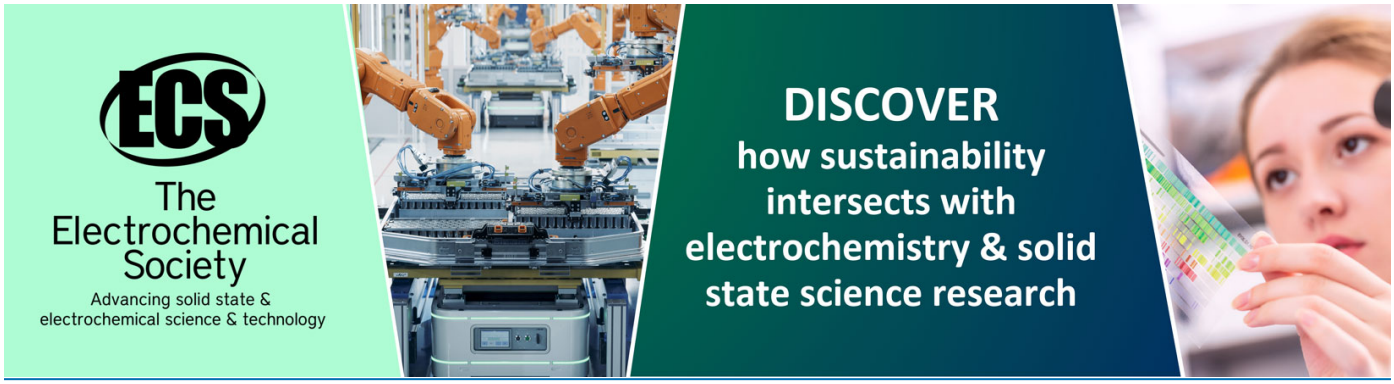
Construction of power industry corpus based on data mining and machine learning intelligent algorithm



IOP Conf. Series: Journal of Physics: Conf. Series **1187** (2019) 022018 doi:10.1088/1742-6596/1187/2/022018

Construction of power industry corpus based on data mining and machine learning intelligent algorithm

Liujun Zhao[1, a], Weizheng Kong[1, b], Qiuling Wang[2] and Lihua Song[2]

[1] State Grid Energy Research Institute CO., Ltd, Beijing 100000, China;

[2] Fujian Yirong Information Technology CO., Ltd, Fuzhou 350000, China.

[a]zhaoliujun@sgeri.sgcc.com.cn, [b]kongweizheng@sgeri.sgcc.com.cn

ABSTRACT: With the advent of the mobile Internet era, the dissemination and diffusion of information has become faster and faster, and the dissemination and generation of information has also increased exponentially. More and more information is generated and diffused in the Internet, and the subsequent problem is that the collection and determination of information increases with its complexity. Therefore, it is necessary to propose and apply a new method to complete the analysis and processing of the information on the Internet. This paper uses data mining technology and machine learning intelligent algorithm to obtain and classify the information data of the power industry on the Internet, so as to construct the power industry corpus.

1. summarize the application of power industry corpus in industry research and industry development.

Corpus is not only the basic resource of corpus linguistics, but also the main resource of empirical linguistic research methods. It can be applied to dictionary compilation, language teaching, traditional language research, statistical or case-based research in natural language processing, etc. The corpus of the power industry is based on documents in the power industry, social news about the power industry, government announcements, research reports of scientific research institutions and other textual information to study policy and technological trends.

1) role in industry research

Constructing corpus can greatly improve the preprocessing ability of complex industry information, screen out time-sensitive and effective data by machine, and then analyze and study the corresponding trends. Through word frequency analysis, we can get the technical heat. Through geographical graphic analysis, we can get the regional technical preference and application degree. More intuitive expression. It will play a guiding role in future business expansion and research prospects arrangement, which is beyond the traditional methods and individual subjective analysis methods. Moreover, corpus method can extract elements quickly and objectively, so long as the algorithm is properly screened, it can achieve high accuracy. Moreover, the accuracy of prediction and analysis can be well quantified and measured. This is difficult to achieve before the popularity of AI technology. It can be said that the establishment and maintenance of corpus today has become a technology that can not be bypassed in the industry.

2) application in industry development

The main function of corpus construction is to study and analyze the situation of the industry. Based on these data, we can judge and analyze which technology and industry applications are being

sought after by the market according to the trend. At the same time, based on the relationship between the correlation and the network hierarchical distribution, the correlation can be representatively expressed.

