Improved Data Analysis Algorithm based on Multi-feature Network Construction

Wenhua Guo

School of Beijing Industry University, Beijing 100000, China.

18612725754@163.com

Abstract. Network information has become an important factor in today's social environment and network environment. With the large coverage of network data traffic and the new generation of network technology, illegal network data is also constantly invading the network environment, which has caused a serious network security threat. Therefore, the analysis and research of network data and the pre-judgment of feature types are of great significance. Based on the existing theoretical techniques, this paper proposes a Data Processing Algorithm Based on The Website Coding Structure (DPA) and a Multi-feature Network Construction Improvement Algorithm (MCIA), processing network data and analysis type prediction features. Experiments show that the data processing algorithm based on the website coding structure has the advantage of targeted processing of website data. It also verifies the validity of the proposed Digital Neighborhood Feature Set Algorithm (DNFA) and the improved algorithm for multi-feature network construction for data analysis. Application, compared with the existing single feature set, the pre-judgment basis for constructing multiple feature sets is more reliable, and the pre-judgment basis is more credible.

Keywords: Network data; Feature set analysis; Data prediction.

1. Introduction

In the 1990s, Bill Inmon[1], the father of data warehousing, proposed the concept of data warehousing, which stimulated the research of data analysis and processing decisions in academia, through data mining[3] and data Analysis[2]gives the value of the value of the data itself. Since then, the door to big data research has been opened. As the main component of big data, network data is the main carrier of network information dissemination. As a reaction to the inherent regularity of data, the feature itself has multi-source heterogeneity. To realize a large amount of data content and data type analysis, it is necessary to abstract unstructured data information into a network architecture, which is composed of data features. For the feature analysis of network data, key steps such as feature selection type and feature extraction method affect the results of data feature analysis. Feature analysis is generally divided into: keywords, word frequency, emotional words [4], part-of-speech tagging [5]. The construction process of feature network requires an efficient and universal automation data [6] processing algorithm to improve data processing efficiency and accuracy.

The modern network is intricate and complex, and the types of network information data are various and difficult to distinguish. The demand for classification, matching and pre-judgment of certain types of information is increasing. The common matching structure of the existing text data is a similarity matching algorithm extended to keywords according to the word similarity algorithm, and the keyword is used as the main feature word for matching calculation. For the study of the characteristics of the general data and the subject reflection, including classification, matching [7], etc., there are certain effects. However, for complex network data, the single keyword feature no longer meets the research needs of network data. Each type of network data has its own unique characteristics in terms of content purpose, expression and text law. The research model cannot fully reflect the unique characteristics of a variety of data, and it is difficult to predict the current changing network data. In the actual research work, there is still a need for a data extraction scheme based on the network coding structure. It also needs an algorithm for diversifying feature research, that is, improving the timeliness of predicting network data and reducing the threat to network data security. The research work of the judgment provides practical significance.

This paper focuses on the current network marketing data as an example and reference data analysis method, from the data extraction efficiency and multi-feature data analysis, combined with

the Data Processing Algorithm Based on The Website Coding Structure (DPA) and the Multi-feature Network Construction Improvement Algorithm (MCIA), the network type of MLM data is prejudged. Through the basic preprocess of data, using machine learning theory and multi-class features of data to aggregate, combined with the proposed feature extension of the digital neighbor feature set, using multiple feature sets as the basis for MLM prediction, and the experimental data Preliminary prediction and reliability assessment. The main contents and innovations are summarized as follows:

First, based on the data processing of the network coding structure (DPA), according to the different structure of the website, the data extraction method makes a corresponding solution, and realizes the different network coding based on the comprehensive utilization of the selenium automation framework, python analysis technology and data local access[8][9][10]. A data processing algorithm for a structured data platform.

Secondly, the digital neighbor word domain feature set (DNFA) is extended, and a network text information set is proposed for a class of digital law or digital mechanism. The large number of numbers appearing in the data content is used as the research object, combining existing keywords, high-frequency feature sets, etc., an improved data analysis algorithm for multi-feature network construction is implemented.

2. Multi-feature Network Construction Improvement Algorithm

This paper takes the data of the network platform as an example. As the basic source of data, through the comprehensive utilization of the main technologies such as selenium automation framework, python parsing technology and data local access, this paper proposes a data processing algorithm based on website coding structure and one for network data. Data feature analysis and feature extension algorithms.

