Design and Implementation of Web Crawler with Real-Time Keyword Extraction based on the RAKE Algorithm

Fei Zhang*, Sunggyun Jang, Inwhee Joe
Dept. of Computer and Software, Hanyang University
*e-mail: zhangfei2010@hotmail.com

Abstract

We propose a web crawler system with keyword extraction function in this paper. Researches on the keyword extraction in existing text mining are mostly based on databases which have already been grabbed by documents or corpora, but the purpose of this paper is to establish a real-time keyword extraction system which can extract the keywords of the corresponding text and store them into the database together while grasping the text of the web page. In this paper, we design and implement a crawler combining RAKE keyword extraction algorithm. It can extract keywords from the corresponding content while grasping the content of web page. As a result, the performance of the RAKE algorithm is improved by increasing the weight of the important features (such as the noun appearing in the title). The experimental results show that this method is superior to the existing method and it can extract keywords satisfactorily.

1. Introduction

Nowadays, with the explosive growth of the network information, how to obtain useful information from massive data quickly, accurately and effectively has become a key issue in the development of the internet. Automatic keyword extraction has always been a hot topic in the field of network information processing both in domestic and overseas. In the past, in order to implement better decisions, people have to deal with many information with deep levels by experience, extensive computation or human brain intelligence. However, with the development of the Internet and the accumulation of network information, the traditional manual processing methods are gradually becoming unrealistic. In particular, a large amount of information is uploaded to the Internet every day which make us have no time to browse and organize. Therefore, how to obtain useful information quickly, accurately and effectively has become a prominent problem in the development of Internet. Data mining or knowledge discovery in databases is a technology which has developed rapidly to meet this need. The generally accepted definition was proposed by Willima J Frawley, Gregory Piatetsk Shapiro [1] and Ussma M Fayyad: Data mining is the extraction of interesting knowledge from large databases. With the development of data mining technology, the object of data mining is far beyond the scope of database. Data sources can also be data warehouses, text data collections, HTML data sets, web documents, or any data collection. Text mining is a branch of data mining, also known as text data mining, which is roughly equivalent to text analysis. It is the process of obtaining high-quality information from text. Most of the researches on keyword extraction in text mining are based on documents and corpora that have been collected and processed. However, in the light of the large amount of information uploaded to the Internet on a daily basis, the purpose of this paper is to establish a real-time keyword extraction system which can capture the keywords of the corresponding text while grasping the text of web page and store them into the database at the same time. Specifically, we design and implement a crawler combining RAKE text mining keyword extraction algorithm which can extract keywords from the corresponding web pages while crawling web pages. Moreover, considering the characteristics of the web news document, the performance of the algorithm is improved by increasing the weight of important features (such as the noun appearing in the title). Finally, the effectiveness of this method is proved by some experiments.

2. Related Words

2.1 Keyword Extraction Method

Keyword extraction is a process of extracting representative keywords from text, and it has important application value in the field of text processing. At present, many scholars and researchers have done a lot of research in the field of keyword extraction and they have proposed a number of representative methods. According to the need to pre-mark the training corpus, the method of keyword extraction can be divided into two categories: supervised and unsupervised. The method of keyword extraction can also be divided into three categories: statistical extraction method, linguistic based extraction method and machine learning based extraction method.

The keyword extraction method which is easiest to consider is to count the frequency and position of candidate words in the text, then the weighted sum of features is weighted. At last, select N words with high

