Acta Technica *62* (2017), No. 5B, 947–952

Analysis and research of forestry key words based on large data

Mao yan-xin[1,2]

Abstract. In this study, according to the analysis of large data we use the digital processing method and keyword extraction to analyze and study the high-frequency words of forestry based on large data. Through the "Forest resources" and "Forest fire prevention" two words of analysis and research, and the statistics of word frequency, weight factor, and by analyzing the frequency and weight of the keywords of large data to reflect the content and degree of concern in the forestry industry.

Key words. Forestry, Keyword extraction, Wildlife.

1. Introduction

Through the analysis of large data, by the year 2015, forest fire prevention ranked the first place in the total ranking of key words in forestry industry, and the frequency reached 140 thousand and 100 times, accounting for 15.19% of the first 100 key words[1]. "Forest resources", "wild animals", "state-owned forest farms", "forestry pests", "collective forest rights", "Forest Park", "prevention and control", "sand control and desertification control", respectively, occupy the top 10 positions.

In the first 100 statistics, the high frequency words about forest fire prevention and management are 11, respectively, forest fire prevention, forest fire prevention, forest fire prevention, forest and grassland fire prevention, forest fire prevention, protection, fire protection, anti-fly, the frequency of a total of 239 thousand and 800 times, accounting for the top 100 keywords frequency 26.01%[2]; keywords related to forestry pest control, forest pest prevention and control, including pine wood line, forest pest control, the total frequency of 90 thousand and 300 times, accounting for 9.80% of the top 100 keyword frequency[3]; keywords relates to afforestation with tree seedlings, planting trees, replanting, national compulsory tree planting, afforestation, greening and beautification, spring planting, operation design, trans-

[1]Workshop 1 - Forest Economics and Development Research Center, State Forestry Administration, Beijing, 100714, China

[2]Corresponding author: Mao yan-xin

formation, low yield forest villages landscaping, greening, the total frequency of 65 thousand and 700 times, accounting for 100. 6.57% of the key word frequency[4]. From the statistical results can be seen that the government departments pay attention to people's livelihood, the key words related to a larger probability, including forest rangers, workers, masses, shantytowns, difficult workers and so on.

In 2012, LI S.X discussed Key words in China from the statistical analysis of research on open access[5]. Based on the data of open access in China collected by China Journal Full-text Data Base, keyword statistical method was carried out on Open Access of the research in China, involved in disciplines, major research institutions, publishing policy, publishing channels and users[6]. To understand our open access research papers on the subject content and research advances the study also pointed out problems in the study for future reference. In 2014, Wang S. studied and optimized the keyword search algorithm for large data[7]. In 2016, Tang X. L. studied the semantic-based Linked Data large data keyword search[8].

2. Techniques

2.1. Data processing methods and processes

In order to complete the study, first of all to get the industry government website public information, it is necessary to develop a web crawler tool, and the data is collected and processed, the formation of special database, and use the Chinese segmentation algorithm, keyword extraction algorithm to extract the high frequency keywords[9].

The first step is to collect sites within the forestry industry and use the home page as the web crawler's entry address[10]. The second step, the development of web crawler (Net Crawler), the main work is data capture all the web pages for each site, and the analysis, to extract the valid data, including the title, URL, parent content, release time, time of grasping the field. When done, save the area to the CSV file. The third step is to standardize the data acquisition, including eliminating the garbled, table arrangement and so on. The fourth step is to build a special database, design database tables and store them in the database. The fifth step of keyword extraction, extraction before, there are two ways, one is through the program of each content field (content) merged into a document, and then use the extraction algorithm to extract; another way is through the process of each content field (content) with export to file, then the keyword extraction of files.

The sixth step, the results are saved to the database, using SQL for various statistics.

2.2. Keywords extraction

Keywords extraction based on large data involves the problem of how to deal with Chinese information. Words are the smallest, meaningful language components that can move independently. In Chinese, there is no delimiter between words, word itself is also a lack of morphological markers, so obvious, unique Chinese information

processing is how to Chinese string into word sequences reasonably, namely Chinese segmentation.

In the Chinese segmentation field, there are a lot of achievements, and the formation of some commercial applications. The core algorithm has two main categories: the first category is the rule method based on linguistic knowledge, such as various forms, at least the maximum matching segmentation method, and comprehensive maximum matching and least segmentation N- shortest path method, the second category is a large-scale corpus based machine learning method, which is widely used currently, effect a better solution. In the segmentation based on automatic keyword extraction can be realized, the main algorithms including TF-IDF (Term Frequency-Inverse Term Frequency), this algorithm requires the support of corpus, another algorithm is based on information entropy (Information entropy) of the keyword extraction algorithm does not need the support of corpus.

In this paper, the Chinese Academy of Sciences NLPDIR Chinese word segmentation algorithm and keyword extraction technology to deal with all the data crawling within the forestry industry, and found that most of the key words are new words (n_new), meaning not in the dictionary.

3. Mathematical analysis

In order to further understand the word within the industry focus on the content of this paper, further to the high frequency words as keywords to the mining of the relevant content, and calculate the frequency of keywords associated with high-frequency words, trying to reveal the relation between.

3.1. Forest resources

Large data from 2006 to 2015 shows, the "forest resources" attention and the amount of information continues to increase. In 2006, the number of "forest resources" documents was 459, and by 2015, it increased to 7027, with an average annual growth rate of 35.41%.

