Automatic Summarization Generation Technology of Network Document Based on Knowledge Graph

Yuezhong Wu[1,2], Rongrong Chen[3](✉), Changyun Li[1,2], Shuhong Chen[4,5], and Wenjun Zou[1]

[1] College of Artificial Intelligence,
Hunan University of Technology, Zhuzhou 412007, China
yuezhong.wu@163.com, lcy469@163.com, 1450793542@qq.com
[2] Intelligent Information Perception and Processing Technology Hunan Province
Key Laboratory, Zhuzhou 412007, China
[3] College of Business, Hunan University of Technology,
Zhuzhou 412007, China
415904214@qq.com
[4] School of Computer Science and Educational Software,
Guangzhou University, Guangzhou 510006, China
shuhongchen@gzhu.edu.cn
[5] School of Computer and Communication, Hunan Institute of Engineering,
Xiangtan 411104, China

Abstract. The Internet has become one of the important channels for users to access to information and knowledge. It is crucial that how to acquire key content accurately and effectively in the events from huge amount of network information. This paper proposes an algorithm for automatic generation of network document summaries based on knowledge graph and TextRank algorithm which can solve the problem of information overload and resource trek effectively. We run the system in the field of big data application in packaging engineering. The experimental results show that the proposed method KG-TextRank extracts network document summaries more accurately, and automatically generates more readable and coherent natural language text. Therefore, it can help people access information and knowledge more effectively.

Keywords: Knowledge graph · Automatic summarization · Automatic annotation · Network document

1 Introduction

In an open network environment, user need to rely on the network in the process of learning, work and life is also growing. The Internet has become an important platform for users to publish and obtain information in which network document resources are included. The information resources and the data generated by its application are growing in a geometric progression, making the "information overload" and "resource trek" problems more and more serious, which has seriously affected users efficiency in

S. Liu and G. Yang (Eds.): ADHIP 2018, LNICST 279, pp. 20–27, 2019.
https://doi.org/10.1007/978-3-030-19086-6_3

information acquisition [1]. Massive and excessive information are presented at the same time, making it impossible for users to easily obtain the resources they actually need. Automatic summarization is an effective mean to solve comprehensive and concise summarization from a large amount of text information. It has a great significance to improve the efficiency for user to obtain information.

However, the traditional extractive automatic summarization directly composes summarization by extracting key sentences in the original text, which, in turn, causes problems like incomplete coverage of the main content, lacking of contextual fluency, and duplication of synonymous information. The existing TextRank algorithm [2] is a graph-based sorting classical algorithm for text, which is mainly used in the fields of keyword extraction and automatic summarization. Knowledge graph [3] can better enrich and represent the semantics of resources and provide more comprehensive summary and more relevant information. Therefore, to deal with these problems, this paper improves the single document summary generation technology based on the extractive method, and proposes the improved extraction algorithm based on knowledge graph and TextRank to enhance the accuracy of summary generation, so that users can quickly and accurately find the network document resources they are interested in.

The remainder of the paper covers background and related work discussions (Sect. 2), the model of automatic summarization generation and detailed illustration (Sect. 3), the experiment and test results (Sect. 4), and the conclusions and future work (Sect. 5).

2 Related Research

Research on natural language generation enables computer to have the same function of expression and writing like human. That is, according to some key information and its expression in the machine, through a planning process, a high-quality natural language text is automatically generated. It is a branch of the field of natural language processing, involving multiple disciplines such as artificial intelligence, computer language processing, cognitive science and human-computer interaction. Currently, the general natural language generation system is a pipelined pipeline architecture, including 3 stages such as content planning, micro-planning and surface generation. In the field of current natural language generation systems, automatic summarization is a hot topic. According to the generation principle, it can be divided into extractive and abstractive; according to the number of input documents, it can be divided into single document summarization and multi-document summarization. Automatic summarization have variety of ways such as statistics-based, graph model-based, latent semantics-based, and integer programming-based currently. Gambhir et al. [4] gave a survey on recent automatic text summarization techniques. Lynn et al. [5] proposed an improved extractive text summarization method for documents by enhancing the conventional lexical chain method to produce better relevant information of the text using three distinct features or characteristics of keyword in a text. To alleviate incoherent summaries and same pronominal coreferences, Antunes et al. [6] proposed a method that solved unbound pronominal anaphoric expressions, automatically enabling the cohesiveness of the extractive summaries. Fang et al. [7] proposed a novel word-sentence co-ranking model named CoRank, which combined the word-sentence relationship with the graph-based unsupervised ranking model. CoRank is quite concise in the view of matrix operations, and its convergence can be theoretically guaranteed.