2.1 Data Processing Algorithm based on the Website Coding Structure (DPA)

Compared with the existing automatic data processing software, the paper is based on the coding structure of different websites, paying more attention to the accuracy of the research process and results, and paying more attention to the basic data processing needs of a wide range of data researchers. The algorithm principle based on the website coding structure is represented by pseudo code as follows:

First call the chrome browser driver, and open the website in the chrome browser, you can set the waiting page load time, avoid network delay or slow website loading affect the data capture (1); secondly by getting each page Index (according to the website paging structure, the page id is turn-page), get the data page structure of the website (2); then get the acquisition path of each link (here, take the absolute path as an example (3) (4)); finally Content parsing (effective content of each data set), the valid data of each link can be controlled by the html tag, such as h1 tag, p tag, span tag, and so on.

```
Selenium.Webdriver.Chrome().get(url)                                    (1)

Select(driver.find_element_by_id("turn-page"))
    .select_by_visible_text(str(index)))                                (2)

xpath = //*[@id='artical_real']/table/tbody/tr[" + str(i) + "]/td[1]/a  (3)

urls = driver.find_element_by_xpath(xpath)                              (4)
```

The data processing algorithm based on network coding structure is more suitable for different website structures, data acquisition and processing methods are different. The algorithm implements a targeted pre-processing scheme for various website coding structure data. The algorithm

Combined with the current popular Python open source tool library, the environment is fast, the operation process is simple, the visual results are clear, and the basic researcher is used for various

types of website data; the data processing is automated throughout the process, and all the results in the process can be automatically saved locally. It facilitates the subsequent query and recording of results; it eliminates the tedious operation of comparing the selection, registration and registration of various data processing software.

2.2 Digital Neighborhood Feature Set Algorithm (DNFA)

This paper proposes a DNFA as an extended feature set of feature networks. Combining the existing feature set, the natural language data set is constructed into a multi-feature network set by data preprocessing and multi-directional feature extraction algorithm, and the feature set is refined and constructed. The result set is used for data type features. Library creation, pre-judgment of data types.

Principle of DNFA: Using the high-frequency words adjacent to the left and right as the feature items, the feature set related to the number is obtained, which not only solves the single defect of the word frequency feature set, but also solves the limitation that the traditional only uses the keyword as the feature.

DNFA, used to augment an existing feature set library, reflecting the portion of the data that contains digital laws or numerical mechanisms. The algorithm implementation does the following:

(1) Using the digital neighborhood words in each data as a research domain, using the regular expression (re) in the python tool library, extract the digital neighboring domain, and record the collection of this domain as: (D_1, D_2, ..., D_n);

(2) Perform a Jieba word segmentation operation on all the sets in each of the obtained data, and change each domain into a set of words.

(3) Calculate the word frequency for the left and right data of each region D_n, and according to the frequency of the word, obtain the high frequency words that can represent each set, as the feature set of each set, synthesize all the set results, and finally get the digital proximity. Word feature set.

(4) The same experimental data is used to calculate the keyword feature matching and the digital neighbor word feature matching respectively, and verify the validity of the feature set of the digital adjacent words.

The DNFA algorithm combines existing feature extraction schemes such as keywords and high-frequency feature sets to form a multi-feature network construction algorithm for solving data feature extraction, data feature analysis, data type classification and prediction.

3. Algorithm Application and Empirical Analysis

The experiment is used to realize the time of data processing, the reliability of feature similarity calculation and the prediction of network data type.

3.1 DNFA Validation Test

The feature correlation can be used to verify the validity and reliability of the feature set. The feature set C of the same data set is similarly calculated[7], and the similarity calculation results of the existing key feature set D are compared. The two result ranges are close. Or if the matching value of the feature set C is higher, the validity and superiority of the feature set C feature representation can be explained.

Each word in each data is adjacent to the word frequency feature as a word frequency vector, then the feature set of the digital neighbor words of each data can be represented as a vector C (f_1, f_2, ..., f_n) where f represents the word frequency value, all experimental data The similarity value, that is, the similarity of the calculated vector, the characteristic correlation formula of any two data is:

$$sim(C_i, C_j) = cos\theta = \frac{\sum_{k=1}^{n}(f_k(C_i) \times f_k(C_j))}{\sqrt{(\sum_{k=1}^{n} f_k^2(C_i)) \times (\sum_{k=1}^{n} f_k^2(C_j))}}$$

Where $dfm(C_i, C_j)$ represents the similarity value of the text information C_i, C_j ,and $f_k(C_i)$ represents the word frequency of the digital word domain of the text information.

The experimental results will be compared with the existing keyword feature correlation calculation results. the results show that the data neighborhood feature set also has the characterization characteristics and the characterization effect is better.

3.2 Predictive Application of DPA

3.2.1 Algorithm Implementation of Data Multi-feature Network Construction

Aiming at the improved Selenium framework-based data multi-feature network construction processing algorithm, the prediction of the data type of the improved data processing algorithm is verified. The experiment is based on the anti-MLM website, using the website case data as the basic source of research, and implementing the improved algorithm. The multi-feature network construction algorithm is used to obtain the pre-judgment basis set of the MLM data, thus obtaining the pre-judgment rule, which can be used as the application of MLM pre-judgment.

The initial data of this experiment comes from the network marketing case in the "China Anti-MLM website". The website is operating normally and the data is constantly updated. It provides a relatively stable data source for the research of MLM pre-judgment. The MLM case is used as experimental data, and the reliability of the data source and the accuracy of the experimental results are also improved.

During the experiment, the data obtained in each stage is saved in a local file to facilitate the retention of experimental data. For the actual large scientific data processing process, the database can be configured to access the data. The data is centered on the website data. At present, the website has a total of 13 pages of data pages, 25 text data per page, a total of 311 text data, and the data is constantly updated. The data is based on the data set. The main experimental objects are the keyword feature set, the word frequency feature set, and the digital proximity feature set. The data details are shown in Table 1:

Table 1. Multi-directional feature set pre-judged experimental data set.

