Keyword Extraction Method for Complex Nodes Based on TextRank Algorithm

Zhou Qingyun[1], Fang Yuansheng[2], Shang Zhenlei[1], Zhong Wanli[2]
[1]The Training and Evaluation Center of GPGC
510520, Guangzhou, Guangdong,
[2]Guangdong Electric Power Science Academe
510520, Guangzhou, Guangdong

Abstract: **Keywords are important way for people to quickly understand content of document and grasp subject, and keyword extraction technology is significant way to quickly obtain core meaning of text information, which has wide range of applications in fields such as intelligence, journalism, information retrieval and natural language understanding. However, traditional TextRank algorithm refers to local co-occurrence relationship among text words, which does not pay much attention to complex network structure characteristics of word graphs. Therefore, structure of network is adjusted by removing nodes to separate sub-networks with layers. Moreover, taking into account complex network structure characteristics of word nodes, method of word node removal is introduced as well. Meanwhile, value of sliding window is increased so that ranking can be obtained through multiple iterations, and then one of the highest ranking keyword nodes will be removed in turn. Besides it, Keyword extraction is then performed on each subtopic where sub-keywords are determined based on ranking of candidate keywords, and key nodes of text network are added to sequence so that keyword extraction can be achieved, which achieves improvement of traditional TextRank algorithm, and accuracy, recall, and F value are all improved as well.**

Keywords: text complex nodes, keywords, TextRank algorithm

I. Introduction

In the field of personalized recommendation, prevalence of the Internet has led to a surge in data volume, which is difficult for people to directly obtain useful information from huge data. What's more, review text not only describes true features of product, but also contains abundant user opinion information that reflects personal preferences. If user preference information is mined from review text, favorite items by users with similar preference characteristics can be recommended to the user. In addition, faced with overloaded review information, merchants urgently need to quickly and accurately grasp user attitudes, and then make targeted responses to improve projects. Moreover, since topic distribution of text is hierarchical, the larger the PageRank value of certain node in complex network is, the larger the PageRank value of all other nodes connected to the node will be. Therefore, text keywords are directly extracted with TextRank algorithm where keywords obtained are adjacent to word nodes with the highest TextRank value, which cannot be fully extracted. However, a Chinese text can usually be divided into different topic areas, and a topic area is called a semantic block in this paper where keywords of text are composed of total keywords and sub-keywords in each semantic block. Based on above assumptions, keyword extraction process of a Chinese article is considered as division of semantic blocks and keyword extraction of each semantic block in this paper, that is, division of sub-graphs and ordering of nodes in complex networks. In addition, when studying the importance ranking of nodes, node ranking method in traditional complex networks is not applicable to text[1-2].

On the basis of original TextRank algorithm, method that removes important nodes multiple times and iterative calculates importance of nodes is adopted so that sub-graph division and node importance ranking of word graph network can be performed, and important nodes in entire word graph network and important nodes in each sub network are used as keywords of text. In addition, text is first needed to be transformed into graph representation to sort text word nodes.

II. Keyword Construction Method

A. Sub-netting of Word Graph Networks

Sub-netting and node ranking algorithm used in this paper is based on following assumptions. If a network contains a series of sub-networks which are related to each other based on some nodes, these sub-networks can be divided by removing these nodes. Additionally, after conducting experiments, it is found that such nodes usually have high PageRank values, and definition of special nodes in word graph is offered in this paper[3-4].

Definition 1: Bridge node is node that connects multiple different communities.

Definition 2: Secondary bridge node is node connecting bridge node and community.

Definition 3: Community leader node is node with the highest PageRank value in a community.

As shown in Figure 1, graph G consists of multiple sub-networks and some connected nodes. According to TextRank algorithm, bridge node A and secondary bridge node B have the highest importance.

