Improvement of Knowledge Search Method for Numeral Terms in Power System Based on Multi-source Fusion

Min Hua*, Xiao-man Qi¹, Ai-qiang Pan¹, San-shan Zhao¹, Xiao-xia Huang², Chao Wu²

¹Shanghai Municipal Electric Power Research Institute, Shanghai 200437, China
²Shanghai Maritime University, Shanghai 201306, China

*Corresponding author’s e-mail: hmhzgb@163.com

Abstract: In the process of energy and power research, it is often necessary to find numeral data information from multiple ways based on searching words, so the domain information can better analyzed and understood, visually discover hidden relationships between information and ultimately improve search efficiency and experience. The article designs a set of numeral entity discovery methods based on NLP and improves a visual search engine. Using CRF for named entity pre-processing extraction, using TF-IDF, etc. to complete multi-source information combination search and information mining method establishment, the power system energy database is used as the information source of the search engine, enabling users to locate the document data of the searching words in the energy database and the context information associated with them quickly, generate summary information, extract the central words based on this, and image the information in the form of word cloud and chart demonstration. Experiments have shown that the searching result are of interest to users, quickly locate the literature and context-related information containing the numeral hotspots, and the search recall rate is significantly improved.

1. Introduction
In the era of big data Internet with information explosion, there is a lot of information about power & energy. Digital energy data scattered throughout the historical literature. It is often difficult to find the required information for a while. Therefore, how to extract numeral and related knowledge from documents is a big challenge. This paper based on the power system information to build a natural language recognition model, combines the needs of the power system to search numeral entities and contexts for high-speed retrieval, and builds a multi-source heterogeneous knowledge map graphical display method to quickly present the digital information contained in the file.

At present, domestic and foreign knowledge intelligent search engines, especially the Internet industry, some units extract entities, relationships, and ontology from data to create a universal knowledge database, and have formed many large-scale knowledge bases, such as DBpedia[1], Yago, Freebase[2], BabelNet, ResearchCyc, WordNet, ConceptNet, KnowItAll, Microsoft Probase[3], Google Knowledge Graph[4], Fudan GDM, XLORE, zhishi.Me, Knowledge Cube (SSCO), etc. The intelligent search engine is divided into two types: "top-down" and "bottom-up"[5]. It is divided into full-text search engines, classified index search engines, and meta-search engines from the structure and formation effect[6]. Current knowledge association systems often use RDF as their underlying layer to implement the application of linked data technology[7], and to achieve the connection between various databases.
2. System

The system mainly provides 4 major capabilities:

The first is the search function, which can quickly locate the literature where the search term is located and display it in the form of a list. The second is the topic extraction function, which can locate the context based on keywords and extract related topic words from them. The third is the relevance calculation function, which can calculate the relevance between the extracted subject words and search words. The fourth is the word cloud display function, which can display the hot words that need to be queried in the form of a word cloud, and list the relevant information and correlation coefficients.

Search result feedback based on the source data, use NER to extract from documents, decompose the text containing numeral entity features, and establish the keyword, use TFIDF algorithm to predict the central word, forming aNER navigation (Navigation bar form based) and central word prediction double-layer index. The user searches for the target digital content, needs to enter a search term, and clicks the central word prediction prompt in the list. The system establishes the event response of the word-oriented graph after the central word is clicked, and automatically navigates to the corresponding digital chart. The system displays the results in the form of a fusion of lists and digital knowledge, combining word-oriented maps, quick result lists, and digital data charts in the form of components to quickly display them. Choose whether to load components based on the presence or absence of relevant content and asynchronously shows the top 100 most relevant messages. Digital data charts provide correlation information mining for search results, displaying knowledge information around search terms. In terms of graphic association, the summary information of the word-oriented graph is used to calculate the key headwords using a variety of correlation algorithms. The graphic components sort the results according to the order of relevance.

2.1 Search structure design

2.1.1 NER navigation construction

Named entity labeling [9] Navigation is to preprocess data according to the principle of named entity labeling, establish data classification indexes, and filter the number types of search results. Dynamically establish the number category labels in the search results by judging the content of the search results, which are ‘Amount’, ‘Ranking’, ‘Percentage’, ‘Number’, and ‘All’. The four fields in the list: document name, sentence, number, number category, and part of speech give accurate information about the search results. The content of the list is obtained by loading JSON preprocessing and then searching. The preprocessing process has extracted the content marked by the numeral NER through the CRF method, and the numeral entity is added to the ‘numeral column. Data is exchanged and shared between lists and graphics via JSON.

