Hotspot Detection and Trend Analysis in the Field of Power Measurement and Data Acquisition

Xiong Li
State Grid Zhejiang Electric Power Research Institute
Hangzhou, China

Jiahong Jin
State Grid Shaoxing Power Supply Company
Shaoxing, China

Shaocheng Liao
State Grid Zhejiang Xinchang Power Supply Co.,Ltd.
Shaoxing, China

Min Yu
Zhejiang Huayun Information Technology Co.,Ltd.
Hangzhou, China

Lei Han
Zhejiang Huayun Information Technology Co.,Ltd.
Hangzhou, China

Ziyan Chen
College of ISEE, Zhejiang University
Hangzhou, China

***Abstract*—In this paper, we study the research hotspot detection problem in the field of power measurement and data acquisition by using a big data and naural language processing based method. We make a combination of the TF-IDF-based and the TextRank-based keyword extraction techniques for preliminary hotspot detection, and further uses the word2vec model to merge synonyms to get more accurate detection result. By applying this method to literature in the field of power measurement and data acquisition, we obtain the annual research hotspots of this field for the past decade. Furthermore, the evolution trend analysis of hotspots is performed. The results can provide reference for practitioners in the field.**

Keywords-power measurement and data acquisition; research hotspot; evolution trend; keyword and keyphrase extraction; synonym identification

I. INTRODUCTION

Power measurement and data acquisition is an important part of the operation of electricity market. Its accuracy and efficiency directly affect the fairness of power settlement and are directly related to the economic interests of power suppliers and consumers [1]. Therefore, how to establish a complete and scientific management system to improve the accuracy and efficiency of power measurement and data acquisition remains an important issue for power companies. Intuitively, comprehending the research hotspots and corresponding evolution trends in this field helps power companies to make correct plans, which can be achieved by manually reading and summarizing relevant literature. However, with the flourish development of the power industry, the large amount of research literature accumulated makes this process not only time consuming, but also labor and resources demanding. Besides, too much human input means too much subjectivity, which can easily lead to irregular processing and large deviation of the result. Therefore, it is of great significance to develop a system for automatic research hotspot detection and trend analysis in the field of power measurement and data acquisition.

To this end, in this paper, we study the research hotspot detection problem in the field of power measurement and data acquisition by using a big data and naural language processing based method, and design the corresponding automatic literature processing system. The method we adopt here is basically a combination of the TF-IDF-based [2-3] and TextRank-based [4-5] hotspot detection techniques. What's more, considering the case that different words represent the same hotspot, we adopt the word2vec model [6-8] to merge the synonyms to further improve the accuracy. Similar methods have been used in other fields [9-10], but to the best of our knowledge, there is no such research in this field. By applying this method to the documents acquired by distributed crawlers, the system implements fully automatic hotspot detection in the field of power measurement and data acquisition. Based on that, the evolution trend analysis of hotspots is performed. The results can provide reference for practitioners in the field.

The rest of this paper is organized as follows. In Section Ⅱ, we introduce the popular TF-IDF-based and TextRank-based keyword extraction techniques, and the combination method. In Section Ⅲ, we introduce the synonym identification method based on word2vec model. The implementation details of the entire system are introduced in Section Ⅳ. In Section Ⅴ, we present the result of research hotspot dection in the field of power measurement and data acquisition. Also, the evolution trend analysis is performed in this section. Finally, conclusions are drawn in Section Ⅵ.

II. KEYWORD EXTRACTION

Research hotspot refers to research topic that frequently appear in the literature over a period of time. In order to get research hotspots, we first need to know the research topics

of each document, which can usually be represented by the "keywords" given by the authors. Then, the frequency of occurrence of these keywords is counted, and the keywords with the highest frequency are selected as the research hotspots. However, these artificially selected keywords are highly subjective and may not fully represent the research topics. In order to get more representative keywords, in this paper, we use a combination of two popular keyword extraction methods, the TF-IDF-based method and TextRank-based method, to automatically and objectively extract keywords from each document.

A. TF-IDF-based method

The TF-IDF-based method calculates the importance of each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in [11]. Specifically, given a document collection C, a word w_i and an individual document $c \in C$, the importance of word w_i in document c is calculated by

$$W_{\text{TF-IDF}}(w_i) = \frac{F_c(w_i)}{N_c} \log\left(\frac{|C|}{|C(w_i)|}\right), \tag{1}$$

where $F_c(w_i)$ equals the number of times w_i appears in c, N_c is the total number of words in c, and $|C(w_i)|$ is the number of documents in C in which w_i appears. Then the words with the highest importance are selected as the keywords of a document.

