Hindawi
Mobile Information Systems
Volume 2021, Article ID 8115551, 10 pages
https://doi.org/10.1155/2021/8115551

Research Article

Logical Intelligent Detection Algorithm of Chinese Language Articles Based on Text Mining

Zihui Zheng

School of Foreign Languages, Anyang Institute of Technology, Anyang, Henan 45500, China

Correspondence should be addressed to Zihui Zheng; 20210026@ayit.edu.cn

Received 8 September 2021; Revised 9 November 2021; Accepted 15 November 2021; Published 16 December 2021

Academic Editor: Sikandar Ali

With the advent of the big data era and the rapid development of the Internet industry, the information processing technology of text mining has become an indispensable role in natural language processing. In our daily life, many things cannot be separated from natural language processing technology, such as machine translation, intelligent response, and semantic search. At the same time, with the development of artificial intelligence, text mining technology has gradually developed into a research hotspot. There are many ways to realize text mining. This paper mainly describes the realization of web text mining and the realization of text structure algorithm based on HTML through a variety of methods to compare the specific clustering time of web text mining. Through this comparison, we can also get which web mining is the most efficient. The use of WebKB datasets for many times in experimental comparison also reflects that Web text mining for the Chinese language logic intelligent detection algorithm provides a basis.

1. Introduction

At present, no matter at home or abroad, there are few research studies on the use of intelligent detection algorithms in detecting Chinese article logic. Text mining is one of the frontier research topics in the field of artificial intelligence; it is a computer technology that processes text based on text data and mathematical statistical analysis [1], combines machine learning and natural language processing [2], extracts information and knowledge contained in text, and selects and extracts samples from text information [3], providing users with easy-to-understand knowledge. For us, Chinese is Mandarin, which we will use in our daily life and study. For example, the words and articles written in books or paper are all Chinese, just like this passage I am writing now. If I read it out, it is actually Chinese, that is, natural language [4]. Of course, it is still Chinese without reading it. As long as people who know this language are working, studying, and living in Chinese, no one can leave it. Putonghua is the standard Chinese language in China and local dialects and national languages of all ethnic groups, that is, local Chinese languages can be translated into standard Putonghua Chinese language.

Whenever and wherever and regardless of any nation and country, our Chinese language can become a common language. There is no natural language understanding problem in interpersonal communication between people [5], but when we communicate with computers, we cannot. How should computers understand our Chinese language and realize text mining from it? At present, we can realize text mining through a series of methods such as information retrieval [6], natural language processing [7], and text information extraction [8]. Comparing web mining with traditional data text mining, we can find that the objects of web mining are distributed and a large number of web documents. Secondly, web text mining is logically a graph composed of document nodes and hyperlinks [9]. Because web documents are unstructured or meaningless, the purpose of data mining may be limited to structured data in the database and use relational tables to store structures [10] for knowledge discovery. Proper usability must be based on the preprocessing of web documents. Therefore, developing new web text mining [11] or preprocessing [12] web files to obtain document features [13] is the focus of web text mining research through three clustering algorithms of

HTML [14] to achieve the realization of logical intelligent detection algorithm and through the dataset [15] to achieve the comparison of experimental data.

Text information mining has always been a research hotspot of scholars. Text information mining based on machine learning, deep learning, and natural language processing technology has also been widely used, such as text information recognition, emotion analysis based on text information, and text similarity detection; at present, there is little research on logic detection and recognition based on text information; especially, for Chinese text, sentence is complex and has multimeaning, logic and grammar are difficult, and sentence composition is complex. How to carry out intelligent recognition of Chinese article logic through text information will be a very useful research topic.

There are few research studies on the use of intelligent detection algorithms in detecting Chinese article logic. The main contributions of this paper are as follows. We proposed a logical similarity degree and mining algorithm of web text and three HTML text clustering-based algorithm (hierarchical method, partition method, and grid method); the use of WebKB datasets for many times in experimental comparison also reflects that the proposed method for the Chinese language logic intelligent detection algorithm provides a basis.