For example, a text serialized in JSON format is as follows:

```json
{"docName": {"id": 1, "name": ""}, "docSentence": {"Word": "", "Pos": "CD","Ner": "NUMBER","Lemma": "", "Sentence": "", "id": 1}}
```

In the preprocessing, the entire sentence containing numbers, the number category of NER
annotation, and the part-of-speech annotation of numeric fields are retained.

### 2.1.2 Keyword prediction
Keyword prediction is to recommend related content of the entire search result for search terms, focusing on several central topics, and it is necessary to establish a central word prediction index list for the search results. List all related headings in related sentences (including search terms). The headings generate a summary based on TF-IDF [10], and a query (Query) establishes keyword summary information for each search word. The association headword is the highest related word in a 5-digit TF-IDF summary word array.

$$\text{TF} - \text{IDF}(w) = I(w) - \text{TF}(w) \cdot \log \frac{N}{\text{TF}(w)} \quad (1)$$

For example: "The reliability of urban and rural power supply can reach 99% and 98%, respectively", and the related head word in this phrase is "reliability". Its key word recommendation is "reliability, rural power network, reach, power supply, city". The user enters "Reliability", and the headword prediction list will be generated and shown, which is as follows.

| Center word | Summary |
|-------------|---------|
| Reliability | (Reliability Rate, Agricultural Network, Reach, Power Supply, City) |
| Power supply | (Power, Reach, Lift, More, Year) |
| Quality | (Quality, Energy, World, Shanghai, Brand) |
| Power grid | (Grid, Feature, World, Distribution Grid, Realization) |
| Brand | (Brand, Red, Heritage, Value, Research and Development) |
| Global | Global, Built, Basic, First, Guided |
| Engineering | (Engineering, Action, Lift, Promotion, Quality) |

### 2.1.3 Graphic association
Word-oriented graphs are an important part of the visualization engine and the core of digital knowledge association construction. It is organized and displayed in the form of numeral entity labeling and word-oriented diagrams. Based on the headword prediction clustering, a three-level structure-oriented structure of "headword-abstract-numbers and descriptions" is obtained, as shown in Figure 2. The headword, the keyword summary, and the phrase containing the digital knowledge result together constitute a three-layer digitally oriented information graph. Its construction object is $O = (E, R)$. Each node contains entities and relationships, and forms a three-layer structure $\{C, D, R\}$.

Figure 2. Numeric descriptions derived from the central word of "sold electricity"

### 2.2 Graphic knowledge fusion

#### 2.2.1 Relationship traceability and color annotation
In addition to effectively expressing the relationship between numeral entities in a document, graphic knowledge search also needs a clear index file and its location to express the relationship between
search terms and documents. Digital color labeling has a good effect on intuitive expression. Through the obvious color-differenced round-cornered rectangle, as shown in Figure 3. It indicates that different documents are used to construct a guide map to identify the data source in the document.

Figure 3. Color-coded documents containing "capacity" and "electricity" numeral content

The document color labeling designed in this article establishes a document and color unique corresponding display, establishes a dictionary Dict<Document, Color>, loops through the objects in the graph, and color-codes the differences in the document attributes, and its complexity is O (n). The first is to establish specific color presets for random selection. The second is to match the input search keywords. The digital objects of the word-directed graph network nodes are color-coded differently.

The regular expressions are used to quickly match the digital object display and highlight it in bold.

\[ \text{Var } rE = \emptyset; \text{//Initialize regular expression} \]
\[ \forall (sN \text{ in } dS) \text{ For all child nodes in the document wrapper collection,} \]
\[ \text{If } (rE.\text{IsMatch}(sN)) \text{If the regular expression matches the child node} \]
\[ \text{If } (sN.\text{Contains}(sText)) \text{If the node contains the search term} \]
\[ sN.\text{SetProperty}(sText, dC[hl]) \text{Highlight text settings in a node} \]
\[ \text{Else} \]
\[ sN.\text{SetProperty}(cText, dC[sN.\text{Incidies}]) \text{Set the node to the same color as the document} \]

Relevant traceability means that the same search content is displayed in different documents, and the relevance of digital information in the documents is visually presented. Different entity sources are distinguished by the color of the connector and the color of the headword node represented, and the reflection mechanism mapping is used as a conditional instruction to execute an external application process to browse external documents, as shown in Figure 3.

Figure 4. Relationship tracing
The word cloud-oriented graph uses the method of reference resolution \([11]\) to reduce redundant information and increase the amount of information in the result. In the same text, they often contain exactly the same digital data, although they are exactly the same digital objects, they point to the same meaning. For the simplest information representation of the digital knowledge graph, which needs to be resolved. The information fingerprint calculation method is introduced. Different FP (D) fingerprint calculation values exist in different documents, and can still be drawn correctly, so that the search results are all contained without omission.