B. TextRank-based method

TextRank is a graph-based ranking algorithm which recursively determines the importance of vertices in a directed graph based on global information. By mapping a document into a graph model, where each word is a vertex in this graph, and then using a voting mechanism to sort the importance of vertex, the Textrank-based method realizes keyword extraction using only the information of the single document itself.

Specifically, given an document, this TextRank-based method first constructs a directed graph model for this document according to certain rules. For the construction details, we refer the readers to [5,12]. Then the importance of a word w_i is calculated recursively by

$$W_{\text{TR}}(w_i) = (1\text{-d}) + \text{d} \times \sum_{w_j \in N(w_i)} \frac{1}{|N(w_j)|} W_{\text{TR}}(w_j), \tag{2}$$

where $N(w_i)$ is the collection of words connected to w_i, and d is a damping factor that can be set between 0 and 1, which allows random transitions from one vertex to another.Similarly, after the algorithm converges, the words with the highest importance are selectedas the keywords.

C. The combination method

In this step, the importance weights obtained earlier are combined to generate a single score, which is then used to make the final decision. Since the importance calculated by various methods usually has different distribution statistics, it is necessary to normalize the weights before combination. Here we use the min-max normalization procedure to fit the data into the range [0, 1]

$$W'_{\text{M}}(w_i) = \frac{W_{\text{M}}(w_i) - \min W_{\text{M}}(w_j)}{\max W_{\text{M}}(w_j) - \min W_{\text{M}}(w_j)}, w_j \in c, \tag{3}$$

where M refers to TF-IDF or TR. Then the final weight is calculated by

$$W(w_i) = \alpha W'_{\text{TF-IDF}}(w_i) + (1-\alpha) W'_{\text{TR}}(w_i), \tag{4}$$

where $\alpha \in (0,1)$ is a weighting factor.

After obtaining the importance ranking of each document, the top five words or phrases are selected as the keywords for the document. Then preliminary research hotspot detection result is obtianed by counting the frequency of these keywords.

III. SYNONYM IDENTIFICATION

Another issue to be considered is synonym identification. In the scientific literature, the same concept can be represented by different words or phrases, and different authors may have different preferences in word selection. For example, data acquiring and data collecting actually represent the same process, and it is unreasonable to calculate their heat seperately. Therefore, it is necessary to synonymize the results of keyword extraction to obtain more reliable hotspots. While synonyms of literal similarity can be easily identified using Levenshtein distance [13], this method has a poor identification result for synonyms of semantic similarity. To this end, we adopt the word2vec model to realize the identificationof synonyms on the semantic level.

Word2vec is a word vector generation model which trains a neural network to obtain the distributed representation of words. Different from the traditional one-hot representation [14] which maps words into high-dimentional and sparse vectors, thedistributed representation maps them into a low-dimensional and dense feature space where words with similar semantics have similar vector representations. Then the semantic similarity bewteen different words is measured by the distance bewteen the corresponding vectors.

Generally, according to the different methods of training neural networks, word2vec can be divided into two models, the CBOW model and the Skip-Gram model [15], as shown in Fig.1. The former trains the neural network by using the words around a certain word to predict this single word, while the latter uses a certain word to predict the words around it. In general, CBOW model is suitable for small corpora, while Skip-Gram model performs better in large corpora. Therefore, for the research hotspot detection problem, the CBOW model is chosen considering the relatively small corpus size.

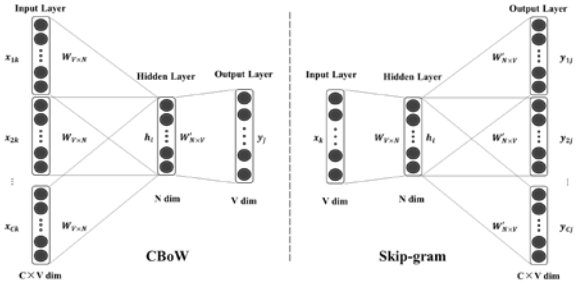

Fig. 1. CBOW model and Skip-Gram model

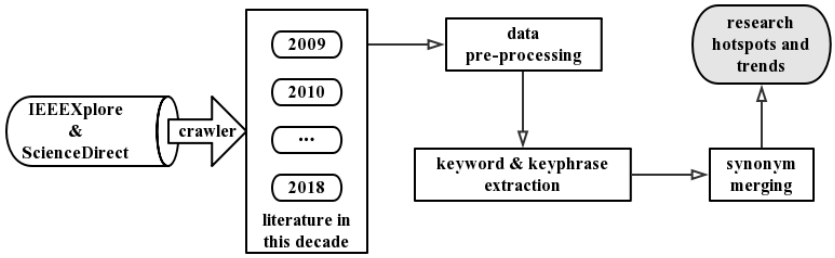

Fig. 2. Overview of system architecture

Based on the obtained synonym model, the recalculation of the hotspot detection result is performed by adding the heat of the relatively low-frequency word in a pair of synonyms to the relatively high-frequency word.