TextRank algorithm is widely used in automatic document summarization. Blanco et al. [8] constructs an unauthorised TextRank network map based on the co-occurrence information of terms in a certain window, which is applied to information retrieval. Combined with the structural features of Chinese text, Yu et al. [9] proposed an improved iTextRank algorithm, which introduced information such as title, paragraph, special sentence, sentence position and length into the construction of TextRank network graph, gave improved sentence similarity calculation method and weight adjustment factor, and applied to automatic summarization extraction of Chinese text.

The application of knowledge graph is coherently born to enrich and represent the semantics of resources. It was proposed by Google in 2012 to describe the various entities or concepts that exist in the real world and incidence relation between them. Knowledge graph is not a substitute for ontology. Ontology describes data schema of the knowledge graph, namely for knowledge graph building data schema equivalent to establishing its ontology. Knowledge graph basing on ontology enriches and expands, and the expansion is mainly embodied in the entity level. The knowledge graph is more accurate to describe the incidence of various relationships in the real world. The knowledge graph is a great promoter of the semantic annotation of digital resources, and promoting the efficient acquisition of knowledge and information. At present, Google, Sogou cubic, Baidu bosom, Microsoft Probase, etc. already preliminarily applied knowledge graph system in the industry. Most of them are general knowledge graph, which emphasizes the breadth of knowledge, and includes more entities. It is difficult to have complete and global ontology layer to unified management, and mainly used in the search business, and not high accuracy requirements. There are some industry knowledge graph, has high accuracy requirements, used for auxiliary complex decision support, the rich and the strict data patterns, etc. Liu, Xu et al. [10, 11] reviewed knowledge graph technology in academia. Hu [12] researched on the construction of knowledge graph based on the application. Li et al. [13] proposed an automatic knowledge graph establishment method and established a knowledge graph of packaging industry. In order to seek semantics support for searching, understanding, analyzing, and mining, Wu et al. [14] proposed a more convenient way which based on domain knowledge graph to annotate network document automatically. The method firstly adopts an upgraded TF-IDF model based on the contribution to quantify instances in knowledge graph, then analyzes the semantic similarity between unannotated documents and instances based on Jaccard distance and lexicographic tree distance comprehensively.

Aiming at the complex relationship of network documents, an automatic summarization extraction method based on knowledge graph is proposed. The entity relationship structure in knowledge graph can reflect the objective knowledge of event development. Applying its semantic characteristics to summarization generation technology can make the summarization more concise and comprehensive.

3 The Model of Automatic Summarization Generation

This paper proposes an automatic summarization generation technology based on knowledge graph and TextRank algorithm named KG-TextRank. Through the knowledge graph, the new meaning of the string is given, and the knowledge system related to the keyword is systematically made, so that the summarization generated by KG-TextRank is more consistent and comprehensive.

3.1 TextRank Algorithm

The classic TextRank algorithm is a graph sorting algorithm, which divides the text into several units, constructs a graph model for the nodes, and uses the voting mechanism to sort the important components in the text.

Let $G = (V, E)$ be a graph structure composed of text units, V is a fixed point set, and E is a edge set. $WS(V_i)$ is the score of the vertex V_i, and the iteration formula is:

$$WS(V_i) = (1 - d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \tag{1}$$

Where d is the damping coefficient, generally as 0.85; $In(V_i)$ is the set of all nodes pointing to node V_i; $Out(V_j)$ is the set of all nodes pointed to by node V_j; w_{ji} is the weight of the edge of node V_j to node V_i.

3.2 Industry Knowledge Graph Construction

The framework of industry knowledge graph construction method is shown in Fig. 1. It includes the lifecycle of domain knowledge graph, which mainly has five processes, namely, ontology definition, knowledge extraction, knowledge fusion, knowledge storage and knowledge application respectively. Each process has own methods and tasks. For example, D2RQ is used to transform the atomic entity table and the atomic relation table into RDF in knowledge extraction; defined the knowledge fusion rules to complete the knowledge fusion task while extracting knowledge with D2R and Wrappers, the tasks are such as entity merge, entity linking and attribute merge.

In this paper, based on the literature [13, 14], the authors obtain the semantic annotation knowledge graph. The semantic annotation helps the generation of sentence text and eliminates the ambiguity and ambiguity of natural language text. The entity in the knowledge graph can be used as a word segmentation dictionary. The semantics of entities, attributes and relationships provide synonymy, inclusion, etc., and remove ambiguity and ambiguity, thus provide standard, concise and comprehensive knowledge information.