2.2.2 Correlation calculation between parts

The process of graphic knowledge fusion uses the keyword abstract extraction ‘Desc,’ to calculate similarity. In the calculation, three similarity algorithms are used to calculate the relationship between the abstract word set of the word-oriented graph and the theme of the digital chart. Since the cosine angle calculation similarity\([12]\) only reflects the number of morpheme occurrences in the abstract, the longest common-principle sequence and SimHash\([13]\) can play a role in processing short information and long information, respectively, so the three different types are combined. The correlation describes the characteristics of the method to more accurately calculate the correlation between elements. Because all three methods can restrict the value of similarity to (0, 1). That is to say, the correlation is exactly the same, and the value obtained by calculating the product can directly reflect the similarity.

\[ \cos_{sim}(a, b) = \frac{\sum_{i=1}^{n} a_i \times b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \times \sqrt{\sum_{i=1}^{n} b_i^2}} \]  \quad (2)

SimHash presets the hash value to 64 bits and constrains the result between 0 and 1 through calculation.

\[ \text{hamming}(a, b) = \sum a_i \oplus b_i \]  \quad (3)

\[ \text{Hamming}_{sim}(a, b) = \frac{\text{hamming}(a, b)}{n} \]  \quad (4)

LCS is the length of the LongestCommonSubsequence, and it is solved by dynamic methods to reduce time complexity.

\[ \text{lcs}(a, b) = \left\{ \begin{array}{l}
0 \\
\max(c[a - 1, b - 1] + 1, c[a - 1, b], c[a, b - 1])
\end{array} \right. \]  \quad (5)

\[ \text{lcs}_{sim}(a, b) = \max\left(\frac{\text{lcs}(a, b)}{\max(\text{len}(a), \text{len}(b))}\right) \]  \quad (6)

Combining multiple similarities by weight can get:

\[ \text{sim}(A, B) = \sum_{k=1}^{n} \text{sim}(a, b) = \frac{f_1}{f_1 + f_2 + f_3} * \cos_{sim}(a, b) + \frac{f_2}{f_1 + f_2 + f_3} * \text{lcs}_{sim}(a, b) + \frac{f_3}{f_1 + f_2 + f_3} \]  \quad (7)

The Sigmoid function is used to optimize the weights, avoid some defects caused by manual selection of the weights, and output a good threshold result.

\[ f(x) = \frac{1}{1 + e^{-x}} \]  \quad (8)

\[ x \in (0, 1) \]  \quad x represents the similarity value obtained by each method, \( f(x) \) for the initial weight of each method, set \( f_1, f_2, f_3 \) respectively.

For example: enter the search term: power supply, corresponding to the production keyword summary vector\([4]\) [power supply, system, general, supervision, audit].

| Name | Summary | Cosine similarity | LCS Similarity | Simhash similarity |
|------|---------|------------------|----------------|-------------------|
| Comparison of coal consumption from the power supply in Shanghai and some | (Fai, Power, Provincial and Municipal, Coal Consumption, | 0.8276 | 0.2 | 0.9375 |
provinces and cities in 2015
Comparison of coal consumption for electricity supply in Shanghai in 2018 with some provinces and cities
Coal consumption for electricity supply in Shanghai and some provinces and cities in 2018
Coal consumption for power supply in Shanghai and some provinces and cities in 2018
Coal consumption and power supply coal consumption in Shanghai
2018 Coal consumption and coal consumption for power generation in Shanghai
Reliability rate of power supply in Shanghai Power Grid City in 2018
Reliability of power supply in Shanghai power grid
Shanghai is available for electricity
Shanghai capacity in 2018

3. Test Evaluation
This system was tested on the company's system. The experimental data totaled 680MB. The data includes world energy data, comprehensive regional energy data, and Shanghai energy data and economic data for a total of 1452 items and 191 data tables, including consulting reports, department reports, learning materials, reference materials and other data, including external energy information, energy data, statistical yearbook and other original information, about 430 MB. The NER word segmentation system performs real-time conversion of energy files in the pre-processing establishment thread. In order to use criteria to evaluate the listing process of search engines, it is judged as to whether the results meet expectations. The accuracy rate $P_L$ and recall rate $R_L$ are used to evaluate the process of search. L represents a set of 100 search results of standard search results.

