Website Information Extraction Method and Algorithms Based on Big Data Technology

Zhang Guoming

Nanchang Institute of Science & Technology, Nanchang 330108, China

Keywords: Data segmentation; Path algorithm; Big data technology; Web site information extraction

Abstract: Based on the rich experience of teaching practice and theoretical analysis, it scientifically explained the cognitive semantic theory in vocabulary teaching methods, and apply them to the practice, in order to test the effectiveness of new vocabulary teaching method. This paper has conducted the empirical research combined vocabulary teaching empirical analysis with questionnaire survey. The results show that, English vocabulary teaching in higher vocational colleges under the guidance of the cognitive semantics theory helps to strengthen the students' vocabulary cognitive thinking, to construct complete and systematic vocabulary semantic network, greatly stimulate students' vocabulary learning interest and initiative, improve the students' vocabulary learning efficiency, to improve students' vocabulary level in a maximum degree.

1. Introduction

Big data technology as the data management technology, has become the main mean in the extraction methods of computer information management. This paper is focus on the web site information extraction and puts forward an inaccurate method of statistical model structure of planning database to carry on the web site information extraction method. For information big data technology, it can record text address and cited as an example to have web site information extraction method. And it establishes a conditional probability distribution of the automatic extraction model of information and data, and gives input sequence, establishes the target demand curve, at last has best segmentation of the data information, to seek out web site information extraction path and the algorithm formula.

2. Characteristics of big data technology

The technology of computer database contains many advantages, it has organization, independence, ability of share, controllable redundancy and flexibility of information and data for data information, its performance is the following aspects:

2.1 The organization of information and data

The existing data in a computer database of information is not without contact, scattered and disorderly, and data files in the same computer database are with mutual relation. At the same time, these data files constitute the overall data with a certain organizational structure in accordance with a specific relationship , have the form of a certain organizational structure, especially the data and information in a collection has similarity.

2.2 The information and data sharing

Big data technology used in the computer contains an important feature is that the data exists sharing, which is an important object to establish a computer database. If the computer in the database does not have the data information sharing, the value of using these data is completely restricted, loses of the value of the information. Computer database sharing is not only for a single person, the various departments, and the whole enterprise to realize data sharing, but also enables the individual, different units and even different regions and countries to share the data resource,

  DOI: 10.25236/iwmecs.2019.074

and then to play the maximization value of the data and information..

2.3 Information data independence

Information data contains the independence characteristic means: first, the logical independence and association of physical independence of the data information; second, logical independence refers to modify information data definition, the changes of new data and the data type, when the conversion between the different data has been changed, but they do not need the modification of original procedure, namely the logical structure and whole architecture of the information database changes; third, physical independence refers to the replacement of data physical storage devices, change the physical storage locations of the data information and other access methods, change the physical structure of data information do not effect the overall logical structure of database, also will not change the application program of the data information.

2.4 Controllability of information data redundancy

If there is duplication of information data, this is called the redundancy of data: first, when each user uses their own private data information, redundancy of data often appears; second, when using the big data technology and after realizing the data sharing, it has to delete the duplicate data.

2.5 Information data flexibility

During data recording, at the same time database management system not only contains the data storage function, but also has a management function of the data information, such as input and output data, query and edit the data and others, there is a greater flexibility, the user can accord to their own conditions and requirements to build the database to have information management and web site information extraction.

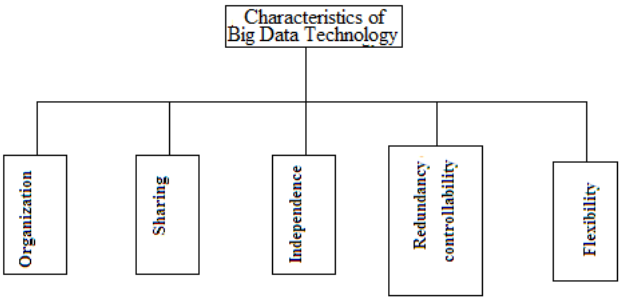

Fig.1 The characteristics of the big data technology

3. Web site information extraction method based on big data technology

For information on big data technology, we believe that unstructured data source of information is a collective independent entity, it can record text address and cited as an example to have web site information extraction method. First, records each structured and unstructured recordings in a target database table as A1, A2 ... AK. The entity row has house number, street name, city name, country name, address and list of authors. In order to prevent the data in one or more columns in unstructured recordings from the lost of the case, we assign a NULL value to it. Therefore, each entity label can appear zero or one time in unstructured recordings. And an unstructured recording can be included the information data that does not belong to any entity. In such a setting, the process of extracting the information data, each segment’s any one of the K entities A1, A2 ... AK data can be regarded as unstructured records of the split word sequence. First, it should set up the automatic extraction model.

