Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 187 (2021) 77–82

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Identification, Information and Knowledge in the internet of Things, 2020

Extracting Keywords from Texts based on Word Frequency and Association Features

Zhenzhen Xu[a], Junsheng Zhang[a,*]

[a]Institute of Scientific and Technical Information of China, Beijing 100038, China

Abstract

With the development of information technology such as mobile Internet and social media applications, network information is growing rapidly and leads to the problem of information overload. Keywords help to filter and find interesting information for users from massive text. Automatic extraction of keywords from text as tags of text help to improve recommendation and keyword-based information retrieval. This paper proposes a novel keyword extraction approach from text that combines features such as word frequency and association. Experiment results show that the precision rate, recall rate and F-measure are all better than those of TextRank and TF-IDF.

Keywords: TextRank; TF-IDF; Keyword extraction; Text

1. Introduction

With the advent of the big data era, information has been increasing exponentially. Traditionally, people acquire information from books, newspapers and magazines. Now they are used to acquire information via the Internet. Texts is one of the main information formats of information. For textual information, a keyword set consists of several words, which can express the meaning of the text. Keywords can help users quickly understand the topics of text. Besides, keyword extraction is the basis of applications such as summarization, information retrieval, text classification and clustering.

In the early stage, keyword are manually extracted from the text [13]. Manual extracting and tagging keywords are time-consuming and labor-intensive. The extraction results are subjective, which is difficult to objectively reflect the meaning of texts. With the quickly increase of information, it is difficult to manually extract keywords. Therefore, it is urgent to automatically extract keywords from texts.

* Corresponding author. Tel.: +86-010-5882455
E-mail address: zhangjs@istic.ac.cn

10.1016/j.procs.2021.04.035

Methods of extracting keywords can be divided into two types: unsupervised keyword extraction and supervised keyword extraction [12].

- Unsupervised keyword extraction does not need manual labeling corpus. Xu [10] proposed a keyword extraction algorithm based on TF-IDF method. Yang [11] proposed an improved TF-IDF algorithm, which combined information gain, discrete quantification and multi-feature fusion methods. Mihalcea et al [7] proposed TextRank method, which has better performance of keyword extraction. Gu [3] proposed a keyword extraction algorithm LTWPR based on the PageRank algorithm, which combines the local and global features of text to get better extraction effect. Li [4] proposed an improved keyword extraction algorithm based on TextRank, which uses the TF-IDF and average information entropy to calculate the importance of words in text, and the results show that the extracted keywords have higher precision rate. Fang [1] proposed a keyword extraction approach for academic texts based on TextRank.
- Supervised keyword extraction requires training corpora to convert keyword extraction to a classification problem. Turney et al. [8] proposed GenEx algorithm, which combines genetic algorithm and machine learning to automatically extract keywords. Witten Et al. [9] proposed the KEA (Keyphrase Extraction Algorithm), which uses Naive Bayes model to construct two classifiers. Since supervised keyword extraction requires training corpora, and the application scopes are limited because of the limitation of corpora. Gu et al. [2] proposed a keyword extraction approach based on LDA and TextRank, which combined the information of the internal structure of documents and the subject information of documents to extract keywords. However, this approach is not fit for texts without obvious topic distribution. Liu [5] proposed a keyword extraction method of current affairs news based on Word2Vec and TextRank, but it requires the training corpus. Ning [6] proposed a method for extracting keywords from news texts that integrates LSTM and LDA, but it does not consider relationships of the semantic importance, coverage and difference of keywords.

TF-IDF is used to measure the importance of words in texts. It is simple to implement, but it is assumed that the words are independent and cannot reflect the relationship between words and the position of words. When extracting keywords, the position of the word information, such as text titles, abstracts, and first and last sentences of paragraphs, contains important information. TextRank is developed based on PageRank, which considers the correlations between words. In the initial stage, the importance of each word is the same.

To improve the effect of keyword extraction, this paper proposes an approach that combines word frequency and association features. It not only considers the frequency of words, but also the relationship between words. The basic idea is to take the intersection of keywords extracted separately by TF-IDF and TextRank as the extracted keywords.

2. Our Method

The basic idea of our method is to use the intersection of the keyword sets extracted by TF-IDF and TextRank as the final keyword set. If the number of final keywords n is greater than or equal to the number of keywords needed m , then the first m of the final number of keywords are used as the keywords of the method. If the final number of keywords n is less than the required number of keywords m, then take $(m-n)$ keywords from the keywords extracted by TF-IDF, and take the union with the final keyword as the keyword set. The experiment uses precision, recall and $F-measure$ as the evaluation indicators of keyword extraction results.