2. Web Text Mining and Classification

Text mining is a combination of computer linguistics, statistical analysis, machine learning, and information retrieval technology. From text information to sample selection and extraction, this is a process of creating understandable knowledge. Web text mining [16, 17] is a process of discovering and extracting potentially useful and hidden information from web files and documents. It analyzes and predicts the content trend of web document set. Although web text mining is similar to flat text mining, tags in web document provide additional information to improve web text mining performance. This is the main research content of excavating teaching materials.

Web mining is a challenging subject. It realizes the web structure and rules of the web access model and realizes the dynamic search of web content. Generally speaking, web mining can be simply divided into three categories, as shown in Figure 1.

3. The Process of Text Mining

The process of text mining is shown in Figure 2. At first, it is the information source of the text, and the final result is the way users acquire knowledge.

3.1. Text Preprocessing. Text mining is the first step of text mining. The writing process may account for 80% of the whole system work.

Compared with traditional structured data, structured files are limited or unstructured. Although it is structured, it will still focus on the form of the document, rather than the content and unstructured of the document. Therefore, data mining is necessary, and it should adopt certain standards.

Most web pages on the Internet are composed of HTML documents or XML documents. In text preprocessing, the first thing to do is to use the web page information extraction module to remove the tags unrelated to text mining, then convert them into TXT text in a unified format, and store them in a folder for subsequent processing. Compared with English text preprocessing, Chinese text preprocessing will be more complicated because the Chinese original word is not a word; the amount of information of this word is relatively low, and there is no inherent separator (such as space) between words in sentences. It can be seen that Chinese texts also need entries.

3.2. Representation of Text. The Boolean logic model, vector space model, latent semantic index, and probability model are commonly used in text representation [18].

The basic idea of the vector space model is to use word bags to represent texts [19, 20]. A key assumption of this expression is that each feature word will correspond to a dimension of the feature space, independent of the order in which the sentences appear in the article. Text is represented as a vector in Euclidean space.

Its core concepts can be described as follows:

(1) Feature items: each word constitutes a document. Document = d (t_1, t_2, t_k, t_n), where t_k represents the kth feature and will be used as a dimension.

(2) Weight of feature items: in the text, each feature element is given a weight to indicate the importance of the feature element in the text.

(3) Vector space model (VSM): after clearing the order information between feature elements, the text is expressed as a vector, which is a point in feature space.

As represented by the text D, $V(d_i) = W_{i1}$, W_{i2}, W_{ik}, W_{im}).

$W_{ik} = f(t_k c_j)$ is a weight function reflecting the importance of feature t_k in determining whether document d_i belongs to class c_j.

(4) Similarity: all documents can be mapped to the space of this text vector. Document information allocation is a vector matching problem in vector positioning. In multidimensional space, the distance between points is measured by the cosine angle between vectors, that is, the similarity between documents. Assuming that the target document is U and the unknown document is V_i, the larger the included angle, the higher the similarity of documents. The similarity calculation formula is as follows:

$$\begin{aligned}\text{similarity}(V_i, U) &= \cos(V_i, U) = \frac{V_i \times U}{V_i \cdot U} \\ &= \sum_{k=1}^{m} w_{ik} \cdot w_i \sum_{k=1}^{m} w_{ik}^2 \sum_{k=1}^{m} w_k^2.\end{aligned} \tag{1}$$

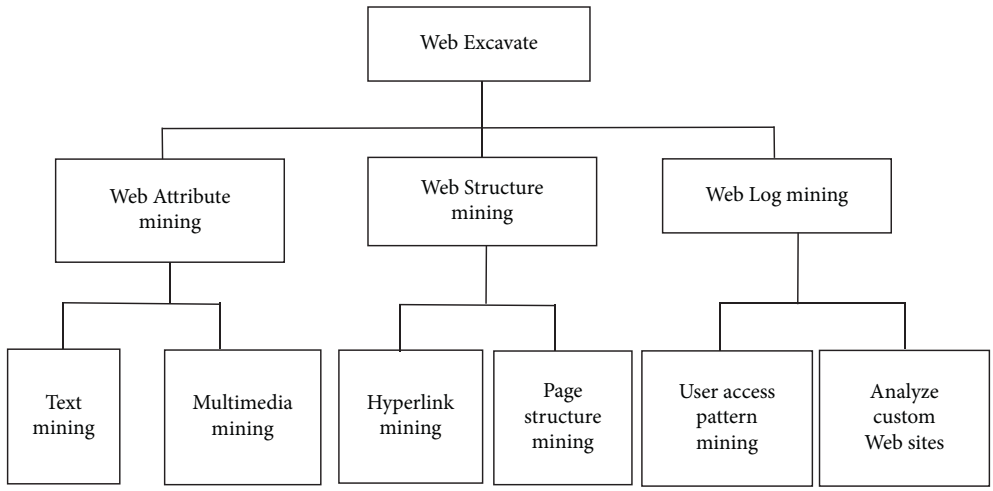

Figure 1: Classification graph of web mining.

