Research on Key Indicators Extraction Technology of Power Transmission Engineering Design Report

Kun Chen
Network planning research center
Guangdong Power Grid Corporation
Guangdong, China
chenkun@gd.csg.cn

Guoyong Li
Beijing Heng Hua Albert Technology Co., Ltd.
Shanghai, China
lgy@ieforever.com

Bochong Pan
Network planning research center
Guangdong Power Grid Corporation
Guangdong, China
panbochong@gd.csg.cn

Zhiming Liu
Beijing Heng Hua Albert Technology Co., Ltd.
Meizhou, China
lzm@ieforever.com

Xiangbing Wang
Network planning research center
Guangdong Power Grid Corporation
Guangdong, China
wangxiangbing@gd.csg.cn

Abstract—In the power transmission engineering design review work, assessors need to repeatedly read the report in order to extract the engineering evaluation index content from a lot of text and tables, then give a comprehensive review comments. In the process, experts need to search the document content and record related indicators repeatedly, which requires expert done manually and greatly affect their efficiency and accuracy. This paper focuses on engineering design report key indicators extraction technology, which extract key indicators information from unstructured engineering report and format a structured review data. It is combined with big data technology. In the Hadoop framework, extract and filter content using MapReduce programming model, which made exploratory research on program extraction technology. We developed a software program using the new program extraction technology, which can system analysis and intelligence extract design report submitted by word. Automatic obtain the key indicator value in the report, format evaluation indicators report. It avoids the evils artificial obtain, improve quality and efficiency.

Keywords-Design review; Engineering design report; Indicators extract; Big data; Regular expressions

I. INTRODUCTION

In recent years, with the rapid growth in demand for electricity, power transmission project construction is also developing rapidly. In the power transmission engineering design review, faced with a large number of engineering design report documents, to improve the efficiency of evaluation experts reviewed, we urgently need some automated tools, which can quickly and accurately find are really needed information in the design report. Key indicators information extraction technology of power transmission engineering design report is raised to solve this problem.

Indicator information extraction technology is different from information retrieval. Indicator information extraction needs to extract key information from the design review report. It broke through the limitation of reading and understanding from the design report and implement automatic search and understanding. Design report is generally unstructured text information. By extracting information on key indicators form formatted form information to facilitate evaluation experts to evaluate, thus improving the efficiency of the project review.

When we need extract technical indicators from a lot of documentation at the same time, we will combine big data technologies and use MapReduce programming model based on a regular expression extraction and filtration machine learning method based on key indicators on Hadoop platform for rapid extraction [1-5].

II. RELATED TECHNOLOGIES

A. Regular expression extraction

A regular expression is a string of special characters. It is used to describe or match a string with a certain distribution law. In the pages of information extraction, web pages need to be treated as a text character stream, and then according to the configured regular expressions to extract the required information [6-8].

B. Filtering method based on machine learning

Machine learning methods have been applied to many fields. For garbage information processing, it access to the rules of indentifying spam mainly through training the history spam and normal information. Then through these rules, we process new data and identify spam.

Algorithms based on machine learning have many ways for classification of spam, such as decision tree, naive Bayes, K nearest neighbors, support vector machine (SVM), Boosting methods, etc. The following mainly introduce multiple linear regression method.

Multiple linear regression method needs vectored data processing, that is statistical processing the key word or words, and then based on the situation key words or word appears, consider which type of the message belongs. In this paper, we identify the spam according to the statistical information of a key word or words. This method of training is slow, but because of easy parallel implementation, combined with Hadoop distributed platform, parallel processing, processing speed is relatively fast.

C. MapReduce programming model

MapReduce is a programming model for parallel computation of large data sets. The concept "Map" and "Reduce", and their main ideas are borrowed from functional programming language, as well as properties from vector programming language. It greatly facilitates the programmer who doesn’t know how to do distribute parallel programming to run his program on distributed systems. Typically a MapReduce job process includes two stages: Map and Reduce stage. These two stages use key / value pairs as a function of input and output. The type of input and output of key / value pairs are decided by programmers based on need. At the beginning, MapReduce will divide input data into independent data block which map function can parallel process completely. MapReduce framework will then classify the output of the map function, as the inputs of reduce function. After reduce function treatment is completed, each reduce task generates an output file [9-11].