Extract useful data from the database of information data is a widely studied topic, many models have been widely proposed. Recently proposed that a promising model can randomly extract information according to the conditions, the country's most advanced extraction method is called probability distribution of segmentation structure in half CRF input sequence.

The information data has the text input and is processed as a sequence X = X1…… XN. And the segmentation of the input sequence X is a sequence of segmentsS1…… SN, the first segment is starting from S1 to the end of the right-segment SJ. Each segment SJ has a starting position Tj , the end position Uj and a label Yj, and establishes the model formula.

First, it should establish a conditional probability distribution of half-CRF model, the given input sequence x is as follows[6]:

$$\Pr(s|X,\Lambda)=\frac{1}{Z(X)}esp(\Lambda\cdot\sum_{j}f(j,x,s))$$

$f(j,x)$ is a vector's feature functions $f_1\cdots\cdots f_n$. In the j section of the S and $\Lambda=(\lambda_1,\lambda_2\ldots\ldots\lambda_n)$ is a weight vector coding. The functional importance of each f is that $Z(X)=\sum esp(\Lambda\cdot\sum_{j}f(j,x,s)$ is the normalized factor, and depending on the token of segment attribute of the front label of a section. Therefore, a functional segment $S_j=(T_j,U_j,Y_j)$ is a function, like $f(Y_j,Y_{j-1},X,T_j,U_j)$ returns a value. It is as the following functional formula[7]:

$$\Pr(s=((T_1,U_1,Y_1),\ldots\ldots,(T_p,U_p,Y_p))|X,\Lambda)=$$
$$\frac{1}{Z(X)}esp(\sum_{j-1}^{p}\sum_{r-1}^{N}\lambda_r f_r(Y_j,Y_{j-1},X,T_j,U_j))$$

During the data information extracting process, the goal is the end of the split S=S1……Sp for input sequence X = X1…… XN, it has to create a target demand graph.

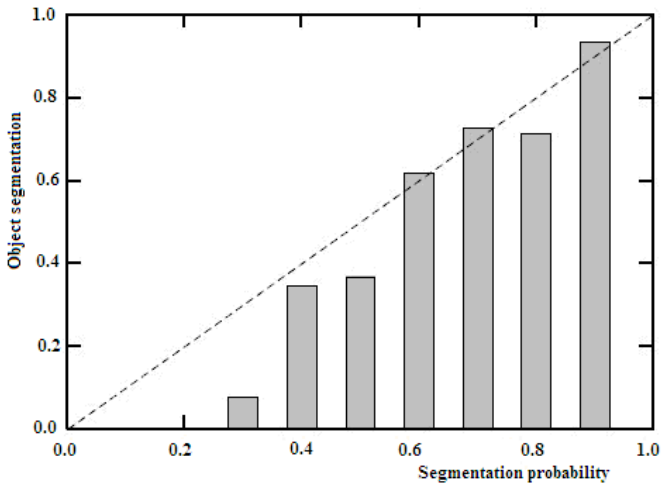

Fig.2 establishing the reliability plots of two data sets

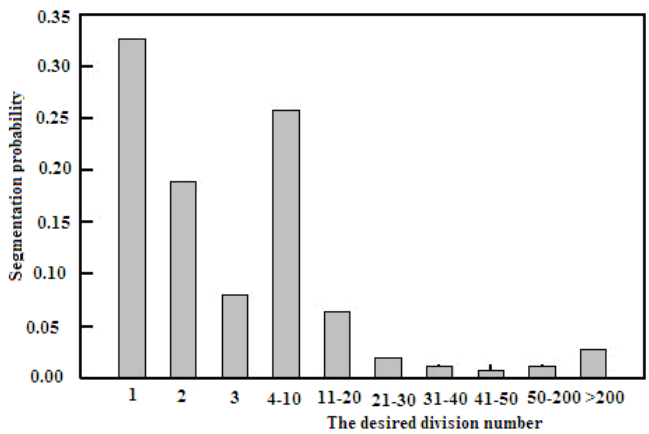

Fig 3 the division number that the quality of probability P desired and used for covering data information of the histogram

Information data model is maximized according to the data division model[9]:

$$\arg\max \Pr(s|X,\Lambda) = \arg\max \Lambda \cdot \sum_j f(Y_j, Y_{j-1}, X, T_j, U_j)$$