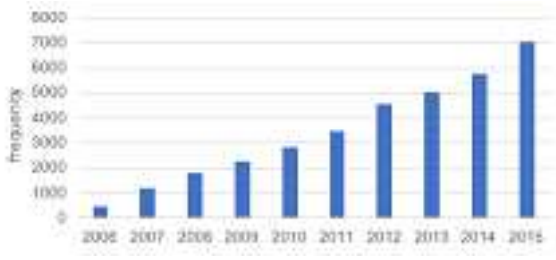

Fig. 1. "Forest resources" information quantity in 2006-2015 years

First, the forest resources survey work, "two investigation", "land change survey", "forest resources assets", "forest coverage rate" as high frequency words reflect the survey of forest resources and forestry basic work content; second, the protec-

tion of forest resources. "The forest fire" as the industry of high frequency words also appeared with "forest resources" has a strong correlation, which is from the forest fire protection. "Forest", "forest harvesting quota", "special action" is the specific measures of protection, third, in the forest resources management process, there are some in the industry pay close attention to the illegal behavior including, "deforestation", "deforestation", "illegal occupation of forest land".

3.2. Forest fire prevention

"Forest fire prevention" as the industry's most attention to high-frequency words, Through the analysis of large data between 2006-2015 years, 9 years in the first place. Related information has been telling growth, with an average annual growth of 42.27% over the past 10 years.

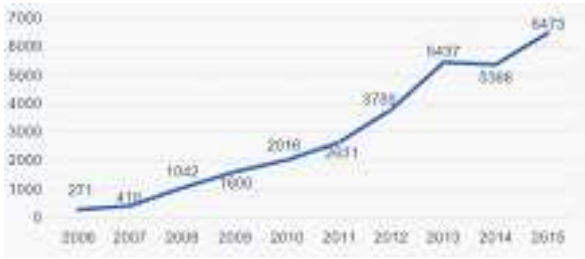

Fig. 2. "Forest fire prevention" information quantity in 2006-2015 years

Further study of the "forest fire", with relevant keywords almost covers the main elements of the work of forest fire prevention, but also reflects the special and professional forest fire prevention, forest fire prevention headquarters "as a row of high frequency words in front of the fire, that must be established in order to force the command center to complete special tasks. The fire is from "forests forest fire prevention headquarters", "forest fire office", "the competent forestry department responsible for the work of the organization, the implementation of" "leaders" responsible, "the chief executive responsibility system". Participate in organizations and individuals including "Rangers", "fire brigade", "Forest Public Security Bureau", "members of the unit command", "staff", "forest armed forces", "emergency unit", "air Ranger station". According to seasonal types of fire prevention, "spring forest fire prevention", "autumn and winter forest fire prevention", "forest grassland fire prevention", "summer forest fire prevention.". "Ganzi" can be ranked as the key word, indicating that the forest fires in Ganzi have a greater impact.

4. Conclusions

Through the analysis of the keyword large data shows that the frequency from the search results, 2009-2015, forestry policy keywords ranking in the top ten academic did not change too much, forestry policy keywords ranking in the top ten social networks has changed greatly, and the search engine ranking before ten forestry policy

and the overall search keywords forestry policy the top ten are basically the same. This shows that search engines are more concerned about the retrieval of forestry policy keywords than academic and social networks. From the distribution of the index weight, we can know that the weight of social influence is higher, and the weight of search influence and academic influence is smaller. This shows that the influence of social network influence on network comprehensive influence is higher than that of search power and academic influence. Comprehensive evaluation results from the influence of the network, social network in the public influence although less than the search engine, there is a certain gap, but the influence of social networks is increasing gradually, but the academic influence declined. Therefore, the forestry industry can advantage in information communication ability and deepen the leveraging social media, social networking platform to expand the scope of radiation, to attract more audience attention, and then improve the network influence of forestry policy.

References

[1] E. J. Kenney: *Keywords-vocabulary of culture and society.* New republic *19* (1976), No. 2, 27–28.

[2] X. S. Zhou, T. S. Huang: *Unifying keywords and visual contents in image retrieval.* Ieee multimedia *9* (2002), No. 2, 23–33.

[3] E. F. Keller, E. A. Lloyd: *Keywords in Evolutionary Biology.* Science *260* (1993), 1153–1154.

[4] B. Amento, L. Terveen, W. Hill: *Keywords exploiting hyperlink structure does "authority " mean quality? predicting expert quality ratings of web documents.* Proc acm sigir (2000), 296–303.

[5] M. A. Andrade, A. Valencia: *Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families.* Bioinformatics *14* (1998), No. 7, 600–601.

[6] A. Ghaziani, M. J. Ventresca: *Keywords and cultural change: frame analysis of business model public talk.* Sociological Forum *20* (2005) 523–559.

[7] R. R. Souza, K. Raghavan: *Extraction of keywords from texts: an exploratory study using noun phrases.* Ophthalmology *111* (2014), No. 4, 2–9.

[8] A. Simitsis, G. Koutrika, Y. Ioannidis: *from unstructured keywords as queries to structured databases as answers.* Vldb Journa *17* (2008), No. 1, 117–149.

[9] O. J. Rutz, M. Trusov, R. E. Bucklin: *Modeling indirect effects of paid search advertising: which keywords lead to more future visits.* Marketing science *30* (2011), No. 4, 646–665.

[10] K. S. Thomson: *keywords and concepts in evolutionary developmental biology.* Bioessays *26*, (2004), No. 2, 214–215.

Received November 16, 2017