A kind of entity recognition algorithm based on Hadoop for power big data

Jun Qi[1,*], *Weichun* Ge[2], *Zhao* Li[1], *Wei* Li[1], *Hongyu* Zhang[2], *Jinghong* Zhao[1], *Chengming* Jin[1], *Liangliang* Yu[1], *Shuo* Chen[2], *Biqi* Liu[1], and *Mingyu* Yang[1]

[1]Information and Communication Branch of Liaoning Province Electric Power Company Limited, Shenyang 110000, China
[2]Liaoning Province Electric Power Company Limited, Shenyang 110000, China

Abstract. With the coming of the era of big data, traditional entity recognition technologies have been unable to effectively finish data pre-processing due to large scale of power grid data and complex volume type features. The rising of Hadoop technologies in these years can deal with big data processings better. Therefore, this paper proposes a power big data entity recognition algorithm based on Hadoop. It applies the discretization algorithm to select higher information accuracy discrete points and put forward a discretization evaluation indicator. In the end, we finish entity recognition of the monitoring data of wind turbines on Hadoop platform. Experimental results show that the proposed algorithm performs well in terms of correctness and breakpoint number experiments and it has a good speed-up ratio. The proposed algorithm can apply to power large data entity recognition processing.

1 Introduction

Along with the advance of information and communication technology, digitization and informatization have been deeply penetrated into every aspect of our lives. Also informatization process in electric power enterprise also get rapid development.Analysis of power effective information in large data processing requirements also enhances unceasingly. How to capture the electric power big data when enterprise decision-makings happen in the era of big data grid enterprises is an important problem in the case of data pre-processing. Entity recognition has always been a key technology of data quality management research which can play a vital role in improving the quality of the data preprocessing. In the power of big data, complex data type, data inconsistent phenomenon is more common. Therefore, entity recognition technology in the power of big data also has a wider application in the future.

Power big data entity recognition accurately identify different entities belonging to the same entity name or attributes and clustering in a given data set. It makes each entity in the decision-making of power grid can be more valuable to identify. It is different from. In literature [1] the author proposes big data entity recognition algorithm based on parallel machines. This algorithm solve the problem that the same object owing different properties

* Corresponding author: qij0427@163.com

by means of "n -Gram". It achieved good results efficiently for large data entity recognition in a short period of time. There have been lots of traditional entity recognition technologies which are mainly focusing on the text in the form about the phrases or relational data. Technologies aiming at different types of data entity recognition research have just started. Literature [2] presents a two-stage associated entity recognition model which fully considers the mode characteristics of the entity and attributing characteristics. And this paper proposes an incremental algorithm of the recognition results based on iteration incremental verification and correction to ensure the accuracy of the results.

Current existing methods researching are mainly to identify the effectiveness. There are seldom studies in entity recognition efficiency of the large data oriented technology now. Most of these methods are aiming at the tuple and string. However, relationships of XML data and graph data discriminant method of unstructured data research is still with less research[3-6]. At the same time, these algorithms are lack of effective evaluation of big data entity recognition results quality theory and public test data set.

Hadoop is a kind of distributed processing of large data infrastructure platform. Its architecture is the underlying Hadoop distributed file system (HDFS) which is mainly responsible for store files on all the nodes on the Hadoop cluster. We presents a large data entity recognition algorithm based on information accuracy (ERBIA) under the background of electric big data. Firstly the algorithm calculates class attribute of similar degree distribution and the value of an attribute in discretization scheme. Then ERBIA algorithm select the information accuracy high discrete points. In the next step we propose an improve discrete evaluation index final decision and obtain results. Finally, we perform experiments for real data sets and random data to multiple sets of contrast test on the Hadoop platform. And we obtain better processing scheme effectiveness and efficiency for power big data.

2 Entity recognition discretization scheme for power big data describing

Chief problem in data processing is the expression of knowledge. In order to facilitate data integration process and improve the efficiency of data pretreatment, we adopt the contingency table for large data attribute formal definition in this paper. Each group of data partition formal definition attributes is showing as follows:

$$S = (\mathrm{U}, \mathrm{V}, \mathrm{C}, \mathrm{f}) \tag{1}$$

In the expression: $U = \{a_1, a_2 ... a_\mathrm{n}\}$ is defined as the data is a not empty finite set, and we call it attribute domain.