Before text preprocessing, keywords need to be labeled manually. Since the keywords are manually labeled, there may be many inconsistencies between different labeling results. For example, keywords may be separated by spaces, Chinese commas, English commas or a space after each keyword. Therefore, the original text needs to be preprocessed, and the subsequent text segmentation and stop-words are performed in TF-IDF and TextRank.

TextRank [7] is adapted from PageRank algorithm. The main idea is to put every document as a node in the graph, and the co-occurrence relationship between words and words as edges, construct a keyword graph $G = (V, E)$, where V is the node set, and E is the edge set. $S(V_i) = (1 - d) + d \times \sum_{j \in in(v_i)} \frac{1}{|out(v_j)|} S(v_i)$, where $in(v_i)$ represents the set of nodes pointing to v_i , $out(v_j)$ represents the set of nodes pointed out from v_j , and d represents the damping factor, d is 0.85 in this experiment .

2.1. Combining TextRank and TF-IDF

The flow chart of the integrated algorithm based on TextRank and TF-IDF is shown in Fig. 1. Kws_TFIDF is the keywords from TF-IDF, $Kws\_TextRank$ is the keywords from TextRank, $len(Kws1)$ is the number of keywords by our method, m is the number of keywords needed, and keywords is the keywords finally extracted by our method.

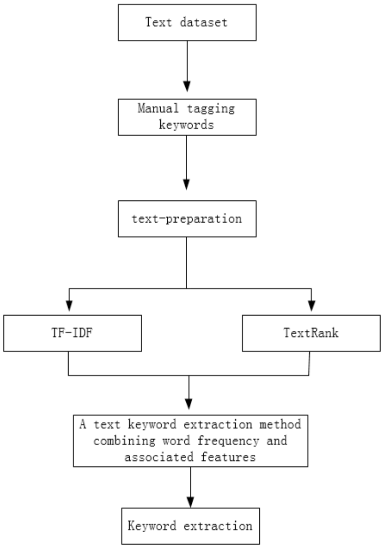


Fig. 1. Combining TextRank and TF-IDf

3. Experiment and Discussion

There are two experiment data sets:

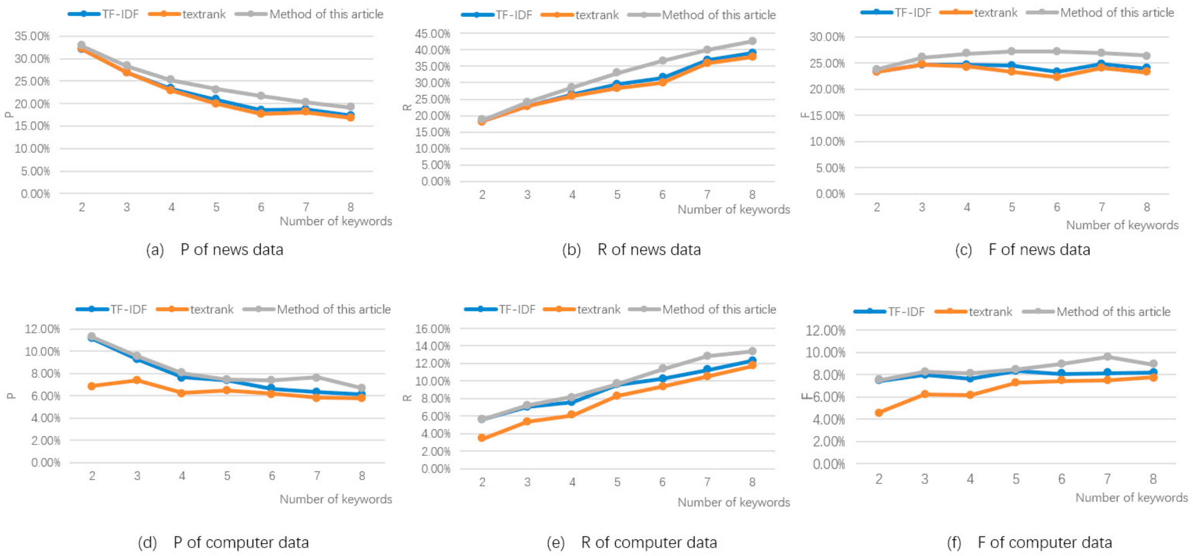
1. News dataset: 1525 news are crawled from Southern Weekend (http://www.infzm.com/content);
2. Computer dataset: it provides the Chinese text data set including 9804 documents of train.zip and 9832 documents of test.zip, and all the documents are divided into 20 categories (https://download.csdn.net/download/number59/11374484). In the experiment, we selects 108 articles in the C19-Computer category. Raw news text data set package for experimental data collection include title, keywords, URL and text. Each piece of data in the original C19-Computer contains title, name, abstract, keywords and text.