Table 3. search results

| NER category | What to search for | Average accuracy | Average recall rate |
|--------------|-------------------|-----------------|--------------------|
| Ranking      | Reliability       | 95.23           | 70.64              |
|              | Capacity          | 90.34           | 76                 |
| Amount       | Reliability       | 92.18           | 70.09              |
|              | Capacity          | 95.7            | 74.06              |
| Percentage   | Reliability       | 93.15           | 70.6               |
|              | Capacity          | 94.6            | 73.45              |
| Digital      | Reliability       | 90.32           | 70.62              |
|              | Capacity          | 93.55           | 74.79              |
Figure 5. NER navigation results

After applying NER navigation, input common search terms in the power system, compared with the first 100 records of traditional full-text search, the recall rate has increased significantly.

Table 4. P, R, F values of search terms

| Search terms     | Accuracy | Recall rate | F    |
|------------------|----------|-------------|------|
| "Reliability"   | 0.8      | 0.32        | 0.457|
| "Power sales"   | 0.556    | 0.192       | 0.286|
| "Power guarantee" | 1.5      | 0.3         | 0.5  |

Figure 6. Comparison of search results for "reliability", "powerguarantee" and "powersale"

4. Conclusions

This topic constructs a multi-source search method, which uses a bottom-up approach and combines the characteristics of a catalog search and a metadata search engine. It changes the manual collection part of the catalog search and uses semi-manual overlay to automatically generate hierarchical cluster navigation. The system retrieval results strengthened the combined representation of multi-source information. The digital data information charts, digital knowledge-oriented maps, and search result correlation information were used to integrate the local data content in a componentized manner, retaining more expandability. The numbers form intelligent recommendations for related information, producing a contrasting effect of similar information from different data sources. For information gathering in energy systems, word-directed graphs are helpful for understanding the scattered digital information in a document. The digital information chart displayed through correlation calculations introduces structured information (authoritative and official sources) to ensure the multi-sourceness of the result information in the energy file, and improves the fusion and supplementation of multi-source information.

The relevance calculation ranking is used to optimize the display of the data content of the energy chart, which has a good practical application effect for quickly finding the relevant digital knowledge in the document. In addition, in terms of optimizing the correlation algorithm, digital direct features can be used as local weights, and the recall rate can continue to be optimized and improved, so there is still room for improvement in the combination of algorithms.
References

[1] Morsey, M. & Lehmann, J. & Auer, S. (2012). DBpedia and the live extraction of structured data from Wikipedia. *Program*, 46(2), 157-181.

[2] Bollacker, K. & Evans, C. & Paritosh, P. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. Proc. Sigmod. 1247-1250.

[3] Wu, W. & Li, H. & Wang, H. & Zhu, K.Q. (2012). Probase: A probabilistic taxonomy for text understanding. Proceedings of the ACM SIGMOD International Conference on Management of Data.

[4] Xu, J & Gan, L & Deng, L. (2015). Dependency parsing based Chinese open relation extraction[C]// 2015 4th International Conference on Computer Science and Network Technology (ICCSNT). IEEE.

[5] Bin, G. (2010) Auxiliary function of clustering in network information retrieval[J]. Information Technology.

[6] Zhifeng, W. (2004). The Reality and Future Development of Search Engine[J]. New Technology of Library and Information Service, 83(13):1759-1764.

[7] Alexopoulos, P. & Zhao, Y. & Pan, J.Z. (2017). Construction of Enterprise Knowledge Graphs (II)*[M]// Exploiting Linked Data and Knowledge Graphs in Large Organisations. Springer International Publishing.

[8] Soderland, S. (1999) Learning Information Extraction Rules for Semistructured and Free Text[J]. Machine Learning, 34(13):233-272.

[9] Guo, J & Zhao, J & Yu, Z. (2012). Research on semantic label extraction of domain entity relation based on CRF and rules. [C]// International Conference on Web Technologies & Applications.

[10] Yupeng, J & Hongxu, H & Ping, Y. (2015). LSA-Based Chinese-Slavic Mongolian NER Disambiguation[C]// 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM). IEEE.

[11] Zhou, L. & Yu, H. & Guo, C. (2015). Research on text similarity algorithm based on improved TF-IDF method. Journal of Taishan University, 37(3).

[12] Chen, E. & Jiang, E. (2017). A review of research on text similarity calculation methods. Data Analysis and Knowledge Discovery.

[13] Mai, F. & Li, D. & Yue, X. (2011). Research on Chinese word segmentation technology based on two-way matching method and feature selection algorithm. Journal of Kunming University of Science and Technology, 36(1).

[14] Xue, S. & Niu, Y. (2016). Research on Chinese text similarity based on vector space model. Journal of Electronic Design Engineering, 24(10).