weight value as the keywords of the article. Karen Sparck Jones (1972) [2] and Salton et al. (1975) [3] laid the foundation for this. The TF-IDF algorithm proposed by Salton et al. is a common statistical extraction method. This method uses frequency features implemented to sequence the candidate words. Later research on keyword extraction applied this result to select keywords for individual documents. Mihalcea R et al. (2004) [4] proposed TextRank algorithm. This method constructs the graph network according to the word co-occurrence and phrases, and then extracts the keywords by sorting. Compared with TF-IDF algorithm, the Text Rank Algorithm not only takes into account the characteristics of word frequency, but also takes into account the location features between words. Matsuo and Ishizuka (2004) [5] also use word co-occurrence to extract keywords. The method first clusters the candidate words into N subsets according to the co-occurrence frequency, and then determines the keywords according to the bias of the calculated candidate words for each subset. Stuart Rose et al. (2010) [6] proposed RAKE algorithm. It is an unsupervised, domain independent, language independent method for extracting keywords from individual documents, and the authors demonstrate the keywords extracted by RAKE algorithm is better than TextRank. With the development of Natural Language Processing technology, linguistic methods are gradually applied to keyword extraction. However, it is worth noting that linguistic methods are often used in conjunction with statistical methods or other methods in keyword extraction. In 1990s, with the emergence and development of machine learning, keyword extraction based on machine learning emerged gradually. Witten I H et al. (1999) [7] propose the extraction method based on Naive Bayes which is KEA algorithm. K. Zhang and H. Xu (2006) [8] calculate the features of candidate words, construct feature vectors, and then extract the keywords using support vector machine (SVM) model. T. Jo et al. (2006) [9] select the frequency characteristics of their candidate words, and determine whether that appears in the title and articles as input features of the neural network for keyword extraction.

With the development of keyword automatic extraction technology, researchers will combine many methods to improve the extraction effect. Due to the different features of the text, the different keyword extraction methods usually have different effects. So far, most scholars fuse different methods together to achieve the purpose of improving the extraction effect. N. G. Ali et al. (2015) [10] used this method.

The RAKE algorithm which is an unsupervised keyword extraction algorithm, is a single text analysis algorithm. And the RAKE algorithm is applied to individual documents and does not have to follow specific syntax rules. In contrast, this algorithm is more applicable to a real-time keyword extraction of web news content.

2.2 Web Crawling

Web crawler is also known as web spider or web information collector. It can crawl web pages and extract web content automatically. Usually it begins operation from a seed called uniform resource locator (URL) list. When accessing these URLs, the crawler identifies all hyperlinks on the page into a "pending list" which is known as the crawl frontier and the so-called crawling territory (crawl frontier). The URL in this territory will be visited recycled in accordance with some strategies and download the page referred to and analyze the page content to extract the new URL into the coexistence of crawling in the URL queue, and repeat the above process until the URL queue is empty or satisfy a creeping termination conditions so that the web traversal is over. [11] This process is called web crawling. In addition, the "pending URL queue" can also be a priority queue, that is, through certain conditions to control reptiles crawling these URLs in sequence.

3. Web Crawler with Real-Time Keyword Extraction

3.1 Web Crawler with RAKE

3.1.1 RAKE

RAKE algorithm was put forward in 2010 in unsupervised keyword extraction method and has been proved to be better than TextRank algorithm. The RAKE algorithm is used for keyword extraction. The RAKE algorithm is used for keyword extraction. In fact, what is extracted is the key phrase, and it tends to longer phrases. In English, the keyword usually includes more than one word, but few punctuation and stopwords are included such as and, the, of and other words which do not contain semantic information. The RAKE algorithm first uses punctuation marks (such as the half period, question mark, exclamation comma) a document into several clauses. Then, each clause is divided into phrases using the stopword as a separator. These phrases are used as the candidate words for the final extracted keywords. So, how to measure the importance of each phrase? We notice that each phrase can be divided into several words by space, and each phrase can be scored by giving each word a score, the score of each phrase is obtained by adding together. A key point is to take into account the co-occurrence of each word in the phrase.

The final formula is defined as

$$\text{wordScore} = \frac{wordDegree(w)}{wordFrequency(w)}$$

That is, the score of the word "W" is the degree of the word (it is a concept of network, every time when it co-occurs with a word in a phrase, the degree is added 1 considering the word itself) divides the word frequency (the total number of words that appears in this document). Then, for each candidate key phrase, each of the word scores is accumulated and sorted. RAKE considers the first 1/3 of the candidate phrases as the extracted keywords.