Weights are generally functions of attribute elements displayed in documents. The frequency with which the feature t_k appears in the document d_i is represented by f_{tk} (DI), and there are several weight functions:

(1) The simplest Boolean form:

$$w_{ik} = \begin{cases} 1, & \text{if } tf_k(d_i) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The text vector consists of 0 and 1.

(2) Word frequency type:

$$w_{ik} = tf_k(d_i). \tag{3}$$

(3) Square root type:

$$w_{ik} = tf_k(d_i)^{1/2}. \tag{4}$$

(4) Logarithmic type:

$$w_{ik} = \lg(tf_k(d_i) + 1). \tag{5}$$

(5) TF-IDF formula:

$$w_{ik} = tf_k(d_i) \cdot \lg\frac{N}{N_k} + 0.5. \tag{6}$$

After normalization, it is

$$w_{ik} = \sum_{j \in d_i}^{w_{ik}} w_{ij}^2. \tag{7}$$

The purpose of normalization is to make different texts have the same length. After the text is segmented by the word segmentation program, the stop word list is first used to delete the words that are not helpful to the classification. We can also use the strategies of feature word correlation analysis, clustering, synonym, and synonym combination and finally express it as the above text vector.

3.3. Feature Set Subtraction. Feature set reduction has three goals. First, to improve the running efficiency and speed of the program; The second point is that tens of thousands of dimensional features have different importance for text classification [21]: some features held in common by all categories make little contribution to the classification required by texts and the features of other specific categories account for a relatively large proportion, while the features of other categories account for a small proportion. The third is to prevent them from overfitting. For each class, the attributes that have little contribution to classification are deleted, and the feature set reflecting the class is filtered out. An effective feature set must have the following two characteristics intuitively:

(1) Completeness: it really reflects the content of the target document

(2) Differentiation: it can clearly distinguish our target document from other documents

When using vector space method to represent documents, the dimension of text feature vector often reaches tens of thousands of dimensions. Even if the stop words in the stop word list are deleted and the low-frequency words are deleted by zip rules, tens of thousands of dimensional features will still remain. Finally, only the best feature is usually selected to execute various text mining operations through it, so it is very important to further reduce the number of features.

Generally speaking, feature subset extraction is to evaluate each feature in the set function. The method is to create an evaluation function, receive the evaluation of each feature, then sort all the functions after evaluation, and select a predefined number of best features as feature subsets. The evaluation function of text feature selection is extended from the perspective of information theory. It is used to mark each individual feature entry, which is a good reflection of the correlation between the entry and the different types.