978-1-7281-5904-1/20/$31.00 ©2020 IEEE
DOI 10.1109/ICCEA50009.2020.00084

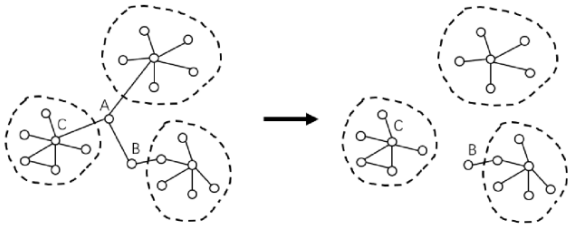


Figure 1 Schematic Diagram of Word Graph Network Structure

However, after bridge node A is removed, connectivity of whole graph changes drastically, and influence of node B will decrease. Moreover, in text network, node A is the hub that connects different semantic blocks, and there is no doubt that it is the keyword of text. However, B is just edge node of a semantic block, which is not so important. On the contrary, after removing node A, ranking of community leader node C has been improved. In fact, position of C can be regarded as center of a semantic block. Therefore, it can be considered that rank of AC is higher than that of B[5].

B. Keyword Node Sorting

Keyword node importance ranking method used in this paper is alternated by the following two steps.

Rank 1 step: Node importance ranking algorithm in complex network is used to perform preliminary importance ranking on nodes in current word node network.

In this paper, TextRank, a common method for text keyword extraction, is selected as method used in each R1 step.

Formula of original TextRank is as follows.

$$ws\left(V_i\right)=\left(1-d\right)+d\times\sum_{v_j\in in\left(v_j\right)}\frac{w_{ij}}{\sum_{v_k\in out\left(v_j\right)}w_{jk}}ws\left(v_j\right) \quad (1)$$

where WS (V) is score value of node V, and Wij is weight term that is increased from original PageRank, which is used to represent weight of weighted edge between two nodes. Since only problem of complex network structure characteristics of word graphs in original TextRank algorithm is addressed in this paper, which does not weight edges of word nodes, weights are all set to 1. Moreover, d refers to smoothing factor, which is used to ensure that each word node has a score greater than 0, and general value of d is 0.85. Besides it, In(v) and Out(v) respectively represent set of nodes that point to node V and nodes that point to node V [6-7].

Score of each word node can be obtained according to formula (1), and then important keyword nodes obtained by each round of Rank steps will be determined according to score level.

Remove 2 step: Node with the highest node importance sequence score at step R1 is removed to generate new graph without this node so that preliminary network division can be completed.

Specifically, G', subgraph of G needs to meet following requirements.

$$G'\left(V',E'\right)\subseteq G\left(V,E\right) \quad (2)$$

where V' and E' respectively indicates node set and edge set after removing node, whose value is determined by following formula.

$$V'=\left\{n\middle|n\in v,\quad n\neq N_r\right\} \quad (3)$$

$$E'=E-E_{N_r} \quad (4)$$

E_{Nr} is obtained by following formula, where $E_{N_r}=\lambda\left(N_r\right)$ represents all connected edges whose value is N_r.

$$E_{N_r}=\lambda\left(N_r\right) \quad (5)$$

N_r is node with the highest importance sequence score obtained in step R1.

$$N_r=\arg\max_{N}PR(G) \quad (6)$$

where PR (G) means to find PageRank value of all nodes in graph G.

Words represented by removed node are put into extracted keyword list. Meanwhile, word graph network after removing nodes is used for next R1 step and R2 step operation until corresponding number of keywords are taken out. According to steps mentioned above, core code of keyword extraction algorithm RemoveRank proposed in this paper is as follows[8].

Table 1 Pseudo Code of Core Part in Keyword Extraction Algorithm RemoveRank

Algorithm 1	RemoveRank Algorithm
Input:	Graph text word graph, TopK keyword extraction number
Output:	Word node, Score
one:	Function Remove Rank (Graph, TopK)
two:	rank←[]
three:	While TopK>0do
four:	rank←Text Rank(Graph)
five:	Graph←Remove Node(Graph, rank[0])
six:	TopK←TopK−1
seven:	Insert rank[0]into Score
eight:	End while
nine:	Return Score
ten:	End function

III. Experiments

A. Experimental Environment

Operating environment in this experiment is as follows.