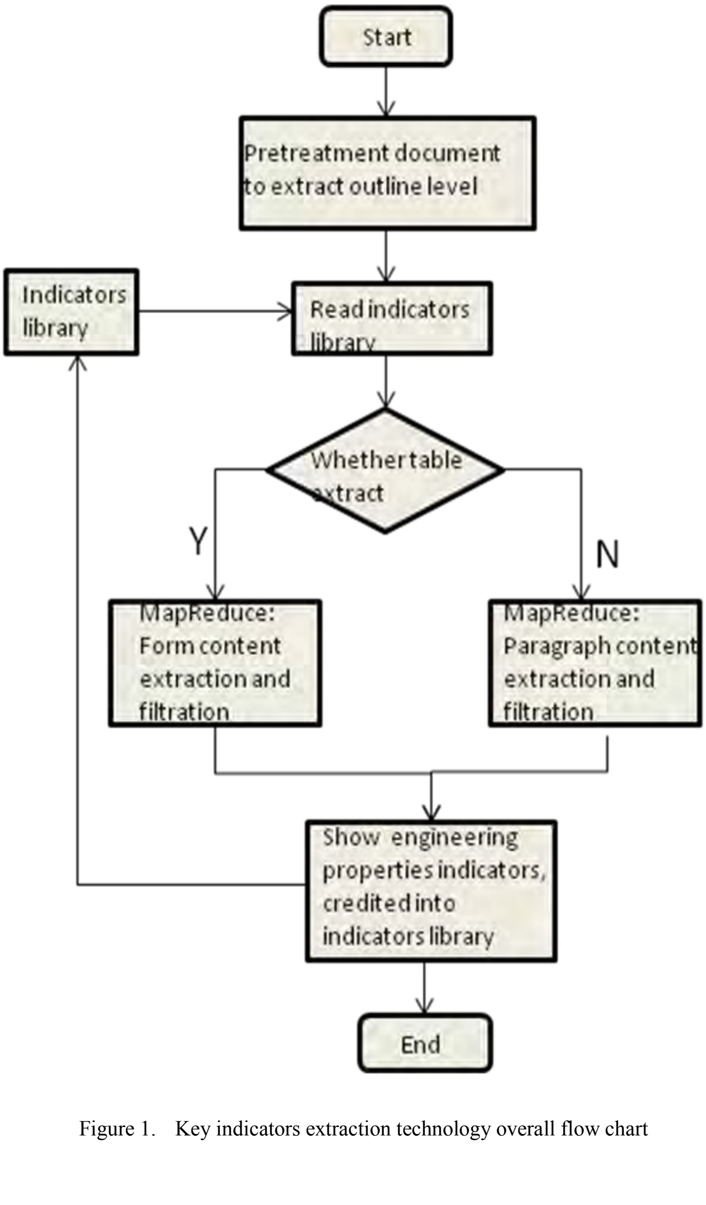

Start
Pretreatment document to extract outline level
Indicators library
Read indicators library
Whether table extract
Y
N
MapReduce: Form content extraction and filtration
MapReduce: Paragraph content extraction and filtration
Show engineering properties indicators, credited into indicators library
End

Figure 1. Key indicators extraction technology overall flow chart

III. KEY INDICATORS EXTRACTION TECHNOLOGY PROGRAM

Key indicators information of transmission and transformation project mainly exists in engineering design report paragraphs of text and tables. Different key indicators are in different chapters, which are described by paragraphs and tabular form. First, define key indicators information of project. Then, vectorize engineering design reports, extract the report outline title, classify description and tables describe, use a variety of methods to extract and integrated display information on the engineering properties, which facilitate evaluation experts view the engineering properties of index information and improve the efficiency of evaluation experts review. The overall flow chart is shown in Fig .1.

Key indicators library is mainly used to store attribute information of engineering properties indicators, including the document title name. If in the table, we need to configure the ranks' title and the cell to extract regular expression. If in the text description, we need to configure the keywords of text description and extraction method.

IV. EXTRACTION TECHNOLOGIES TO ACHIEVE KEY INDICATORS

A. Documentation pretreatment to extract outline level

In word, every paragraph has an outline level attribute: body text or specific level, such as level 1, level 2, level 3 and so on. When editing an article word document, people can use the word comes paragraph headings, bullets, etc., collectively, the "outline level". Word comes outline level is a kind of tree-structured data; at the same time, you can also write the paragraph number directly. By setting some common identification numbers and letters we can distinguish different paragraphs headings, which is called "Custom outline level", such as " title 1" or " title a". Word document itself does not recognize custom outline level. Therefore, when extracting the word document to outline level, we need to consider the word itself outline levels and custom outline level for extraction.

When extracting the document outline levels, we need to record the document paragraph number where outline level is, the outline-level and the outline-level where the table is.