$\mathrm{V} = \bigcup \mathrm{V}_a (\forall a \subset \mathrm{C})$ is defined as the effective information of the range of values of the function f.

C is defined as the attribute domain and $\mathrm{C} \neq \varnothing$.

$\mathrm{f} = \{\mathrm{f}_a : \mathrm{V} \to \mathrm{V}_a\}$ presents associated list information function and f_a is information function of attribute a.

According to the above definition, power big data set S can be expressed as the attribute domain element of number N for the list. The relationship in the power of large data sets a property value. The attribute i has a value of $a_i \subset V$, and its domain is C_i. In the set S of values a_i can be expressed as $a^i(\mathrm{u}) = \{a_1^i, a_2^i ... a_n^i\}$, in which n presents the quantity of a_i.

We assume the data is set of continuous attribute a, and the continuous attributes in each has a discretization scheme R. The set of threshold value for the attribute domain is divided into an intersection zero interval $R:\{[c_0,c_1),(c_1,c_2),...,(c_{n-1},c_n]\}$. The range of values of the attribute a. We plan the values in the order and form the corresponding breakpoint set $\{c_0,c_1,...,c_n\}$. Owing to the breakpoint set and proposed corresponding discretization scheme, we can use one in two to express attribute discretization. According to the above definition of correspondence we can establish some attribute a discretization scheme D corresponding to the Table 1.

Table 1. Some attribute a corresponding discretization scheme D corresponding table.

Attribute category	Discretization intervals $[(c_0,c_1)\cdots(c_{i-1},c_i)\cdots(c_{n-1},c_n)]$					Decision attribute
a_1	q_{11}	…	q_{1i}	…	q_{1n}	d_1
⋮	⋮		⋮		⋮	⋮
a_i	q_{i1}	…	q_{ii}	…	q_{in}	d_2
⋮	⋮		⋮		⋮	⋮
a_m	q_{m1}	…	q_{mi}	…	q_{mn}	d_3

From the above definition we can see that the proposed discretization algorithm for big data sets entity recognition is essentially based on choosing appropriate continuous interval attribute sets of data. So that we can avoid the problem in traditional data entity recognition method which is usually used for single entity model features or based on the method of the single type entity attribute of the correlation in the data measured. The problem is to effectively integrate with both of them. Here comes a Hadoop platform on a big data entity recognition algorithm based on information accuracy.

3. Hadoop platform on big data entity recognition algorithm based on information accuracy

Traditional attribute discretization algorithm is mainly used for decision making in areas such as knowledge discovery and knowledge, and examining the main effect of discretization of index to be performed by information entropy. The concept of information entropy works as a measure of the amount of information and it can be more carefully for discretization intervals. Also it makes the discretization between the information more clearly. But the disadvantages of evaluation index based on information entropy is that discreting interval differentiate too elaborate lead to scale of calculate process too large although classification of the content in information contained is more concisely. Moreover the algorithm's efficiency and hardware consumption are affected. And it is not conducive to the follow-up data processing process [7-9]. Therefore, in view of the large power data attributes, in this paper we propose a big data entity recognition algorithm based on information accuracy (ERBIA) on the basis of information theory.

3.1 Definition of information accuracy

The essence of power big data in attribute discretization is to discrete demarcation points within the range of values of the attribute set. And the attribute of the domain is divided into interval. At last point with an integer value represents each division of property values.

So the first problem to solve is to study the selection of demarcation points for us. In this paper,classification point selection standard is defined as information accuracy. We assume that there is an information table S ,and information accuracy Q_i of attribute $a_i(i=1,2,3,...,n)$. We apply $Q_i=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{l_i}$ to present system information accuracy of attribute a_i in decision. When the attribute a_i values for l_i , it namely means the decision attribute a_i taken to the number of values.