3.1.2 Scrapy

Scrapy is an open source web crawler framework developed with Python which can be used to quickly grab web sites and extract structured data efficiently from pages. [12]

Scrapy is based on the Twisted asynchronous network library for processing communications. The architecture of Scrapy is clear, and it includes a variety of middleware interfaces which can flexibly complete various needs. The working principle is as follows: Firstly, from the seed URL, the scheduler will send it to the downloader to download, and then it will be handed to the crawler for analysis to process separately according to the results. If it needs the further crawling links, these links will pass for a callback; if it is the data that needs to be saved, it is sent to the project pipeline component for later processing including detailed analysis, filtering, storage and so on. In addition, data flow channels are also allowed to install various middleware for necessary processing.

Scrapy can be widely used in data mining, monitoring and automated testing which provides the base class for many types of reptiles, such as BaseSpider, SitemapSpider and so on.

3.2 Web Crawler with Advanced RAKE

Compared with common documents, news web pages have their own unique content. It is embodied in the following aspects: ①Usually, news pages are relatively short and important words or phrases will appear repeatedly. ②News headlines usually embody the subject matter of news. Because the news page has the unique content of the above features, we can improve the classic RAKE formula. Based on the premise that the score of candidate keywords is unchanged under the original algorithm, we improve the weight of news headlines and nouns, construct a comprehensive score calculation formula as follows and then extract keywords according to the importance degree.

The formula is:

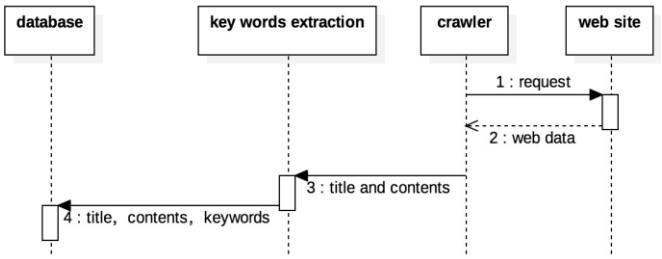
$$\text{wordScore} = \frac{wordDegree(contents) + 2 \cdot wordDegree(title)}{wordFrequency(contents) + wordFrequency(title)}$$

4. Experiment

We chose the YAHOO news web page as the final target page for extracting keywords. Specifically, the extraction experiments are carried out for the news in YAHOO news science column. During this experiment, we crawl the pages of the text. As we all know, there are a lot of noise in news pages including navigation, advertising, related links, copyright notices and so on. Therefore, must filter and extract the text before extracting keywords. And while crawling, each of the news has implemented a keyword extraction. In the experiment process, we realized separately the crawler which combines the RAKE algorithm and the crawler which combines the improved RAKE algorithm.

News web keyword extraction includes the following core processing steps: ①Extract web text and filter out the noise content of the web, including advertising, related links, copyright and so on. ②Filter the words that will not be keywords such as the stopwords and vocabulary of keywords, and establish the initial set of candidate keywords. ③Calculate the composite score and reduce the candidate keywords set according to the size of the score. Default the score to the first 1/3 words or phrases to form the keyword set. Figure 4.1 shows the interaction of the web crawler with keyword extraction.

<Figure 4.1> Interaction of the web crawler with keyword extraction



In this paper, the extraction experiments are carried out for the 100 news in YAHOO news science column. The number of keywords to be extracted each set 6, and manually annotated keywords for each news. We compared the results to manually annotated keywords for the keywords extracted by crawler which combines the RAKE algorithm and the crawler which combines the modified RAKE algorithm.

At present, there is no uniform evaluation method for the quality and evaluation standards of keyword extraction in domestic and overseas. This paper uses a common evaluation method of extracting keywords. This method is to compare the result of the computer automatic extraction with manual annotation, using the precision rate P, the recall rate R and F1 value to evaluate the effectiveness of the improved algorithm.