The evaluation indicators used are precision rate (P), recall rate (R), and the F-measure is used as an evaluation index.

3.1. Extracting keywords with different numbers

We use TF-IDF, TextRank and our algorithm to collect keywords from experimental data. The average number of keywords in the original tags of the news data set and the C19-Computer dataset are 3.63 and 4.74, respectively , so the number of keywords extracted by the three algorithms is 2~8, and precision rate ***P***, recall rate ***R*** and ***F*** value are compared, as shown in Fig. 2. The extraction results of TextRank+TF-IDF integrated algorithm are better. With the increase of the number of extracted keywords, the precision of extraction results is decreasing, while recall rates rise, and ***F*** value seems stable. ***F*** value is an indicator that comprehensively considers precision rate and recall rate.

From Fig. 2(c), ***F*** value of news text data shows that the optimal number of keywords extracted by TF-IDF and TextRank is 7. ***F*** value of C19-Computer data in Fig. 2(f) shows that the optimal number of keywords extracted by TF-IDF is 5. The number of optimal extracted keywords in TextRank increases with the number of extracted keywords. It can be seen form the analysis of Fig. 2(d) and (e) that when the number of extracted keywords is 5 , the precision of extraction is the largest. The precision rate decreases with the number of keywords is increased.



(a) P of news data

(b) R of news data

(c) F of news data

(d) P of computer data

(e) R of computer data

(f) F of computer data

Fig. 2. Combining TextRank and TF-IDf

3.2. Impact of the number of keywords extracted by separate algorithm on final extraction result

The number of keywords extracted by a single algorithm is an important factor affecting the final keyword extraction effect. In order to verify how the number of keywords extracted by the separate algorithm affects the final keyword extraction effect, this experiment sets the final extracted keywords as Q. The prerequisite for a successful experiment is to ensure that the number of keywords extracted by a single algorithm is greater than the number of final keywords extracted, Therefore the number of keywords extracted separately is $Q+1$, $Q+2$, $Q+3$, $Q+4$ and $Q+5$. Taking into account the average number of keywords of the news data set is 3.63, and the average number of keywords in the original tag of the C19-Computer data set is 4.74. Take Q as 3, 4 and 5 for experiment.

Table 1 shows the impact of the number of keywords extracted by a separate algorithm of news text on the final extraction results. From the results, when the number of keywords collected by a single algorithm is 7, no matter how much Q is taken, F value is the largest, and the extraction effect is the best. The reason for the analysis is that the last experiment shows that the optimal number of keywords extracted by

TF-IDF and TextRank is 7. Under the optimal number of extracted keywords, the integration method is also optimal.

Table 1. Impact of the number of keywords extracted from news dataset on final extraction results

#Selected Keywords=3					
#keywords	4	5	6	7	8
P	27.72%	27.36%	27.38%	29.63%	29.61%
R	23.62%	23.35%	23.33%	25.25%	25.22%
F	25.51%	25.20%	25.19%	27.27%	27.24%
#Selected Keywords=4					
#keywords	5	6	7	8	9
P	24.76%	23.91%	25.98%	25.83%	25.69%
R	28.12%	27.14%	29.45%	29.26%	29.11%
F	26.33%	25.42%	27.61%	27.44%	27.29%
#Selected Keywords=5					
#keywords	6	7	8	9	10
P	22.28%	23.82%	23.46%	23.38%	23.13%
R	31.57%	33.81%	33.23%	33.10%	32.75%
F	26.12%	27.95%	27.50%	27.40%	27.11%

Table 2 shows the effect of the number of keywords extracted by the C19-Computer text algorithm on the final extraction result. If the view when $Q = 3$, the number of keywords in a separate extraction is 5 or 6 under the best extraction; When $Q = 4$, the extraction effect is best when the number of keywords extracted separately is 5; when Q=5, the extraction effect is best when the number of keywords extracted separately is 8. The reasons for the analysis are as follows: the optimal number of keywords extracted by TF-IDF is 5, but TextRank cannot determine the optimal number of extracted keywords, so when the number of extracted keywords is greater than or equal to 5, the optimal number of keywords extracted cannot be determined.