2. data mining technology based on web crawler technology

1) introduction to web crawler technology

Web crawler technology is an application program or script program that automatically captures Internet information based on certain rules. It is widely used in Internet search engines or other news portals. It can automatically collect all the page content that it can access in order to obtain or update the content of these websites. Search mode. Functionally speaking, crawlers are generally divided into three parts: data acquisition, processing and storage. The traditional crawler starts with one or several URLs of the initial web page and obtains the URLs of the initial web page. In the process of crawling the web page, it constantly extracts new URLs from the current page and puts them into the queue until it meets certain stopping conditions of the system. The workflow of focused crawler is complex. It is necessary to filter topic-independent links according to a certain web page analysis algorithm, retain useful links and put them in the waiting URL queue. Then, it will select the next page URL from the queue according to a certain search strategy, and repeat the process until it reaches a certain condition of the system. In addition, all web pages captured by crawlers will be stored by the system, analyzed, filtered, and indexed for subsequent query and retrieval; for focused crawlers, the analysis results obtained in this process may also provide feedback and guidance for future crawling process.

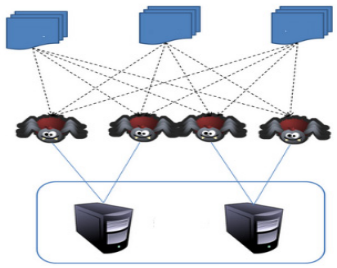


Fig. 1. graphical representation of web crawler

2) construction of web crawler

The basic workflow of web crawler is as follows:

(1) Firstly, according to the information we pay attention to, we adopt appropriate strategies to select some carefully selected seed URLs.

(2) put these URL into URL queue to be grabbed.

(3) Take out the URL queue to be crawled, parse the DNS, get the IP of the host, download the corresponding pages of the URL, and store them in the downloaded webpage library. In addition, put these URL into the URL queue that has been grabbed.

(4) Analyse the URL in the grabbed URL queue, analyze the other URLs, and put the URLs in the queue to be grabbed, so as to enter the next cycle.

(5) add the information concerned to the database.

3) preservation and updating of information data.

Web crawlers crawl down pages that are large text, and can design a storage method to store large-scale data. It should not be appropriate to store it in relational databases such as MySQL or sqlserver. First of all, the pages are relatively independent, basically no relationship, only the simple relationship of URL or describing text corresponding pages, and relational database system in order to support relations and efficient query will increase a lot of additional costs, which is not worth the cost. Moreover, crawlers should be highly efficient in crawling pages. If a relational database is used to store pages, a large number of data will be inserted into I/O in a short time. Insertion is bound to be a bottleneck problem, which is also a big pressure for database maintenance network and physical disk. Therefore, I think it is appropriate to store data text directly. Open-source larbin crawler also uses text-based storage, but it defaults to store a physical file for each page. In my personal opinion,

frequent file creation, writing, flush, shutdown, and system overhead are also relatively large. Considering comprehensively, I designed a scheme, that is, a physical file stores multiple pages. In order to support proper search, segmentation and merge operations, the data file will correspond to an index file. In this way, in the operation project, it can be re-indexed in the file, the index file is much smaller than the data file, traversing or querying will be very fast. Not only if, when data is merged, only index files need to be merged, which will be much more convenient. The specific format is shown in Figure 2.

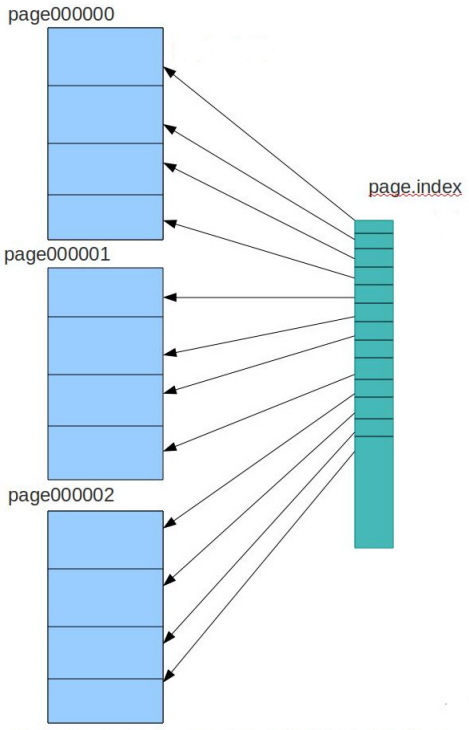


Fig. 2. storage format and structure design

3. Natural Language Processing and multi-level classification of databases

1) Natural Language Processing technology brief introduction

Natural Language Processing (NLP) is a field of computer science, artificial intelligence, and linguistics that focuses on the interaction between computers and human (natural) languages. Therefore, Natural Language Processing is related to the field of human-computer interaction. Natural language processing faces many challenges, including natural language understanding. Therefore, natural language processing involves the area of human-computer interaction. Many challenges in NLP involve natural language understanding, i.e., computers derive from the meaning of human or natural language input, and others involve natural language generation.

Modern NLP algorithm is based on machine learning, especially statistical machine learning. Machine learning paradigm is different from the previous attempt to deal with language. The implementation of language processing tasks usually involves the large set of rules that are directly encoded by hands.