Table 2 Experimental Running Environment

Operating environment	Configuration and application range
processor	Intel I7 9700k
RAM	32GB

operating system	Ubuntu18.04LTS
Programming environment	Python3.8
Experimental procedure of algorithm	Based on network package and related open source programs
Chinese word segmentation	Open source *jieba* word segmentation package

B. Processing of Experimental Data

Since experiments in this paper are based on automatic extraction algorithm of web page text, data set provided by which does not contain content of "culture, economy" and other sections, news and economics sections of Southern Weekend website is re-crawled in this paper to extracted content and keywords of text so that meaningless web page structure information can be removed to obtain more than 1,500 Chinese articles.

Secondary annotations is filtered as follows.

The number of keyword tags for each article is at most five.

2. Keywords of article do not include phrases.

3.The number of words in a single article is 200-600.

4. Keywords should be mentioned in article, which is in line with meaning of text.

After removing a certain number of articles that do not meet standards as well as secondary tagging, Chinese keyword extraction data set with keyword tagging of about 1,000 articles is finally formed.

Relevant statistical information of experiments in this paper is shown in Table 3.

Table 3 Relevant statistics Description of Data Set

Statistical indicators	size
Total number of documents	1027
average word nodes in document	336
average sentences in document	73
annotation keywords in Document	3.6

Note: The number of word nodes refers to the number of nodes forming word graph after words not used any more are removed.

C. Experiment Design and Results Analysis

TextRank algorithm and Tf-idf algorithm are selected to simultaneously compare with centrality and closeness of centrality, which are typical methods of node importance ranking in complex networks.

In order to facilitate comparison with existing algorithms, accuracy rate P, recall rate R and F values which are commonly used indicators in information retrieval are used to evaluate effectiveness of keyword extraction algorithm in this paper.

$$p = \frac{w_c}{w_e} \tag{7}$$

$$R = \frac{w_c}{w_s} \tag{8}$$

$$F = \frac{2PR}{P+R} \tag{9}$$

Wc is the number of keywords extracted correctly, We refers to the number of all keywords extracted by certain method, and Ws indicates standard number of keywords provided by document. In addition, TextRank smoothing factor is 0.15 in experiment.

(1) Comparison on sliding window values

Value of sliding window is quite important for method used in this paper. In previous work, few experiments have evaluated some attributes of text network graphs, such as clustering coefficients. However, when graph method is used for keyword extraction, extraction result obviously depends on graph features of word graph. Therefore, in this paper, following experiments are performed for keyword extraction task with a number of 5 and sliding window is selected from 2-10 for key extraction with RemoveRank algorithm proposed in this paper. Meanwhile, clustering coefficients of lexical nodes network are calculated, and comparison results are shown in Table 4.

Table 4 Effects Comparison on Different Values of Sliding Window k

Sliding Window	clustering coefficient	P	R	F
2	0.03	0.289	0.286	0.382
3	0.27	0.336	0.397	0.320
4	0.28	0.288	0.332	0.334
5	0.39	0.399	0.286	0.287
6	0.33	0.334	0.398	0.288
7	0.29	0.287	0.333	0.397
8	0.40	0.401	0.287	0.507
9	0.33	0.398	0.398	0.334
10	0.69	0.334	0.333	0.287

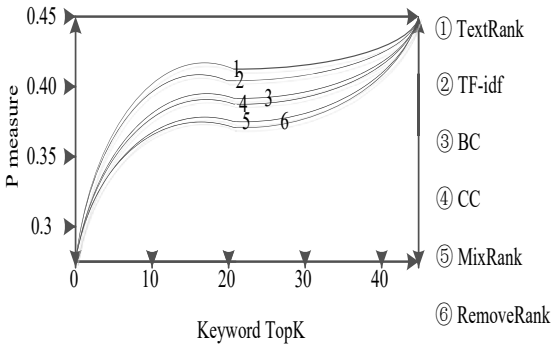
According to Table 4, change of sliding window from 2 to 10 is the most obvious when sliding window value ranges from 2 to 4. However, accuracy rate, recall rate, and F value achieve their maximum values when window is 4. Therefore, algorithm proposed in this paper can get better results when clustering coefficient is relatively reasonable.