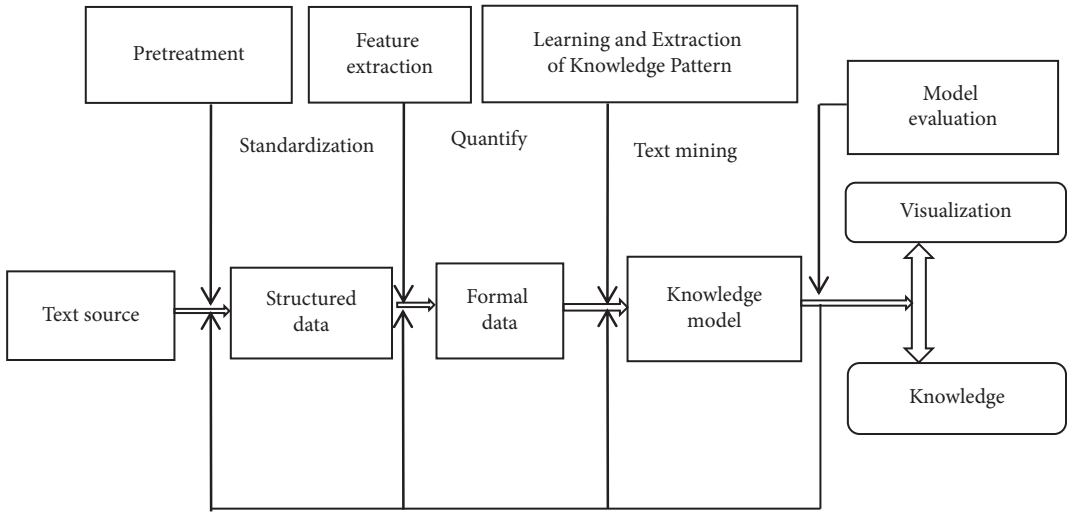

Figure 2: Schematic diagram of the text mining process.

Commonly used evaluation functions include document frequency, information gain, expected cross entropy, mutual information, word weight, text evidence weight, and probability ratio.

4. Logical Intelligent Detection Algorithm of Chinese Language Articles

Articles are written languages that reflect objective things and constitute chapters [22]. It is the product of a certain social life and is reflected in people's thoughts. In this case, the study must be inseparable from the general logic (hereinafter referred to as logic) that studies the logical form and law of thinking. In a certain sense, logic is related to the merits and demerits of an article, and logic determines the life of an article.

4.1. Logical Similarity Degree and Mining Algorithm of Web Text. The calculation of text similarity [23] and logic based on HTML structure belongs to web text mining, which depends on the logical structure feature, which is to calculate the text similarity from the nesting and path of tag structure, so as to obtain whether the logic of Chinese language is reasonable or not.

Tags are the smallest parts in HTML. Using tags to measure the similarity of web texts, the similarity is judged mainly by considering the proportion of the number of tag pairs to the total number of tags. The calculation formula is as follows:

$$\mathrm{Sim}(T_1, T_2) = \frac{N(T_1) \cap N(T_2)}{N(T_1) \cup N(T_2)}, \tag{8}$$

where T_1 and T_2 are two HTML document trees and $N(T_1)$ and $N(T_2)$ are the set of tag elements of the two document trees. In this way, the depth of tags should be combined, and the deeper the tags, the more they can reflect HTML document information.

Therefore, when calculating text similarity, each tag element should be weighted.

The path matching method reflects the similarity between HTML texts by comparing the ratio of the same path in two HTML texts to all the paths in these two HTML texts. Path refers to the sequence from root node to leaf node. Costa Gianni and others first proposed path matching and used Jaccard coefficient [24] to express text similarity. The calculation formula is as follows:

$$\mathrm{Sim}(T_1, T_2) = 1 - \frac{\left|\mathrm{Path}(T_1) \cap \mathrm{Path}(T_2)\right|}{\mathrm{Max}(\mathrm{Path}(T_1), \mathrm{Path}(T_2))}. \tag{9}$$

It can be seen from the formula that the greater the Jaccard coefficient, the smaller the similarity. However, this method has its limitations, and most of them are applied by partial matching in practical applications.

Whether it is text similarity calculation based on tag structure or tag path matching, the effect is not very ideal in the subsequent processing, but its operation efficiency is high, so it is also widely used.

Through the calculation of similarity, we can also get the value of logical reliability. The Chinese text with high similarity will have relatively high logical reliability, and at the same time, it will have the value of being mined.

4.2. HTML Text Clustering Algorithm. Text clustering algorithms are based on the hierarchical method, the partition method, and the grid method, each of which has its own advantages. At present, *K*-Means algorithm is widely used because of its simplicity and high efficiency, in order to fully consider the topic information of web text.

4.2.1. Clustering Algorithm Based on Partition. *K*-means algorithm is a basic partition algorithm [25]. On this basis, the *k*-means + algorithm is improved. This is the basic principle in the initialization process. The central point of view is that the distance between the central points of each cluster can solve this shortcoming as much as possible. In the

clustering process, the center point of each cluster is randomly initialized to a certain extent. Firstly, the algorithm randomly selects a data point ($n = 1$) as the initial point of the first cluster, then selects the initial cluster center of $n+1$ point of the data point of the first n data point, and calculates the distance between samples and the cluster center as follows:

$$d(x) = \sqrt{\sum_{i=1}^{m} (x_i - c_i)^2}. \tag{10}$$

4.2.2. Hierarchical Clustering Algorithm. The hierarchical structure of hierarchical clustering algorithm begins with a single object in the cluster; these objects are associated with other objects in the cluster, and they are located in one or more clusters in the cluster. This is an aggregate, which is a part of the Agnes hierarchy, which is the internal structure of an aggregate and produces an inverted binary tree in the process of merging. On the contrary, all objects are treated as a cluster first and then divided into multiple clusters according to a certain similarity law, and then, this partition step is repeated until each object cannot be partitioned, or a certain termination condition is reached; then, further partition is stopped. The above is the split hierarchical clustering (DIANA) process.

The HTML tag nesting contains a hierarchy, and the HTML text of a simple point can be represented by a tree diagram, as shown in Figure 3.

Web pages are represented as a subset of root-to-connection paths in the corresponding DOM (Document Object Model) tree, and Figure 3 shows the values of the four parts that need to be considered in the DOM tree. A collection of links is commonly referred to as such a DOM Root-to-link path and all the parent paths (upper level paths) that share the path. In the figure, a page is specially described through its set of links. For example, the left branch can be represented as HTML-HEAD-TITLE-title, where title is the value of the TITLE tag. Based on this, the commonly used algorithm is clustering hierarchical clustering algorithm (AHC), which merges two most similar clusters in each generation. The three most commonly used merging strategies are single link, full link, and average link, in which the distance between clusters is calculated as the nearest distance, the farthest distance, and the average distance between objects, respectively. In HTML text clustering, the partition method and the hierarchical method are used. The characteristic of structured clustering algorithm is that it defines the similarity measure for document grouping. One of the basic methods to calculate document similarity is the tag-only method, which measures the number of common tags between each pair of documents. However, for documents with little difference in the number of structural responses, this method is poor in estimating the similarity between documents. Using full link, average link, or partial methods (such as k-means), the same results are obtained in terms of clustering quality.

PathHP (HTML Pattern Path) is proposed on the basis of Apriom algorithm. In PathHP, if the characteristic f is the maximum frequent path, it is called a pattern. If the document in D contains at least minsup percentage, where minsup is the minimum support parameter defined by the user, and f is not a subpath of any other frequent path:

$$\begin{aligned} &\text{frequent}(f) \Longleftrightarrow \exists_{D'} \forall_{d' \in D'} f \in d' \wedge \frac{|D'|}{|D|} \geq \text{minsup}, \\ &\text{pattern}(f) \Longleftrightarrow \text{frequent}(f) \wedge \exists_f (\text{frequent}(f) \wedge f(\omega f)). \end{aligned} \tag{11}$$

The algorithm first sets the initial value of minsup to $1/k$ (where k is the expected number of clusters). Next, input dataset D for maximum frequent path mining is set. Before the number of paths found is greater than or equal to the expected number of clusters, minsup is divided by 2 and the mining process restarts. The resulting pattern set P is then grouped into K profiles using the complete link AHC algorithm, and the pattern similarity measure is defined as

$$\text{sim}(p1, p2) = \frac{\sum_{d \in D} \min\{m(p1, d), m(p2, d)\}}{\sum_{d \in D} m(p1, d) + m(p2, d) - \min\{m(p1, d), m(p2, d)\}}. \tag{12}$$

Finally, each document is assigned to the profile with the highest connection strength. In PathHP, we define the connection strength calculation formula of document d to meta as follows:

$$\text{str}(d, \pi_i) = \frac{\sum_{p \in \pi_i} m(p, d)}{|\pi_i|}. \tag{13}$$

Therefore, the negative influence caused by the size of π_i will be eliminated.

5. Experimental Analysis

5.1. Introduction of Dataset. WebKB data are used in the experimental dataset. The original dataset should be decompressed under Linux system, and the decompressed experimental data are shown in Table 1.