It has been proved in the literature [10] that the importance of the attributes on the probability and statistics are independent each other. In the information table it can be defined as discrete points total accuracy $Q_o=\prod_{i=1}^{n}Q_i$. When the amount of data tends to infinity, we can see that all Q_i are equal, and we mark it as Q. And $Q_o=Q^n$, $Q=\sqrt[n]{Q_o}$.

3.2 Improved discretization of the evaluation index

After determining the definition of information accuracy, in this paper we put forward an improved discretization based on information entropy evaluation index. We use it to measure some attributes of the discretization scheme discreting effectively.

Traditional information entropy is defined as following:

$$H(X)=\sum_{i=1}^{n}p_i\log_2 p_i, p_i=n_i/|X| \tag{2}$$

In it $|X|$ means base of X , and n_i presents the number for instance attributes i .

On each interval d of information entropy is expressed as $H_D(X)=\sum_{i=1}^{d}p_i\log_2 p_i$ in this discretization scheme in this article. If the discrete points d can divide collection X into two subsets and points d to a collection X of information entropy can be defined as:

$$H(X)=(|X_{d-x}|(H(X_{d-x})/|X|+|X_{d+x}|(H(X_{d+x})/|X|)/\log_2(n) \tag{3}$$

For the proposed improvement discretization of evaluation index are defined as follows:

When the value of $H(X)$ is greater, it means the accuracy of the information handled by the continuous attribute discretization is higher. And there is a higher quality of divided in the discretization scheme.

In this paper, we use $\log_2(n)$ as an operator to discrete interval number limited in a reasonable range as far as possible. It is used to avoid to interval discrimination too rough or too precise.

When the range X is zero, we can conclude that all class interval distribution is even, and $H(X)$ takes the minimum value.

4 Experimental analysis

In order to validate the effectiveness of the proposed algorithm based on information accuracy of big data entity recognition, we use one company's on-line monitoring data of grid wind turbines as an example to analysis the algorithm on the aspects of breakpoint number ,the correctness and speedup ratio.

4.1 Correctness

This article selects some operation monitoring data in December 2015 of which several operating parameters are class attributes. As decision conditions, we choose six different temperature as input data of wind turbines to measure the effect of discrete. They are NCC300 temperature a_1, NCC320 temperature a_2, side semiconductor temperature a_3, environment temperature a_4, network side semiconductor a_5, and gear box bearing temperature a_6. In order to facilitate its showing, in this paper, the decision results are expressed by three kinds of coding, respectively normal with 00, qualified with 10 and unqualified with 11. There are monitoring data of attribute value from capture part in Table 2 (in Celsius).

Table 2. Part of the values of the monitoring data attribute.

S_D	a_1	a_2	a_3	a_4	a_5	a_6	d_1
x_1	25.8	27.8	32.8	21.6	30.8	22.8	00
x_2	26.3	29.3	32.9	22.6	31.6	24.1	10
x_3	27.9	29.7	33.6	23.6	32.4	24.8	00
x_4	27.1	30.6	31.8	23.4	33.1	24.1	00
x_5	30.5	33.5	40.6	26.4	34.3	29.1	11
x_6	31.2	34.5	46.1	28.6	35.6	28.0	00

In the Eclipse environment we use the algorithm of ERBIA after discretization of the attribute data and show the result in Table 3.

Table 3. Part of the values of monitoring data of attribute ERBIA discretization algorithm.

S_D	a_1	a_2	a_3	a_4	a_5	a_6	d_1
x_1	3	3	2	4	3	2	00
x_2	2	2	3	1	0	0	10
x_3	4	4	3	3	4	2	00
x_4	4	5	5	4	3	3	00
x_5	5	2	1	2	0	0	11
x_6	5	6	5	4	5	3	00

It can be seen that to the same set of data, applying ERBIA discretization processing algorithm has the same effort in the calculation with the regular one. But in conventional algorithms its adopting the integral calculation of the average algorithm can make the individual attribute evaluation get rougher deviation, and it make the decision results and the actual operation get deflection.

4.2 Breakpoint number analysis

For the same set of data we firstly use CAIM discretization processing algorithm to deal with the data. CAIM algorithm is a kind of global, static, top-down supervised discretization algorithm. The algorithm is based on maximizing the attribute correlation and minimum break point as the goal. It has the advantage of breakpoints. So we use the proposed ERBIA algorithm comparing with CAIM algorithm in terms of number of breakpoints. It can be seen from Table 4 that breakpoints of the ERBIA algorithm significantly reduced.