IV. System Implementation

Based on the keyword extraction method and synonym identification method introduced above, we develop the corresponding literature analysis system to automatically and efficiently explore the research hotspots and development trends in the field of power measurement and data acquisition. The overall architecture of the system is shown in Fig.2.

In the corpus acquisition module, the system takes IEEEXplore and ScienceDirect as data sources and crawls all academic papers and patents published in the filed of power measurement and data acquisition since year 2009.Considering that a complete academic paper or patent is relatively long and might contains too much useless and confusing information, which can make the process time-consuming and lead to poor results, in this system, the corpus contains only the abstracts of papers or patents, which highly summarizes the content of the full text.

Based on the corpus obtained, the system first performs data preprocessing on the text data., which includes stemming, lemmatization and removal of stop words [16]. Then for the keyword and keyphrase extraction module, the combination method introduced in Section Ⅱ is used, where we set the damping factor d to be 0.85, and set the weighting factor α to be 0.5. Next for the synonym identification module, the CBOW model is used to train the language neural network for word-to-vector mapping.

Finally, the research hotspot detection and the evolution trend analysis are performed based on the results of previous modules. Taking year as time interval, the top 10 words or phrases are selected as the research hotspots of that year.

V. Results

In this section, we present the results of research hotspot detection and trend analysis in the filed of power measurement and data acquisition.

The temporal variation of the number of academic literature in the filed of power measurement and data acquisition is shown in Fig.3. From Fig.3 we can find that the quantity of literature in this field rises steadily during the past decade, which indicates the rapid development of this field and demonstratesthe necessity of the work in this paper.

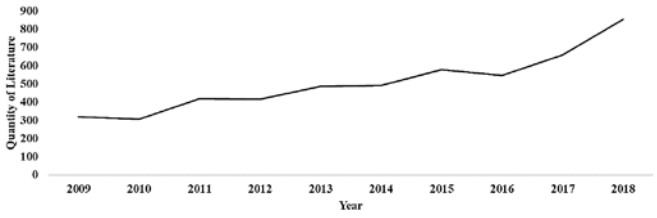

Fig. 3. Temporal variation of the number of literature in this decade

The annual research hotspots detected by the system from year 2009 to 2018 are listed in Table 1. Through statistical analysis of the hotspot detection result, we obtain the following several results.

As shown in Fig.4, during the past decade, the frequency of the keywords “smart grid”, “fault diagnosis”, “power quality”, “phasor measurement unit” and “state estimation” has remained at the top ten almost every year, and the

corresponding rankings are relatively high, which suggests that research on these topics has always been the focus of the field for the past decade. We expect that the research heat of these topics will continue forseveral years in the future.

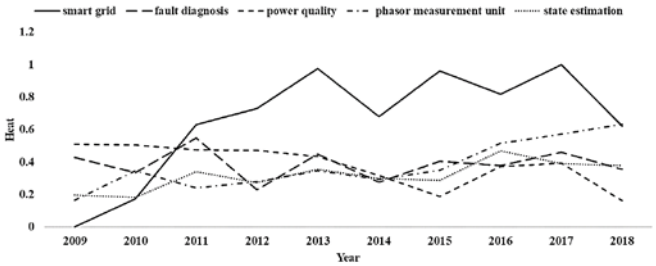

Fig. 4. Heat trends of the most common research hotspots in this decade

On the other hand, as shown in Fig.5, the frequency of the keywords “fault detection”, “microgrid”, “demand response”, “electric vehicle” and “internet of things” has shown a clear upward trend in the past decade. Among them the keyword “Internet of Things” has risen rapidly after its first appearance in 2013, and for the first time became the annual hotspot in the year 2018. We expect these research topics to become more mainstream in the future, thus we suggest power companies and practitioners pay attention to the development of related topics.