WebKB includes web page texts of computer science departments of four universities, which are Course Faculty Student class pages and so on; to develop a probabilistic, symbolic knowledge base that mirrors the content of the

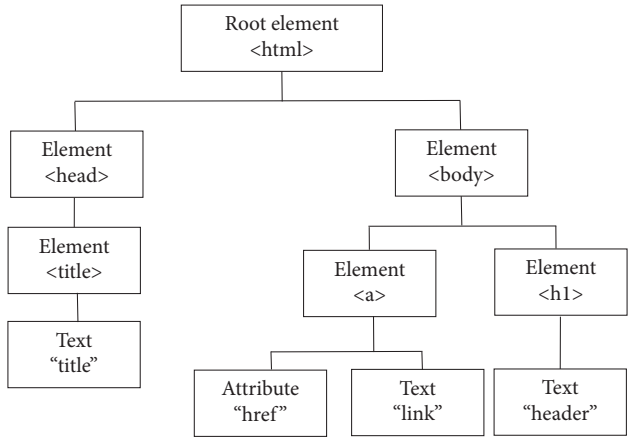

Figure 3: HTML DOM tree diagram.

Table 1: WebKB datasheet.

WebKB	Course	Department	Faculty	Project	Staff	Student	Other	Total
Cornell	44	1	34	20	21	128	619	867
Texas	38	1	46	20	3	148	571	827
Washington	77	1	31	21	10	126	939	1205
Wisconsin	85	1	42	25	12	156	942	1263
Misc	686	178	971	416	89	1080	639	4113

World Wide Web is used. This will make text information on the web available in computer-understandable form, enabling much more sophisticated information retrieval and problem solving. MISC contains pages from other universities; WebKB dataset is usually used for classification and clustering tasks. In the experiment, in order to process data effectively, it is divided into two datasets; one is to remove the web pages in misc files, a total of 4162 HTML files, and the other is to remove 8275 HTML files of WebKB as a whole (in fact, 8282 files, 7 duplicate files are removed), which are recorded as WebKB4162 and WebKB8275, respectively.

5.2. Experimental Results and Evaluation. Because clustering is not based on knowledge to judge the classification, most of the time we cannot evaluate the clustering results. Commonly used evaluation strategies are evaluating the differences of data objects in categories, such as purity and recall rate; there are also differences between evaluation classes, mainly calculating entropy. The specific performance of the evaluation strategy is shown in Figure 4.

F-Score is a clustering evaluation method used in this paper, which depends on accuracy and recall rate. Accuracy is also called accuracy, and recall rate refers to the maximum percentage of the correctly classified text in the total number of classified network texts. Recall rate refers to the percentage of correctly classified texts in the total number of online texts. These two indicators can also reflect whether the logic of Chinese articles is reasonable and clear. The accuracy precision is calculated by the following formula:

$$\text{precision} = \frac{\sum_i x_i}{\sum_i x_i + \sum_i y_i}, \quad i = 1 \ldots k, \tag{14}$$

where X is the number of web texts that are correctly clustered, y is the number of web texts that are incorrectly clustered into other classes, and k is the number of clustering clusters. The recall rate recall formula is calculated as follows:

$$\text{recall} = \frac{\sum_i x_i + \sum_i y_i}{|T|}, \quad i = 1 \ldots k, \tag{15}$$

where T is the total number of texts in the dataset. Although these evaluation methods are derived from supervised learning, they can be used when evaluating experimental results because all the web text data used contains tags with original categories. A web text may belong to both Class A and Class B. Calculate the likelihood that a document T in Class A belongs to Class B. The selected correlation calculation formula is as follows:

$$\text{Association degree}\left(T_y\right) = \frac{\sum Y/X}{|T|}, \tag{16}$$

where T is the total number of Web texts, x represents the number of categories in which web texts T_y belong, and y represents the number of similar categories of T_y. Figure 5 shows the class time of WebKB8275 dataset.

Although there is a certain gap in clustering time between knowledge base clustering and traditional vector space model, it is better than the latter in clustering accuracy and recall rate.

Time spent for text mining using an HTML clustering algorithm versus text mining without an algorithm or using another algorithm is shown in Figure 6.