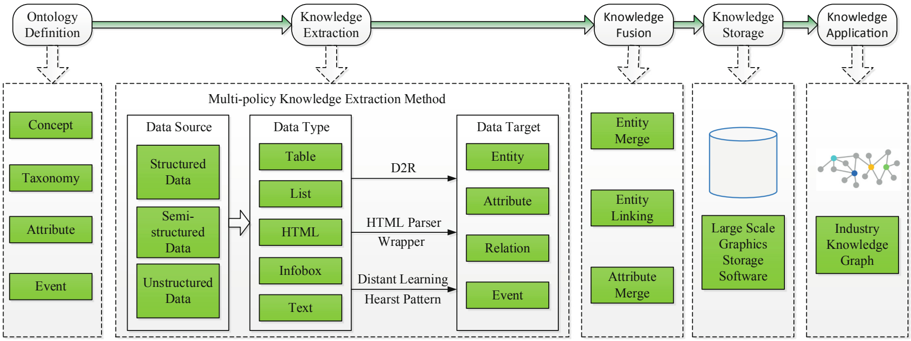

Fig. 1. The framework of industry knowledge graph construction method.

3.3 The Description of Implementation Algorithm

The authors improve the single document summary generation technology based on the extractive method named TextRank, and propose the improved extraction algorithm based on knowledge graph. The implementation flow chart of improved algorithm is as shown in Fig. 2.

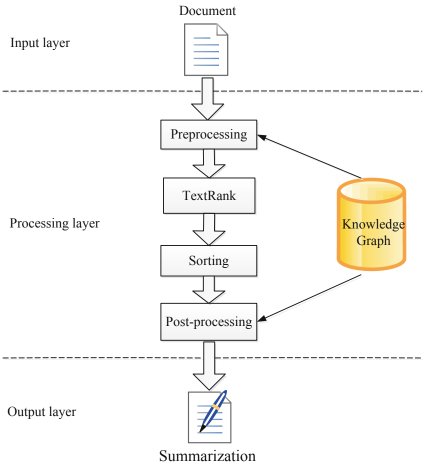

Fig. 2. The flow chart of implementation algorithm.

The computing process is as follows:

Firstly: Processing the input document and parsing into text T; then performing the word segmentation by entity dictionary of the industry knowledge graph and the general dictionary, and constructing the feature vector matrix of the text T; then extracting the keyword and computing the correlation degree of sentences.

Secondly: Constructing a TextRank network diagram and performing iterative calculation using TextRank algorithm.
Thirdly: Sorting and Selecting the top t sentences.
Finally: In the post-processing stage, simplifying and supplementing sentences based on industry knowledge graph, then outputting a concise and comprehensive summarization A_a according to the order of the original text.

4 Experiment and Evaluation

4.1 Constructing Packaging Knowledge Graph

We construct packaging knowledge graph, which is as shown in Fig. 3. For example, the knowledge graph includes the following basic concepts, namely, “packaging knowledge point”, “Company”, “Product”, “Organization”, “Patent”, “Paper” and “Event”. Major relations include “has product”, “upstream”, “downstream”, “has patent”, and “executive”.

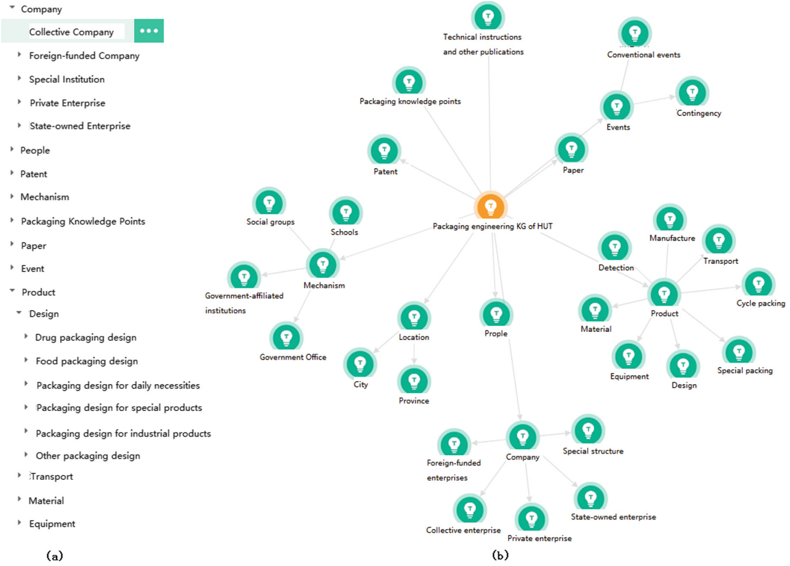

Fig. 3. Packaging knowledge graph.

4.2 Algorithm Evaluation

In this paper, what is adopted is an evaluation method that using artificial abstract as a standard to calculate the precision rate and recall rate. Set up artificial abstract sentences sets as $A_m = \{S_{m1}, S_{m2}, \ldots, S_{mt}\}$, automatic summarization sentences sets as $A_a = \{S_{a1}, S_{a2}, \ldots, S_{at}\}$, where t is the sentence number of generating summarization.