4. Conclusion

This paper presents an appraoch extracting keyword text term frequency associated with the fusion characteristics. The intersection of TF-IDF keywords and TextRank keywords extracted separately is taken as the final keywords. This integration approach takes into account both word frequency and association characteristics between words.

Experiments were carried out on two datasets, and results showed that the precision and recall rates of our method are improved compared with the extraction of keywords using TextRank and TF-IDF alone. The disadvantage is that when TextRank and TF-IDF were selected, parameters in each algorithm are not modified in detail, and the optimal parameter values are selected for experiments. In addition to the keyword extraction algorithm used in this paper, there are many other methods that can be used for integration, such as TextRank and LDA integration, TF-IDF and LDA integration, and these issues can be further deepened in future research.

Acknowledgments

This research work was partially supported by National Natural Science of China (Grant Number 71503240) and the National Key Research and Development Project (Grant Number 2019YFA0707201).

Table 2. Impact of the number of keywords extracted from computer dataset on extraction results

#Selected Keywords=3					
#keywords	4	5	6	7	8
P	9.57%	9.88%	9.88%	8.95%	9.57%
R	7.30%	7.56%	7.56%	6.64%	7.10%
F	8.28%	8.57%	8.57%	7.62%	8.15%
#Selected Keywords=4					
#keywords	5	6	7	8	9
P	9.26%	8.33%	7.41%	7.64%	7.64%
R	9.60%	8.52%	7.41%	7.64%	7.64%
F	9.43%	8.42%	7.41%	7.64%	7.64%
#Selected Keywords=5					
#keywords	6	7	8	9	10
P	7.78%	7.41%	7.78%	7.22%	7.22%
R	10.02%	9.68%	10.09%	9.29%	9.29%
F	8.76%	8.39%	8.79%	8.13%	8.13%

References

[1] Fang, J., Cui, H., He, G., Lu, W., 2019. Keyword extraction of academic text with textrank model based on prior knowledge. Information Science 37, 77–82.

[2] Gu, Y., Xia, T., 2014. Study on keyword extraction with lda and textrank combination. New Technology of Library and Information Service 30, 41–47.

[3] Gu, Y., Xu, M., 2017. Keyword extraction from news articles based on pagerank algorithm. Journal of University of Electronic Science and Technology of China 046, 777–783.

[4] Li, Z., Pan, S., Dai, J., Hu, J., 2020. An improved textrank keyword extraction algorithm. Computer Technology and Development 030, 77–81.

[5] Liu, Q., Shen, W., 2018. Research of keyword extraction of political news based on word2vec and textrank. Information Research 000, 22–27.

[6] Ning, S., Yan, X., Zhou, F., Wang, H., Zhang, J., 2020. A news keyword extraction method combining lstm and lda differences. Computer Engineering and Science 042, 153–160.

[7] Rada, M., Paul, T., 2004. Textrank: Bringing order into texts, in: Proc Conference on Empirical Methods in Natural Language Processing, pp. 404–411.

[8] Turney, P.D., 2002. Learning to extract keyphrases from text. CoRR cs.LG/0212013. URL: http://arxiv.org/abs/cs/0212013.

[9] Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G., 1999. Kea: Practical automatic keyphrase extraction, in: Proceedings of the Fourth ACM Conference on Digital Libraries, Association for Computing Machinery, New York, NY, USA. p. 254¨C255. URL: https://doi.org/10.1145/313238.313437, doi:10.1145/313238.313437.

[10] Xu, W., Wen, Y., 2008. A chinese keyword extraction algorithm based on tfidf method. INFORMATION STUDIES: THEORY & APPLICATION 31, 298–302.

[11] Yang, K., 2015. Research on Automatic Keyword Extraction Algorithm Based on Improved TFIDF. Ph.D. thesis. Xiangtan University.

[12] Zeng, P., Tan, Q., Yan, Y., Xie, Q., Xu, J., Cao, W., 2017. Automatic keyword extraction using word embedding and clustering, in: 2017 International Conference on Computer Systems, Electronics and Control (ICCSEC), pp. 1402–1408. doi:10.1109/ICCSEC.2017.8447033.

[13] Zhao, J., Zhu, Q., Zhou, G., Zhang, L., 2017. Review of research in automatic keyword extraction. Journal of Software 28, 2431–2449.