Calculation formula as follow:

$$P = \frac{Num(correct\ keywords\ of\ extraction)}{Num(keywords\ of\ extraction)}$$

$$R = \frac{Num(correct\ keywords\ of\ extraction)}{Num(keywords\ of\ document)}$$

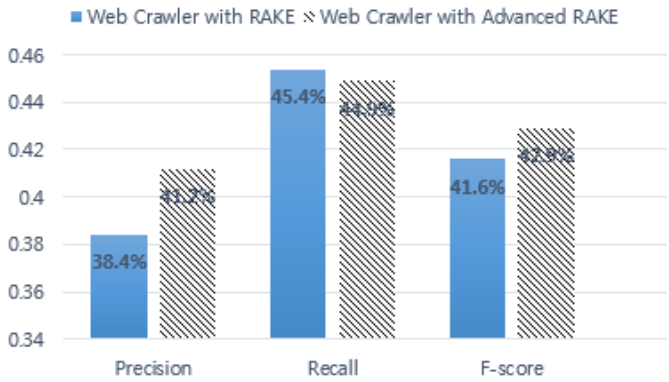
$$F1 = \frac{2PR}{P+R}$$

Table 4.1 and Figure 4.2 shows the results.

<Table 4.1> Comparison of results about real-time keyword extraction on 100 news

Num. of Keywords		6	
Method		Web Crawler with RAKE	Web Crawler with Advanced RAKE
Evaluation value	P (precision)	0.384	0.412
	R (recall)	0.454	0.449
	F_1 (F-score)	0.416	0.429

<Figure 4.2> Comparison of results about real-time keyword extraction on 100 news



As can be seen from Table 4.1 and Figure 4.2, the improved algorithm is better than the original algorithm. The precision, recall rate and F1 measure value both have a corresponding increase. It proves that our modified algorithm is effective.

5. Conclusion

In this paper, a real-time keyword extraction system is proposed, to achieve the extraction of keywords while grasping the news web massage. Through mining news content, we can find the current news hot spots and give feedback to the users in time. This paper designs and implements a web crawler combining RAKE text mining keyword extraction algorithm. On the basis of the content characteristics of news web pages, the author improves the classical RAKE algorithm formula and constructs a candidate keyword scoring formula for news texts. The performance of the algorithm is improved by increasing the weights of important features. The experimental results show that the improved RAKE algorithm can improve the accuracy of keywords compared with the improved algorithm. This method is obviously superior to the existing method and can extract keywords satisfactorily.

Of course, this method also has shortcomings and needs improvement and optimization. The author encounters some problems when using RAKE algorithm to process news in other languages, such as the effect of Korean to use the stopwords to divide a sentence into phrases is far less effective than English. In this regard, how to extract keywords from multi language news will be the next step for the author's research.

Reference

[1] Frawley W J, Piatetsky-Shapiro G, Matheus C J. Knowledge discovery in databases: An overview[J]. AI magazine, 1992, 13(3): 57.

[2] Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of documentation, 1972, 28(1): 11-21.

[3] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[4] Mihalcea R, Tarau P. TextRank: Bringing Order into Text[C]//EMNLP. 2004, 4: 404-411.

[5] Matsuo Y, Ishizuka M. Keyword extraction from a single document using word co-occurrence statistical information[J]. International Journal on Artificial Intelligence Tools, 2004, 13(01): 157-169.

[6] Rose S, Engel D, Cramer N, et al. Automatic keyword extraction from individual documents[J]. Text Mining: Applications and Theory, 2010: 1-20.

[7] Witten I H, Paynter G W, Frank E, et al. KEA: Practical automatic keyphrase extraction[C]//Proceedings of the fourth ACM conference on Digital libraries. ACM, 1999: 254-255.

[8] Zhang K, Xu H, Tang J, et al. Keyword extraction using support vector machine[J]. Advances in Web-Age Information Management, 2006: 85-96.

[9] Jo T, Lee M, Gatton T M. Keyword extraction from documents using a neural network model[C]//Hybrid Information Technology, 2006. ICHIT'06. International Conference on. IEEE, 2006, 2: 194-197.

[10] Ali N G, Omar N. A hybrid of statistical and machine learning methods for Arabic keyphrase extraction[J]. Asian Journal of Applied Sciences, 2015, 8(4): 269-276.

[11] Cho J. Crawling the web: discovery and maintenance of large-scale web data[J]. A Thesis Nov, 2001.

[12] Scrapy A. Fast And Powerful Scraping And Web Crawling Framework[J]. Scrapy. org. Np, 2016.