2) participle technology for data classification

In text analysis, word segmentation technology occupies a very critical position. Its main purpose is to divide continuous text into specific lexical elements with meanings. In terms of participle comprehension, Chinese is much more complicated than English. English sentences consist of words, and spaces are used as natural delimiters between words, while Chinese sentences and paragraphs do not have obvious delimiters.

Only words, sentences and paragraphs can be simply delimited by clear demarcation marks. Only words do not have a formal demarcation mark. Although English also has the problem of phrase demarcation, at the word level, Chinese is much more complicated and difficult than English.

These segmented words will be sent to the backstage dictionary for matching, and then the matching results will be brought to the computer, so that the meaning contained in the text can be properly understood.

3) multi level division of word segmentation technology

Using word segmentation technology to establish preliminary partitioning data, at the same time, according to the correlation and frequency of word segmentation, the second or even multiple partitioning is carried out. Specific methods include frequency allocation, weight allocation, correlation degree allocation and logical allocation.

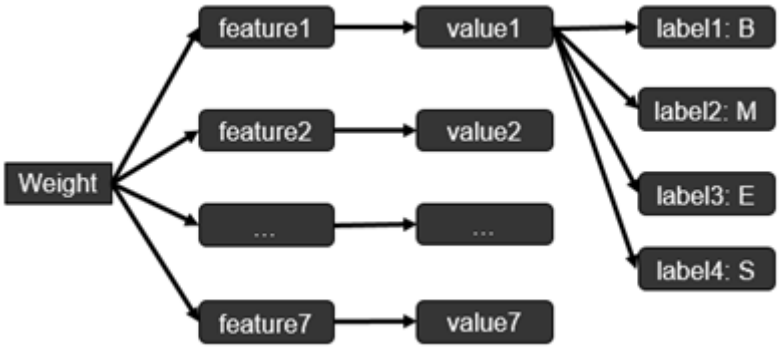


Fig. 3. multilevel division of participles

4. construction and display of knowledge map based on R language

1) introduction to R language text processing technology

Although the ability of processing text in R language is not strong, it can greatly improve the efficiency of work if used properly. At the same time, R language package is a processing ability with high statistical characteristics, so some text operations have to be processed with it. Regular expressions are indispensable for efficient text processing. Although R is inherently inefficient in this respect, it uses regular expressions for most functions dealing with strings.

Regular expressions are expressions used to describe / match a text set. The specific ways of implementation are as follows:

(1) All English letters, numbers and many displayable characters themselves are regular expressions to match themselves. For example,'a'is a regular expression matching the letter'a'.

(2) Some special characters are not used to describe themselves in regular expressions. They have been "escaped" in regular expressions. These characters are called "metacharacters".

Square brackets denote the selection of any one of the square brackets (e.g. [a-z] denotes any lowercase character); ^ Put at the beginning of an expression to denote the beginning of the matching text, and at the beginning of a square bracket to denote any character in non-square brackets; Brackets denote the number of repetitions of previous characters or expressions; | denotes optional items. That is to say, the expression before and after is selected.

(3) using a reference sign (or code change symbol), usually a backslash "/". It should be noted that in R, two backslashes, i.e.'\\', are used. If parentheses are to be matched, they should be written as'\\(\\\\)'

(4) Different languages or applications (in fact, many rules are common) define special metacharacters to represent certain types of characters.

2) knowledge network construction

Using web crawler technology to acquire and update data, at the same time, according to the built machine learning algorithm, natural language processing and word segmentation are carried out. The processed text is segmented twice or more by using R language to construct the database. According to

the systematic classification of power industry websites and technical journals, the statistical and relevance-based word segmentation libraries are obtained. The two libraries are analyzed separately, and the first 200 keywords are screened out in R language for display on the web.

3) dynamic network updating and learning

Finally, the automation technology is used to crawl the Internet network data regularly, and a new database is constructed. Compared with the previous version, the words appearing many times and ranking first are added to the corpus. So as to achieve the purpose of dynamic learning.

References:

[1] Zhao Xiaoliang, Liu Yu Zhang. The method of establishing the thesaurus of fiscal classification [J]. library and information guide, 2002, 12 (4): 31-32.
[2] Zou Qimeng, Liu Qing, Yin Xianjun. Establishment method of classification model, keyword selection method and device of SEO thesaurus: CN106294416A [P]. 2017.
[3] Liao Liang. System design and Implementation Based on Bilingual thesaurus retrieval and classification [D]. Kunming University of Science and Technology, 2017.
[4] Liu Kaiying, Guo Bingyan. Natural Language Processing [M]. Science Press, 1991.
[5] Tang Yincai. R language and statistical analysis [M]. higher education press, 2008.
[6] Wu Yingliang, Wei Gang, Li Haizhou. A Chinese word segmentation algorithm based on N-gram model and machine learning [J]. Journal of Electronics and Information, 2001, 23 (11): 1148-1153.
[7] Liu Jun. 1. Summary of research on machine learning algorithms in the field of artificial intelligence [J]. digital communications world, 2018 (1).