WebKB4162 removes the data in misc, and only the relevant data of web texts of computer departments of Comell, Texas, Washington, and Wisconsin are retained in

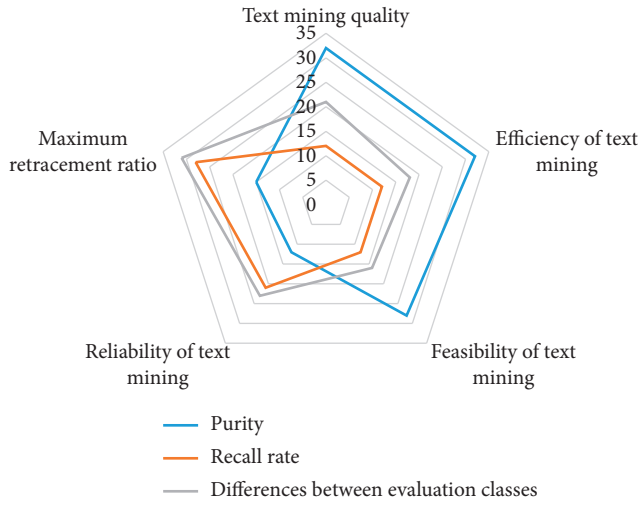

Figure 4: Specific performance of evaluation index.

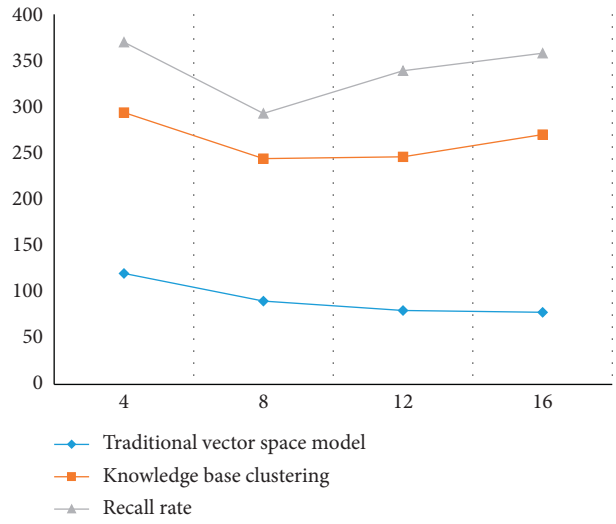

Figure 5: WebKB8275 dataset class time.

the dataset. The time comparison between knowledge base clustering and vector space model clustering is shown in Figure 7. Clustering based on knowledge base will have obvious difference in running time when K is different, but the way of the vector space model basically does not have obvious difference in running time when K is different.

Through the above process, the representation description logic reduces the dimension of web text and discovers its potential semantic relationship, which improves the efficiency of data clustering in description knowledge base and the reliability of mining Chinese language logic. Compared with the traditional vector space computing method, the knowledge representation based on description logic knowledge base has certain advantages in the accuracy of clustering problems. Because of the existence of its reasoning algorithm, it has disadvantages in running time. The specific operational efficiency of different web mining operations is shown in Figure 8.