The main steps for extracting word document outline levels are as follows:

1) Initialize documents, record the number of paragraphs each table occupied;

2) Traverse every paragraph in the document, parse the paragraph attribute information, record every few paragraphs;

3) Determine whether the paragraph attributes in the table: if in the table, skip the paragraph number, record the sequence number that the table appears in the document and the outline title which the table is in. Return to step 2;

4) Determine the paragraph properties. If the property value is not "body text" paragraph, remove paragraph outline level value directly. If it is "body text," then set outline level to body text, return to step 2;

5) Determine the outline level as "body text", use regular expressions to parse the content of the paragraph, parsing rules are as follows:

a) the characteristics of custom outline paragraph is beginning with numbers and letters, outline numbering spilt with "." . And if the outline is a digital start, then there must be a space to distinguish between them, such as "1 110kV power distribution device";

b) Filtered the paragraph which start with a number and is not the outline, such as "220 kV substation main power supply range Tong Mei Hui District in the southwest region" for paragraph begins. In this situation, we filter the digital followed by "kV, mA, kV" and other special characters;

c) According to the figures, letters using regular expressions parsed paragraph outline level.

B. Extraction and filtering table content

In the Hadoop framework, do the following steps using MapReduce programming model:

1) Get the chapter headings where engineering properties indicators are and the ranks of the title and the expression that extracting forms needs.

2) Matches the outline level extraction results, get the corresponding outline title and child nodes below outline form and form number;

3) Locate the corresponding form number directly according to the table;

4) Traverse each table, determine the unique cells depending on the ranks title, extract of cell contents or use an expression to extract the contents directly;

5) According to the statistics of the key word or words, distinguish spam and normal information, filtering spam.

C. Paragraph content extraction and filtration

In the Hadoop framework, do the following steps using MapReduce programming model:

1) Get the section title , content extraction of keywords and synonyms, content extraction methods where the engineering properties indicators are;

2) Matches the outline level extraction results, get the corresponding outline title and the paragraphs contents of the child nodes below;

3) Use keywords and synonyms, content extraction method to extract the contents of a paragraph. Extraction methods are:

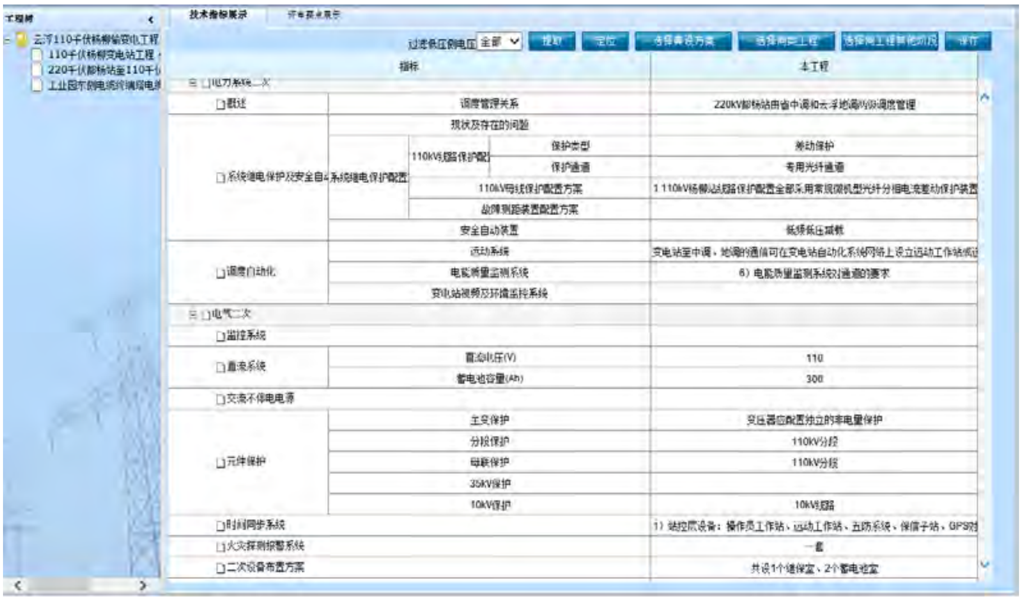
- Expression: extract the contents of the relevant heading, according to the configuration regular expressions to extract index information;
- Punctuate method: extract the contents of the relevant heading, extract statement where index resides according to the configuration information and its synonyms keyword;
- Exactly matching method: extract the contents of the relevant heading, according to the configuration keywords and their synonyms, using flexible matching algorithm BPD algorithm for rapid extraction;

- Extract chapters: extract the contents of the paragraph in the section where the outline of the title is;

4) Statistical processing the key words or words. Then according to the situation of these key words or word appears, considering the category to which the information belongs. In this text, according to the statistical information of the key words or the words, distinguish spam and normal information, filtering spam.