Table 4. Two kinds of algorithm comparing in breakpoint number.

Algorithm			Number of the breakpoints			
	a_1	a_2	a_3	a_4	a_5	a_6
CAIM	97	115	235	356	373	312
ERBIA	6	4	3	5	6	7

4.3 Speedup ratio

Speed ratio is used to measure the performance and effect of parallelization. It can be defined as in a single run time and the ratio of the running time in the cluster. This paper provides the test data set volume of 2G, respectively working in the node number of different cluster of 2,4,6,8. The experimental data is shown in Table 5.

Table 5. Speedup on different node of the cluster.

Number of nodes	2	4	6	8
Speedup ratio	1.77	3.58	4.35	5.04

It can be seen that with the increase of number of nodes, the running time significantly declines. Operation speed of the algorithm is also improved. So we can conclude that the proposed algorithm obtain a good speedup and it is well applied in big data environment.

5 Conclusion

Traditional entity recognition algorithm can only realize relationship identification, such as simple naming. With the coming of the era of power big data, problem in relationship between complex data attributes in the big data entity recognition is imminent [11-12]. We proposed ERBIA algorithm in this paper, aiming at solving the shortcomings of the existing entity recognition algorithm. This paper proposed a discretization scheme based on information accuracy, and it put forward an improved discrete evaluation index to evaluate algorithm. Finally we finished the experiment on a Hadoop cluster. Experimental results showed that the validity of the algorithm in this paper and the advantage of discrete breakpoint number and speedup ratio. Our next focus is the study of large data sets redundant and related analysis. We look forward to the preprocessing of large data sets to provide support for the final decision in the power grid.

Acknowledgement

Supported by Science and Technology Project of Liaoning Province Electric Power Company Limited(2018YF-56)

References

1. Fan Wenfei, Huai Jinpeng. Querying Big Data: Bridging Theory and Practice[J]. J. of Computer Science & Technology, (2014),05:849-69.
2. Li Mingda, Wang Hongzhi, Zhang Jiancheng, Li Jianzhong, and Gao Hong. PEIF:Parallel Entity Resolution on Big Data. Journal of Computer Research and Development, (2013). 211-20.
3. Whang S E, Garcia-Molina H. Incremental entity resolution on rules and data[J]. The VLDB Journal—The Int. J. on Very Large Data Bases, (2014), 23(1): 77-102.
4. Dasu T, Johnson T. Exploratory Data Mining[J]. Exploratory data mining and data cleaning, (2003): 17-68.
5. Ayat N, Akbarinia R, Afsarmanesh H, et al. Entity resolution for probabilistic data[J]. Information Sciences, (2014), 277: 492-511.
6. Li L, Li J, Gao H. Rule-Based Method for Entity Resolution[J]. Knowledge and Data Engineering, IEEE Transactions on, (2015), 27(1): 250-63.
7. Li L, Wang H, Gao H, et al. EIF: A Framework of Effective Entity Identification[J]. Lecture Notes in Computer Science, (2010):717-28.
8. Denk M. A Framework for Statistical Entity Identification in R[J]. Studies in Classification Data Analysis & Knowledge Organization, (2008):335-42.
9. K02pcke H, Thor A, Rahm E. Evaluation of entity resolution approaches on real-world match problems[J]. *Proceedings of the Vldb Endowment*, (2010), 3:484-93.
10. Sun Fengying. Method research based on rough set classification.[D].Jilin University,(2011).
11. Qu Zhaoyang, Chen Shuai, Yang Fan, Zhu Li. An Attribute Reducing Method for Electric Power Big Data Processing Based on Cloud Computing Technology[J]. Automation of Electric Power Systems,(2014),08:67-71.
12. Li Hui, Hu Yaogang, Tang Xianhu, Liu Zhixiang. Method for On-line Operating Conditions Assessment for a Grid-connected Wind Turbine Generator System[J]. *Proceedings of the CSEE*, (2010),33:103-09.