second-level domain can also indicate that the webpage is not news.

C. *Structure Attributes*

The structure and content of a news webpage affect classifier performance in addition to the URL attributes. The Web news pages contain rich structure information that can enhance the accuracy of recognition, if used correctly. With this starting point, we found that some structure attributes contribute to page recognition, including a webpage's title and subtitle, labeled with <title> and <Hn> tags, and tags that carve up a webpage's hierarchy.

By traveling the DOM tree, we can pick up all the <title> tags in an HTML file. The content picked up includes the start bit, the end bit, and the tag's content. By analysing the distilled content, we found that the title tags are similar in format—for example, "Scientist made time machine theoretic model shuttle space time may come true – 163 news center." The content extraction includes the webpage's title and the website information, which is spaced out by the conjunction and can be an important attribute of the news webpage.

The text's title is usually labeled by an <Hn>. By traveling the DOM tree, we can assert that if the <Hn> exists it might be a classification attribute in the HTML content. Moreover, we have also found the hierarchy labeled by between the text and the title of Web news, including the time attributes. The tag is an important attribute for classifying Web news, just like year, month, date, hour, and minute.

D. *Content Attributes*

A webpage classification sets an HTML file as an object. We get relevant HTML content from the URL address and pick up some keywords as characters. These keywords should affect the accuracy of recognition of the news webpage. (To help readers, we translated the content attributes of Chinese Web news into English phrases.) After reviewing 881 HTML files of non-Web news and 1,087 HTML files of Web news, which were randomly selected from our 10 news websites, we found that the frequency of the keyword news in a webpage is an important attribute for the recognition of a news webpage. We observed that, whether the keyword news appears twice can discriminate a news webpage from a non-news webpage with an accuracy of 82.56 percent.

Therefore, the appearing frequency of the keyword news must be greater than or equal to two as another attribute of a news webpage. We also selected the following keywords as content attributes of a news webpage: news center, report, reporter, editor, and relative news. Table 1 lists the Web news attributes we extracted from the popular news websites of 163, Sohu, Sina, Tom, and QQ.

III. Experimental Results and Analysis

We used our randomly selected 1,087 news Web pages as our experimental objects. We then used C4.5 to perform inductive data mining on them. We evaluated the performance of the induced C4.5 classifier in terms of precision. Depending on the attributes from the different websites, we conducted three groups of experiments accordingly. For the purposes of reporting our experimental results, we denoted the sites accordingly: QQ (S1), Sina (S2), Sohu (S3), 163 (S4), Tom (S5), New China net (S6), Chinese news net (S7), CCTV (S8), Chinese net news (S9), and Chinese youth (S10).

For the first group of experiments, we extracted the attributes of Web news from five websites and then conducted training and testing with the same websites. Table 2 shows the results of these experiments. For the second group of experiments, we extracted the attributes of Web news from five different websites that were not used for training and testing. Table 3 shows the results of these experiments. For the third group of experiments, we extracted the combined attributes of Web news from five websites and evaluated the results using five different websites for training and testing.

```
Time attributes = 0
|  First level catalog attributes = 0: 0
|  First level catalog attributes = 1
|  |  the end of URL with "/" = 0: 1
|  |  the end of URL with "/" = 1: 0
Time attributes = 1
|  bbs = 0
|  |  sum up the number of times of "news" appears in the HTML page>2 = 0:0
|  |  sum up the number of times of "news" appears in the HTML page>2 = 1
|  |  |  includes news feature in <div> tag or not = 0
|  |  |  |  related news = 0
|  |  |  |  |  comment = 0
|  |  |  |  |  |  second level domain attributes = 0: 1
|  |  |  |  |  |  second level domain attributes = 1: 0
|  |  |  |  |  comment = 1: 1
|  |  |  |  related news = 1: 1
|  |  |  includes news feature in <div> tag or not = 1: 1
|  bbs = 1: 0
```

Figure 2: Web Page Content Detected

In the second group of experiments, we selected attributes from the 163, Sohu, Sina, Tom, and QQ websites and used the other websites for training and testing. Figure 2 shows the decision tree from the Tom website for feature selection, and the 163, Sohu, Sina, Tom, QQ, CCTV, China news net, Chinese news net, New China net, and Chinese youth websites for testing. We can convert this decision tree into a set of production rules for the automatic recognition of Web news.

For example, time attribute = 0 ∧ first-level catalog attribute = 1 ∧ a URL ending in "/" = 1 → non-news Web page. In contrast, time attribute = 1 ∧ bbs = 0 ∧ sum up the number of times of "news" appears in the HTML page = 1 ∧ includes news feature in tag or not = 1 → news Web page. From these experimental results, we can see that from a comprehensive usage of Web news URL, structure, and content features in the selection of attributes we can get an accurate C4.5 classifier when identifying Web news. The experimental results indicate that the accuracy of our proposed automatic recognition method is more than 96 percent.