V. Application of Technology to Extract Key Indicators

Accreditation Unit has provided key indicators of concern in the assessment process, we developed a software program, which can system analysis and intelligence extract design report submitted by word that the unit submitted. Automatic obtain the key indicator value in the report, format evaluation indicators report. It avoids the evils artificial obtain, improve quality and efficiency. Show the extracted corresponding technical indicators in accordance with the number of engineering works and in tabular form for expert review and inspect. Software interface is shown in Fig. 2.



指标				本工程
⊟ □电力系统二次				
□概述	调度管理关系			220kV郎杨站由省中调和云浮地调两级调度管理
□系统继电保护及安全自	现状及存在的问题			
	系统继电保护配置	110kV线路保护配置	保护类型	差动保护
			保护通道	专用光纤通道
		110kV母线保护配置方案		1 110kV杨柳站线路保护配置全部采用常规微机型光纤分相电流差动保护装置
		故障测距装置配置方案		
	安全自动装置			低频低压减载
□调度自动化	远动系统			变电站至中调、地调的通信可在变电站自动化系统网络上设立远动工作站或远
	电能质量监测系统			6）电能质量监测系统对通道的要求
	变电站视频及环境监控系统			
⊟ □电气二次				
□监控系统				
□直流系统	直流电压(V)			110
	蓄电池容量(Ah)			300
□交流不停电电源				
□元件保护	主变保护			变压器应配置独立的非电量保护
	分段保护			110kV分段
	母联保护			110kV分段
	35kV保护			
	10kV保护			10kV线路
□时间同步系统				1）站控层设备：操作员工作站、远动工作站、五防系统、保信子站、GPS对
□火灾探测报警系统				一套
□二次设备布置方案				共设1个继保室、2个蓄电池室

Figure 2. Figure 2. Software interface

VI. Conclusion

In recent years, with the rapid growth in demand for electricity, grid work has expanded rapidly. To ensure the construction quality, we must increase the project feasibility study and design review efforts. Power transmission engineering design report key indicators information extraction techniques extract key indicators information from unstructured engineering report and format a structured review data. It is convenient for engineering design review expert review and improves the efficiency of project evaluation, while stored the key index information into the database, research and analysis of the data further. By means of information, assist assessors complete the assessment efficiently and quality, escort our grid project.

References

[1] Peng Wang. “The key technology and application examples of cloud computing” [M].Beijing: People Post Press,2009.

[2] Yan Guo. “Network information extraction technology research” [J]. IT Letters,2008(6) : 15 －23.

[3] Amazon Incorporation..Amazon elastic compute cloud[EB/OL]. http: / / aws. Amagon. com/ec2. [2010 －04 －30/2012 －09 －06］.

[4] Shaoling Sun,Zhiguo Luo,Meng Xun,etc. “Research and implementation of cloud computing application” [J]. Telecommunications Engineering Technics and Standardization,2009(11) : 2 －7.

[5] Qinghua Zheng,Jun Liu,Feng Tian,etc. “WEB knowledge mining: Theory”, Methods and Applications［M］. Beijing: Science Press,2010.

[6] Intelligence Science. “Hadoop-based distributed parallel data mining platform PDMiner” [EB/OL]. Http: //www. intsci. ac. cn/pdm/msmirer. hml［2010 － 06 －23/2012 －09 －06］.

[7] Mobile labs. “Research on parallel data mining tools platform based on mobile computing” [EB/OL]. Http: //labs , chinamobile. com［2009 － 03 － 25/2012 － 09 － 06］.

[8] Kang Chen,Weimin Zheng. Cloud Computing: “Examples of systems and research status” [J]. Journal of Software,2009(5): 1337 －1348.

[9] Miao Cheng. “WEB data mining based on cloud computing” [J]. Computer Science,2011,38(B10): 146 －149.

[10] Xiao J,He L. “Keyword weight adjusting schema based on domain repository”[C].Chengdu, China:3rd IEEE International Conference on Computer Science and Information Technology, 2010:221-225.

[11] Zhang R,Yasuda K, Sumita E. “Improved statistical machine translation by multiple Chinese word segmentation” [C].Proceedings of the Third Workshop on Statistical Machine Translation. Columbus, Ohio, USA: Association for Computational Linguistics Stroudsburg,2008:216-223.