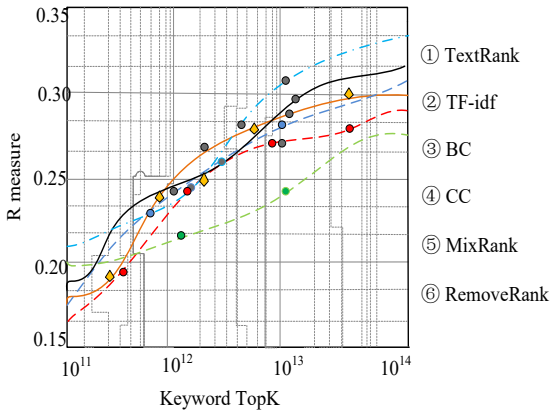
(2) Comparison of different methods on keyword extraction

In order to ensure effectiveness of algorithm, 6 methods is used in this paper to extract 3, 5, 7 keywords under default parameters, and evaluate performance of different algorithms on accuracy rate, recall rate and F-value with the help of data set in this paper. Meanwhile, change of F value under different numbers of keywords is offered in this paper. Results are shown in Figure 2 and Table 5.

Table 5 Comparison of RemoveRank with Other Methods in Southern Weekend Dataset

Draw number	method	P	R	F
3	TextRank	0.366	0.304	0.332
	Tf-idf	0.376	0.313	0.342
	Intermediate centrality(BC)	0.356	0.296	0.323
	Closeness centrality(CC)	0.337	0.281	0.306
	MixRank	0.374	0.311	0.339
	RemoveRank	0.382	0.318	0.347
5	TextRank	0.273	0..379	0.318
	Tf-idf	0.274	0.380	0.319
	Intermediate centrality(BC)	0.262	0.364	0.305
	Closeness centrality(CC)	0.246	0.341	0.286
	MixRank	0.274	0.381	0.319
	RemoveRank	0.291	0.405	0.339
7	TextRank	0.215	0.418	0.284
	Tf-idf	0.219	0.425	0.289
	Intermediate centrality(BC)	0.207	0.403	0.274
	Closeness centrality(CC)	0.197	0.383	0.260
	MixRank	0.215	0.418	0.284
	RemoveRank	0.226	0.439	0.298





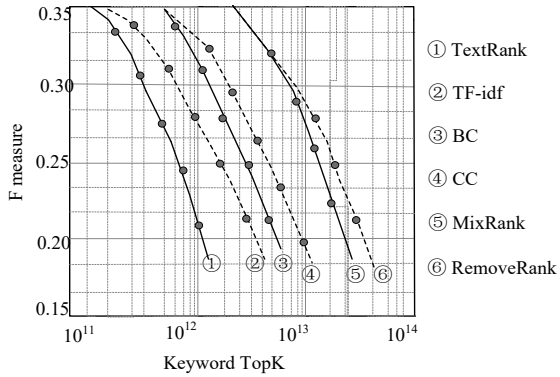


Figure 2 Variation Curves of Accuracy Rate, Recall Rate, and F Value When TopK is from 3 to 10

According to Figure 2 and Table 5, it can be seen that RemoveRank algorithm based on graph structure characteristics proposed in this paper is slightly better thanthat of TextRank and Tf-idf algorithms in P, R, F and other indicators. Compared with TextRank, when TopK = 1012, RemoveRank proposed in this paper has an increase of 6% in F-value index. Moreover, when the number of extracted keywords is 4-7, RemoveRank algorithm improves more than TextRank algorithm since according to strategy of RemoveRank algorithm, keyword extraction is performed by removing nodes, and for extraction of a small number of keywords, such as 1-3, the number of removed nodes is relatively small, which can not cause a large impact on entire word graph network, and the effect also approaches to TextRank algorithm.

(3) Comparison on Typical Keyword Extraction Results

In order to show characteristics of RemoveRank proposed in this paper in detail, some sample documents are selected to extract keywords with TextRank method and RemoveRank method. Moreover, keywords in this paper are compared with keywords marked in document, and extraction results are shown in Table 6.