TABLE I. RESEARCH HOTSPOTS IN THIS DECADE

Year	Research Hotspots
2009	power quality; signal processing; fault diagnosis; wireless sensor networks; induction motors; neural networks; analog-to-digital converter; field programmable gate arrays; real time; spectral analysis
2010	power quality; real-time; phasor measurement unit; fault diagnosis; neural networks; kalman filter; wireless sensor networks; frequency estimation; parameter estimation; image processing
2011	smart grid; fault diagnosis; power quality; state estimation; wireless sensor networks; field programmable gate arrays; induction motors; distributed generation; signal processing; condition monitoring
2012	smart grid; power quality; neural networks; induction motors; transmission lines; phasor measurement unit; state estimation; kalman filter; distributed generation; signal processing
2013	smart grid; fault diagnosis; power quality; state estimation; phasor measurement unit; distribution system; kalman filter; neural networks; real-time; wireless sensor networks
2014	smart grid; neural networks; power quality; wireless sensor networks; state estimation; phasor measurement unit; fault diagnosis; demand response; induction motors; energy management
2015	smart grid; fault diagnosis; microgrid; phasor measurement unit; state estimation; condition monitoring; energy management; demand response; distribution system; distributed generation
2016	smart grid; phasor measurement unit; state estimation; induction motors; fault diagnosis; power quality; demand response; fault detection; real-time; energy storage
2017	smart grid; phasor measurement unit; fault diagnosis; fault detection; induction motors; power quality; state estimation; microgrid; condition monitoring; permanent magnet
2018	phasor measurement unit; smart grid; state estimation; fault diagnosis; neural networks; distribution system; fault detection; internet of things; microgrid; electric vehicle

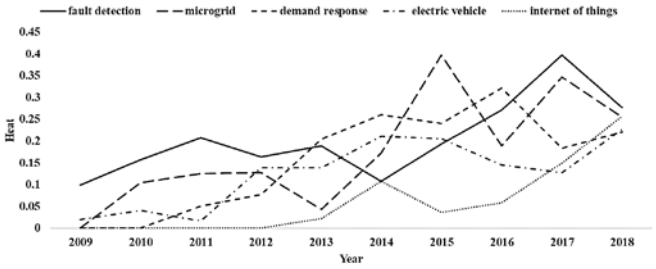

Fig. 5. Heat trends of the rising research hotspots in this decade

VI. CONCLUSION

In this paper, we studied and developed a research hotspot discovery system for the field of power measurement and data acquisition. By applying a hybrid keyword extraction technique and a word2vec-based synonyms identification technique to the literature in this field, we obtained the annual research hotspots for the past decade. On this basis, we analyzed the evolution trend of research hotspots. The results can provide reference for practitioners in the field.

REFERENCES

[1] M. Kutz, Handbook of measurement in science and engineering. Wiley, 2013.

[2] G. Salton, C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information processing & management, 1988, 24(5): 513-523.

[3] Z. Zhu, J. Liang, D. Li, H. Yu, and G. Liu, “Hot Topic Detection Based on a Refined TF-IDF Algorithm,” IEEE Access, 2019, 7: 26996-27007.

[4] R. Mihalcea, P. Tarau, “Textrank: Bringing order into text,” Proceedings of the 2004 conference on empirical methods in natural language processing. 2004: 404-411.

[5] C. Mallick, A. K. Das, M. Dutta, and A. Sarkar, Graph-Based Text Summarization Using Modified TextRank. Soft Computing in Data Analytics. Springer, Singapore, 2019: 137-146.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

[7] Q. Le, T. Mikolov, “Distributed representations of sentences and documents,” International conference on machine learning. 2014: 1188-1196.

[8] Y. Goldberg, O. Levy, “Word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method,” arXiv preprint arXiv:1402.3722, 2014.

[9] Y. Wen, H. Yuan, and P. Zhang, “Research on keyword extraction based on word2vec weighted textrank,” 2016 2nd IEEE International Conference on Computer and Communications (ICCC). IEEE, 2016: 2109-2113.

[10] K. Hu, H. Wu, K. Qi, J. Yu, S. Yang, T. Yu, J. Zheng, and B. Liu, “A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model,” Scientometrics, 2018, 114(3): 1031-1068.

[11] J. Ramos, “Using tf-idf to determine word relevance in document queries,” Proceedings of the first instructional conference on machine learning. 2003, 242: 133-142.

[12] A. N. Langville, C. D. Meyer, Google's PageRank and beyond: The science of search engine rankings. Princeton University Press, 2011.

[13] G. Navarro, “A guided tour to approximate string matching,” ACM computing surveys (CSUR), 2001, 33(1): 31-88.

[14] D. Harris, S. Harris, Digital design and computer architecture. Morgan Kaufmann, 2010.

[15] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” Advances in neural information processing systems. 2013: 3111-3119.

[16] C. Fox, “A stop list for general text,” Acm sigir forum. ACM, 1989, 24(1-2): 19-21.