We have also built recognition classifiers with naïve Bayes and C4.5 on the same features. As Figure 3 shows, the experimental results demonstrate that the two classifiers built by naïve Bayes and C4.5 both have a high precision for recognizing Web news, while the precision of decision trees by C4.5 is better than that of naïve Bayes. The main contribution of this recognition component is the establishment of the features in Table 1 for Chinese news webpages, based on our empirical studies from 10 well-known Chinese news websites.

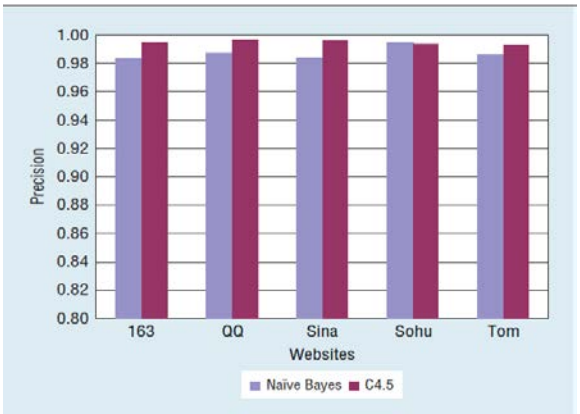

Figure 3: Classification Precision for C4.5 Classifier Algorithm

This means we can actually use any existing classification method to perform recognition with these attributes. Based on our extensive experiments, the URL attributes and structure attributes of news Web pages have most frequently occurred in the constructed decision trees. Time attributes, second-level domain, first-level catalog, and URLs ending with the term index and "/" are the most predictive attributes. The keyword news in the <title> tag and a time feature in the tag generally contribute to a positive classification. In addition, the keywords "main body" and "report" as well as the number of times of the term news occurs on an HTML page are comparatively important for content attributes of news webpages.

A. Key-Phrase Extraction Based on Semantic Relation

Key phrases are mainly the nouns in academic journals.9 However; verbs also play a key role in representing news topics. Therefore, our NS system considers all phrases except stop words in the Web news pages. Based on the word similarity and lexical chai, our Key-Phrase Extraction based on Semantic Relations (KESR) algorithm is designed as follows.

1. Filter non-news content in a news webpage. Segment the words, and remove the stop words.
2. Compute the TFIDF of each word $\omega_i$by

$$TFIDF_i = \frac{tf_i \times \log(N/n_i)}{\sqrt{\sum_j (f_i \times \log(N/n_i))^2}}$$

Where tf_i is the frequency of word ω_i in the given webpage, N is the number of the documents in the corpus, and ni is the number of documents in the corpus that contain wordω_i.

3. Select the top n words {ω_1, ω_2, … , ω_n} from all the words segmented, except stop words by TFIDF as candidate words.
4. Compute the word similarities based on How Net and word co-occurrence frequencies of candidate words.
5. Select the first candidate word w1 to construct the first lexical chain L1, based on lexical cohesion among sequences of related words.10
6. Select a candidate word ω_i. If the word similarity between ω_i and some word in the lexical chain L_j exceeds the threshold t1 or the word co-occurrence frequency exceeds the threshold t2, then ω_i is filtered into L_j, else use ω_i to construct a new lexical chain. The t1 and t2 are predefined parameters by the user.
7. Repeat step 6 until all the candidate words are computed.
8. Compute the weight of each phrase ω_i by

$$\text{Weight}(w_i) = a \times TFIDF_i + b \times \left|\text{chain}_i\right|$$

Where TFIDFi is the TFIDF value of ω_i, |chaini| is the length of the chain that contai wi and a and b are two parameters that can be adjusted.

9. Select the top m words as the key phrases extracted from the candidate words by their weights.

Because there are no standard news webpages for key-phrase extraction, we selected 120 Web news pages with core hints from the 163 website (www.news.163.com) as our experimental data to test KESR. We compared the key phrases extracted with the phrases in the news title and the phrases in the core hints provided by the editor. We used recall and precision as measures of extraction performance. The title recall R and core hint precision P are defined as follows:

$$R = \frac{\text{no. of key phrases extracted matching title phrases}}{\text{no. of title phrases}}$$

$$P = \frac{\text{no. of key phrases extracted matching core hint phrases}}{\text{no. of key phrases extracted}}$$

We compared the KESR experimental results with TFIDF (a key phrase extraction implementation based on TFIDF) and KELC (a key phrase extraction method based on lexical chai)11 by conducting two sets of experiments. In the first set, we *removed* the news title and the core hints for each news webpage and then compared the title recall and the core hint precision of different methods.