Table 6 Comparison on Keyword Extraction Case Results

Documentation	method	Tag keywords	Extract keywords
6	TextRank	Highway, cars, speed	Highway, cars, speed
	RemoveRank	Highway, cars, speed	Highway, car, speed, speed limit, speed measurement, penalty
1008	TextRank	Internet, fiber optic cable, server	Internet, connection, server, website, network
	RemoveRank	Internet, fiber optic cable, server	Internet, fiber optic cable, server, website, connection
1364	TextRank	Qinghai Lake, Ta'er Temple, Rape Flowers, Yak	Rape flower, Qinghai Lake, Huangjiao, Yak, known as
	RemoveRank	Qinghai Lake, Ta'er Temple, Rape Flowers, Yak	Rape flowers, Qinghai Lake, Ta'er Temple, Yak, Huangjiao

It can be seen from Table 6 that compared with TextRank algorithm, word node importance ranking with Rank Step for

the first time in this paper can be seen as the number one keyword obtained by original TextRank algorithm. Therefore, in the first keyword extraction, RemoveRank algorithm proposed in this paper is consistent with original TextRank, where effects have a large overlap. However, there has been certain change from the second keyword, which is consistent with the idea of removing nodes in this paper. Taking Document No. 1364 as an example, after word nodes are removed twice, words that have the most impact on document changed significantly. "Huangjiao" is no longer the most important word for document, and "Thar Temple" is more important. Besides it, influence of words "canola flower" and "Qinghai Lake" is removed by RemoveRank algorithm, and important words are filtered from remaining words, the effect of which is better.

IV. Conclusion

TextRank algorithm is improved, and Chinese keyword extraction method based on removal of complex network nodes is proposed in this paper, which improves effectiveness of Chinese keyword extraction algorithm. Meanwhile, the number of TextRank iterations is reduced during actual operation, and time for text keyword extraction is shortened as well. Moreover, experimental results show that for unsupervised keyword extraction algorithm, full consideration on complex network characteristics of word nodes in document itself can affect the importance of a word in text to a certain extent. Therefore, method of node removal in complex networks can better extract keywords.

References

[1] X. H. Qiu, G. J. Li, M. Xiao. Analysis on the Evolution of Research Themes of Foreign Library Science and Information Science——A Case Study of Co-word Analysis [J]. Journal of Information, 2013,32 (12), pp: 110- 118.

[2] W. Wei, X. P. Sun. Key phrase extraction of biomedical literature combining statistics and TextRank [J]. Computer Applications and Software, 2017, 34 (6), pp: 27-30.

[3] X. G. Hu, X. H. Li, F. Xie, et al. Keyword extraction method for Chinese news webpages based on lexical chain [J]. Pattern Recognition and Artificial Intelligence, 2010, 23 (1), pp: 45-51.

[4] J. S. Zhao, Q. M. Zhu, G. D. Zhou, et al. Review of research on automatic keyword extraction [J]. Journal of Software, 2017, 28 (9), pp: 2431-2449.

[5] J. K. Sparck. A statistical interpretation of term specificity and its ap-plication in retrieval [J]. Journal of Documentation, 1972, 28 (1), pp: 11-21.

[6] H.W. Hui, C.C. Zhou, S.G. Xu, F.H. Lin, A Novel Secure Data Transmission Scheme in Industrial Internet of Things, China Communications, vol. 17, no. 1, 2020, pp: 73-88.

[7] F.H. Lin, Y.T. Zhou, X.S. An, I.You, K.R. Choo, Fair Resource Allocation in an Intrusion-Detection System for Edge Computing: Ensuring the Security of Internet of Things Devices, in IEEE Consumer Electronics Magazine, vol. 7, no. 6, 2018, pp: 45-50. doi: 10.1109/MCE.2018.2851723.

[8] J.T. Su, F.H. Lin, X.W. Zhou, X. Lv, Steiner tree based optimal resource caching scheme in fog computing, China Communications, vol. 12, no.8, 2015, pp: 161-168.