Data ID	Feature Type	Num	Data description
ID1	Keyword	311	Keyword feature set
ID2	Word frequency	311	Word frequency feature set
ID3	DNFA	311	DNFA feature

3.2.2 Pre-judging Algorithm and Rules for Multi-feature Network Construction

For the MLM pre-judgment algorithm of multi-directional feature set, this paper designs three similar feature matching experiments: keyword feature set matching, high word frequency feature set matching, and digital proximity feature set matching.

(1. Perform the similarity matching calculation on the keyword feature set, the high frequency word feature set and the digital proximity feature set of each text data;

(2. For the similarity matching calculation of the keyword feature set, n text data is represented as T_1, T_2..., and T_1, T_2, ... are represented by respectively corresponding vectors (V_1, V_2...);

(3. Calculate the cosine similarity of V_1, V_2... relative to all vectors except itself, denoted as Cos_1, Cos_2...

(4. Analyzing all cosine similarities, obtain the minimum value min and the maximum value Max, and calculate the mean value average as the pre-judgment value range of the keyword feature set.

(5. According to the algorithm principle (1, the experiments of high word frequency feature set matching and digital proximity feature set matching are completed respectively, and the pre-judgment range of each feature set is obtained.

The respective pre-judgment ranges A, B, and C of the keyword feature set, the word frequency feature set, and the digital neighborhood feature set are obtained. When an experimental data to be tested is obtained, after the data is preprocessed, a keyword set U_k to be tested is obtained. The word frequency feature set U_f to be measured, the digital proximity feature set Un to be tested, and the existing feature set are used for the multi-directional feature set of the pyramid scheme pre-judging algorithm experiment, and three ranges K, F, and N are obtained:

(1. If the result range K∩A of the keyword feature set U_k to be tested is not empty, it is judged that the text data may have a pyramid selling property;

(2. If the result range F∩B of the keyword feature set U_f to be tested is not empty, it is judged that the text data may have a pyramid selling property;

(3. If the result range N∩C of the keyword feature set U_f to be tested is not empty, it is judged that the text data may have a pyramid selling property.

4. Conclusion

Network data extraction is the basis of data processing. The data extraction method directly affects the efficiency and result of data processing and analysis. The data processing algorithm based on website coding structure (DPA) proposed in this paper has a great deal in the processing of different types of network data. Actual effect. Data feature analysis is one of the important methods of data information correlation analysis and data clustering. The accuracy of feature selection also affects the results of data information processing analysis. This paper proposes a multi-feature network construction improvement algorithm (MCIA), DNFA and a feature set pre-judgment algorithm. The automatic extraction framework of network data features improves the processing speed and accuracy, and the obtained feature set also increases the network data type prediction. The credibility provides a data type pre-judging algorithm scheme, which provides a new reference algorithm for network data extraction, data pre-judgment and other research fields.

The next research work paper will focus on the performance and operational efficiency of the algorithm, and continuously improve and expand other extreme performance problems in the algorithm, so that the experimental results predict the matching value more accurately.

References

[1]. Yuan Shuhan, Xiang Yang, E Shijia. The Understanding and Development Trend of Text Big Data Based on Feature Learning[J]. Big Data,2015,1(03):72-81.

[2]. Cheng Xueqi, Lan Yanyan. Text Content Analysis of Network Big Data[J]. Big Data, 2015, 1 (03): 62-71.

[3]. Crawford M, Khoshgoftaar T M, Prusa J D, et al. Survey of review spam detection using machine learning techniques. Journal of Big Data, 2015, 2(1):1-24.

[4]. ZHAO Wei, QIN Bing, LIU Ting. Text Emotion Analysis[J]. Journal of Software, 2010, 21 (08):1834-1848.

[5]. HAN Pu, WANG Dong-bo, LIU Yan-yun, SU Xin-ning. Study on the Influence of Part of Speech on Chinese and English Text Clustering[J]. Journal of Chinese Information Processing, 2013, 27 (02):65-73.

[6]. Zhao Jingsheng, Zhu Qiaoming, Zhou Guodong, Zhang Li. A Review of Research on Automatic Keyword Extraction[J].Journal of Software,2017,28(09):2431-2449.

[7]. Li Hui. Summary of Research on Word Similarity Algorithm[J]. Modern intelligence, 2015,35(04):172-177.

[8]. Pei Ying Zhang. Word Similarity Computation Based on WordNet and HowNet [J]. Applied Mechanics and Materials,2013,2491(336).

[9]. Wu Ling Ren, Jin Ju Guo. Word Similarity Algorithm Based on WordNet And HowNet[J]. Applied Mechanics and Materials,2012,1682(155).

[10]. Jun Hou, Richi Nayak. The Heterogeneous Cluster Ensemble Method Using Hubness for Clustering Text Documents[M]. Springer Berlin Heidelberg:2013-06-15.