In the second set of experiments, we kept the news title while removing the core hints and then compared the core hint precision of the different methods. The parameters of n, a, and b selected are respectively 20, 1, and 1 by experiments. According to our experiments, n should be between 20 and 50; if it is smaller than 20, the advantages of semantic relation would be evident, and if it is greater than 50, the importance of word frequency to the attracted key phrases would be

reduced. The values of a and b have followed the choices in previous research.

Especially when the number of key phrases extracted is 3, the recall and the precision improve 20.93 and 26.97 percent, respectively. The semantic relation of phrases is considered in KESR on the basis of term frequency. The aim of KESR is to extract those words that have a low frequency but provide a great contribution to the text subject, and to filter out those words that have a high frequency but do not contribute to the text subject.

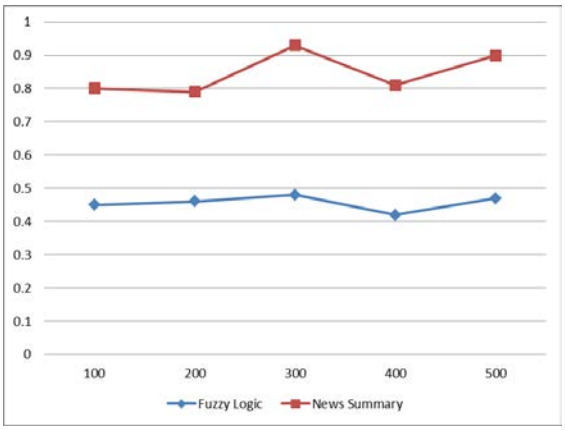

Figure 4: Accuracy of Classification of the NS against the Fuzzy Logic Clustering

IV. CONCLUSION

Our NS system differs from existing commercial news systems such as Google News in significant ways. As Google's automated news aggregator, Google News does not provide filtering and summarization function. The NS system recognizes Web news pages automatically, filters the non-news content by using preconfigured extraction rules, and summarizes the news in lexical chai.

Because Chinese characters differ from English words, one of the NS system's key challenges is to deal with news written in Chinese. To address this, the recognition component selects elaborate features to build a classifier for identifying Web news, especially for Chinese news Web news pages, while traditional webpage classification focuses on classifying general webpages. The carefully selected attributes, such as time attributes, second-level domain attributes, and first-level catalog attributes about URL attributes are the original attributes for recognizing Web news pages.

REFERENCES

[1] J.Y. Jiang, R.J. Liou and S.J. Lee, "A fuzzy self-constructing feature clustering algorithm for text classification", IEEE transactions on knowledge and data engineering, Vol. 23, No. 3, Pp. 335-349, 2011.

[2] D.L. McGuinness, P. Fox, B. Brodaric and E. Kendall, "The Emerging Field of Semantic Scientific Knowledge Integration", IEEE Intelligent Systems, Vol. 24, No. 1, Pp. 25–26, 2009.

[3] Z. Gong and Q. Liu, "Improving Keyword Based Web Image Search with Visual Feature Distribution and Term Expansion", Knowledge and Information Systems, Vol. 21, No. 1, Pp. 113–132, 2009.

[4] O'H. Kieron and S. David, "The Devil's Long Tail: Religious Moderation and Extremism on the Web", IEEE Intelligent Systems, Vol. 24, No. 6, Pp. 37–43, 2009.

[5] A. Saiiuguet and F. Azavant, "Building Intelligent Web Application Using Lightweight Wrappers", Data and Knowledge Eng., Vol. 36, No. 3, Pp. 283–316, 2001.

[6] L. Liu, C. Pu and W. Han, "XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources", Proc. 16th IEEE Int'l Conf. Data Eng (ICDE), 2000.

[7] G. Salton, A. Wong and C.S. Yang, "On the Specification of Term Values in Automatic Indexing", J. Documentation, Vol. 29, No. 4, Pp. 351–372, 1973.

[8] Z.D. Dong and Q. Dong, "How Net and the Computation of Meaning", World Scientific Publishing Company, 2006.

[9] Q. Liu and S.J. Li, "Word Similarity Computing Based on How-net", Computational Linguistics and Chinese Language Processing, Vol. 7, No. 2, Pp. 59–76, 2002.

[10] D. Chun, "On Indexing of Key Words", Acta Editologica, Vol. 16, No. 2, Pp. 105–106, 2004.

[11] J. Morris and G. Hirst, "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text", Computational Linguistics, Vol. 17, No. 1, Pp. 21–48, 1991.

[12] H.G. Suo, Y.S. Liu and S.Y. Cao, "A Keyword Selection Method based on Lexical Chai", J. Chinese Information Processing, Vol. 20, No. 6, Pp. 25–30, 2006.