According to the above formula, web site information extraction path can be calculated by using dynamic programming. Let L be the maximum segment length, so that data: Y represents the partition of all parts of i from 1 (the sequence index), set an end position i of the final section. Let V (y) represent the maximum value, and the best split of the following recursive calculation is[9]:

$$V(i,y) = \begin{cases} \max Y_j, i = i-l, ..., i-1, \Lambda \cdot \sum_j f(Y_j, Y_{j-1}, X, T_j, U_j) \\ 0 \\ -\infty \end{cases}$$

The data information has the best segmentation, and then it is corresponded to the tracking path $\max V(|X|, Y)$. The best segmentation probability can be calculated. Tasks are extracted in a typical real-life, the highest score before traction is not necessarily correct extraction. The divided probability as the output of the information data can be presented by p; there is a 100% chance of correct extraction. We illustrate this is the extraction methods for these two data sets through the reliability extraction. Here we show the scores of the true accuracy computation that output by classification probability model, in this case, big data technology in web site information extraction method is more advantageous.

4. Combined with the data integration and web site information extraction method

It can base on the above data extraction method to have the data integration and segmentation technique, according to the collected data applications in free text form to have the second value calculation. The data is stored as text of PSD application, there are two main affects[10-13]:

(1) It is impossible to know all the query information data in advance. The captured text represents an intuitive user provided all the attempt of this information might be relevant. The structured data of this model covers all currently known formats and the text portion that used when there is web site information extraction method.

(2) In the traditional database management system, because restriction data of dynamically generating schema is captured as text, they only need to add a column to a major task of an existing table in the production system.

In addition, the entity described in the related information data in and the text is part of the structured data entities and relationships. IE browser combined to application program provides favorable conditions for data integration. Ideas of filling predefined templates and data integration based on web site information extraction can provide a global model used as the template. Global mode / template binding mode can be created by structured data of other metadata source and with the body. Metadata from the data source can be used to help the IE semiautomatic to assist the process of creating the required input IE module. Data integration using a system of low-level diagram based on the common data model can be extended mode to become the new entity. The template that IE process filled in will provide a new data source added to the global schema to support the new query and the data information.

5. Conclusion

Big data technology as the data management technology, has become the main mean in the extraction methods of computer information management. This paper is focus on the web site information extraction and puts forward an inaccurate method of statistical model structure of planning database to carry on the web site information extraction method. For information big data

technology, it can record text address and cited as an example to have web site information extraction method. And it establishes a conditional probability distribution of the automatic extraction model of information and data, and gives input sequence, establishes the target demand curve, at last has best segmentation of the data information, to seek out web site information extraction path and the algorithm formula.

Acknowledgement

The work was supported by the Science and Technology Research Project of Jiangxi Education Department with the project number GJJ171099 and the project name *Construction and Research of User Online Learning Behavior Analysis Model in Big Data Environment.*

References

[1] Hassan A. Sleiman, Rafael Corchuelo. A class of neural-network-based transducers for web information extraction[J].Neurocomputing, Volume 135, 5 July 2014, Pages 61-68.

[2] Giuseppe Della Penna, Daniele Magazzeni, Sergio Orefice. Visual extraction of information from web pages[J].Journal of Visual Languages & Computing, Volume 21, Issue 1, February 2010, Pages 23-32.

[3] Aditya Vyas, Urmil Kadakia, Pokhar Mal Jat. Extraction of Professional Details from Web-URLs using DeepDive[J].Procedia Computer Science, Volume 132, 2018, Pages 1602-1610.

[4] Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, Robert Baumgartner.Web data extraction, applications and techniques: A survey[J].Knowledge-Based Systems, Volume 70, November 2014, Pages 301-323.

[5] Yih-Ling Hedley, Muhammad Younas, Anne James, Mark Sanderson.Sampling, information extraction and summarisation of Hidden Web databases[J].Data & Knowledge Engineering, Volume 59, Issue 2, November 2006, Pages 213-230.

[6] Attiya Kanwal, Sahar Fazal, Aamir Iqbal Bhatti, Mukhar Ullah, Muhammad Arslan Khalid. PubMedInfo Crawler: An innovative extraction process that leads towards biological information mining[J].Meta Gene, Volume 20, June 2019, Article 100550.

[7] Agnieszka Konys. Towards Knowledge Handling in Ontology-Based Information Extraction Systems[J].Procedia Computer Science, Volume 126, 2018, Pages 2208-2218.

[8] Juan D. Velásquez.Web mining and privacy concerns: Some important legal issues to be consider before applying any data and information extraction technique in web-based environments[J].Expert Systems with Applications, Volume 40, Issue 13, 1 October 2013, Pages 5228-5239.