The Precision Rate calculation formula is as follows:

$$P = \frac{|A_m \cap A_a|}{|A_a|} \tag{2}$$

The Recall Rate calculation formula is as follows:

$$R = \frac{|A_m \cap A_a|}{|A_m|} \tag{3}$$

The F-measure Value calculation formula is as follows:

$$F_1 = \frac{2 \times R \times P}{R + P} \tag{4}$$

Our experiment material are from own development system named China packaging industry large data knowledge graph system. We pick up a part of the academic papers from the system for the summarization generation, and choose number of each paper summarization within the four sentences, and process comparative analysis through computing precision rate (P), recall rate (R) and F- measure value. The Experimental Results are as shown in Table 1.

Table 1. Experimental results.

Algorithm	Precision rate	Recall rate	F_1 value
TextRank algorithm	0.467	0.333	0.389
KG-TextRank algorithm	0.554	0.448	0.495

From the results of Table 1, we can see that the improved KG-TextRank algorithm has higher precision rate, recall rate and F-measure value than the traditional TextRank algorithm, indicating that the use of domain knowledge graph helps to automatically generate summarization closer to manual summarization. At the same time, it was found that these values decreased with the increase of summarization sentences.

5 Conclusions

How to make sentences to extract more in line with the thinking of artificial screening in the field of automatic summarization research is research hot spot and focus. In this paper, based on the classic TextRank algorithm, the authors, by joining the industry knowledge graph, provide a technical implementation scheme for the automatic summarization. The experimental results show that the improved algorithm proposed in this paper improves the quality with a certain degree in the generated summarization.

The next step is to apply deep learning to learn and text eigenvector, researches on abstract summarization generation, experiments in packaging large-scale corpus, and constructs a complete packaging large data summarization system.

Acknowledgements. This work is supported in part by the National Natural Science Foundation of China under grant number 61502163, in part by the Hunan Provincial Natural Science Foundation of China under grant numbers 2016JJ5035, 2016JJ3051 and 2015JJ3046, in part by the Hunan Provincial Science and Technology Project under grant number 2015TP1003, in part by the Project of China Packaging Federation under Funding Support Numbers 2014GSZJWT 001KT010, 17ZBLWT001KT010, in part by the National Packaging Advertising Research Base and Hunan Packaging Advertising Creative Base under grant number 17JDXMA03, in part by the Intelligent Information Perception and Processing Technology Hunan Province Key Laboratory under grant number 2017KF07.

References

1. Wu, Y., Liu, Q., Li, C., Wang, G.: Research on cloud storage based network document sharing. J. Chin. Comput. Syst. **36**(1), 95–99 (2015)
2. Mihalcea, R., Tarau, P.: TextRank: bringing order into texts. In: Proceedings of EMNLP 2004, pp. 404–411. ACM, Barcelona (2004)
3. Amit, S.: Introducing the Knowledge Graph. Official Blog of Google, America (2012)
4. Gambhir, M., Gupta, V.: Recent automatic text summarization techniques: a survey. Artif. Intell. Rev. **47**(1), 1–66 (2016)
5. Lynn, H.M., Chang, C., Kim, P.: An improved method of automatic text summarization for web contents using lexical chain with semantic-related terms. Soft Comput. **22**(12), 4013–4023 (2018)
6. Antunes, J., Lins, R.D., Lima, R., Oliveira, H., Riss, M., Simske, S.J.: Automatic cohesive summarization with pronominal anaphora resolution. Comput. Speech Lang. (2018). https://doi.org/10.1016/j.csl.2018.05.004
7. Fang, C., Mu, D., Deng, Z., Wu, Z.: Word-sentence co-ranking for automatic extractive text summarization. Exp. Syst. Appl. Int. J. **72**(C), 189–195 (2017)
8. Blanco, R., Lioma, C.: Graph-based term weighting for information retrieval. Inf. Retrieval **15**(20), 54–92 (2012)
9. Yu, S., Su, J., Li, P.: Improved TextRank-based method for automatic summarization. Comput. Sci. **43**(6), 240–247 (2016)
10. Liu, Q., Li, Y., Duan, H., Liu, Y., Qin, Z.G.: Knowledge graph construction techniques. J. Comput. Res. Dev. **53**, 582–600 (2016)
11. Xu, Z.L., Sheng, Y.P., He, L.R., Wang, Y.F.: Review on knowledge graph techniques. J. Univ. Electron. Sci. Technol. China **45**, 589–606 (2016)
12. Hu, F.H.: Chinese knowledge graph construction method based on multiple data sources. East China University of Science and Technology, Shanghai (2014)
13. Li, C., Wu, Y., Hu, F.: Establishment of packaging knowledge graph based on multiple data sources. Revista de la Facultad de Ingeniería **32(**14), 231–236 (2017)
14. Wu, Y., Wang, Z., Chen, S., Wang, G., Li, C.: Automatically semantic annotation of network document based on domain knowledge graph. In: 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications, pp. 715–721 (2017)