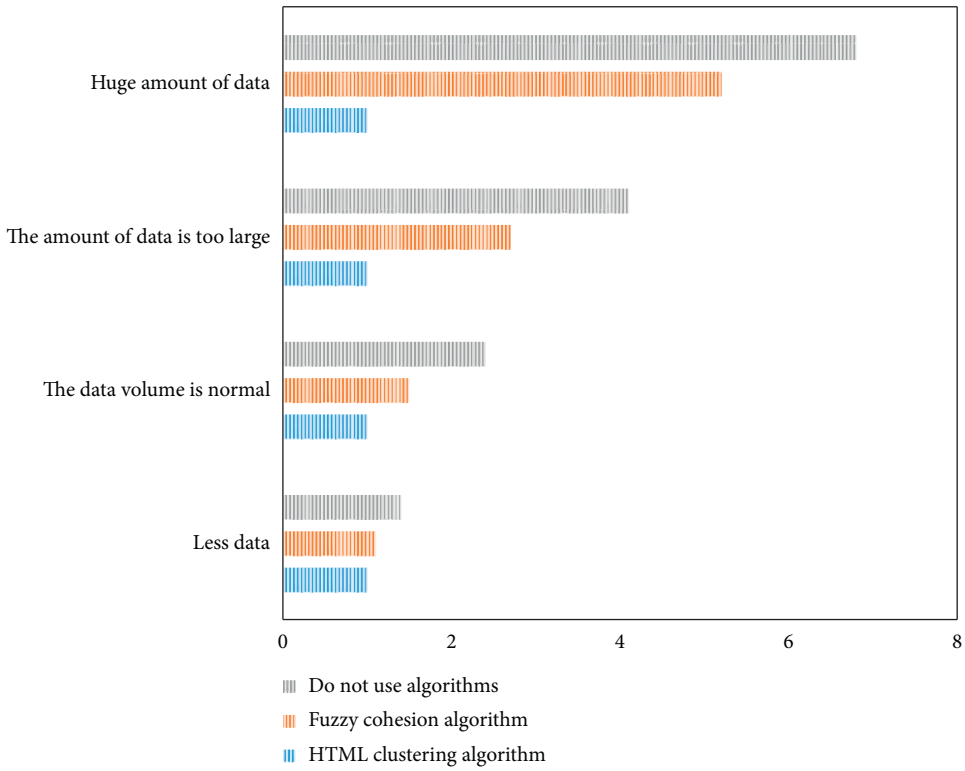

Figure 6: Time spent on text mining in various cases.

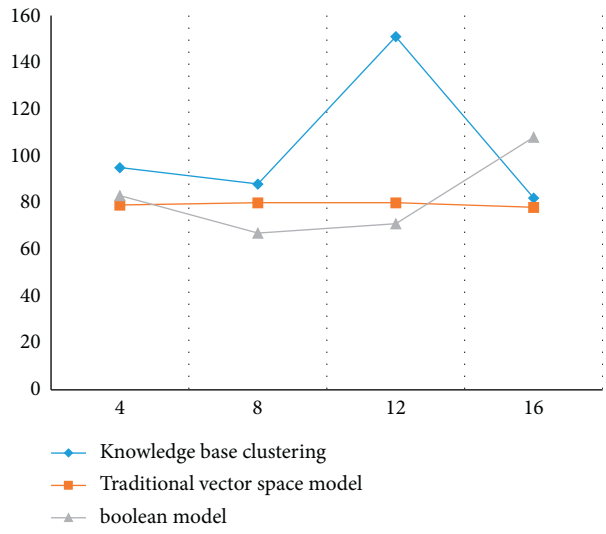

Figure 7: WebKB4162 clustering time.

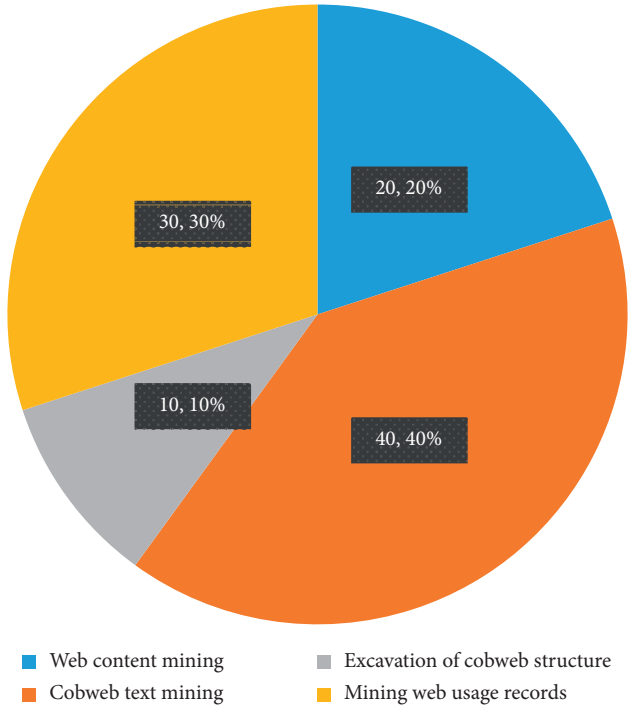

Figure 8: Percentage of different web mining operation efficiencies.

6. Conclusion

With the development of the Internet, the communication between computers and people no longer depends solely on binary. Nowadays, the ways of communication are more and more diverse, just like text mining is a technology introduced in China, and the algorithm of using text mining to realize the logical intelligent detection of Chinese language can better help us save manpower. This paper proposed a logical similarity degree and mining algorithm of Web Text and three HTML text clustering-based algorithm (hierarchical method, partition method, and grid method), and the use of WebKB datasets for many times of experimental comparison also reflects that the proposed method for the Chinese language logic intelligent detection algorithm provides a basis. This intelligent algorithm not only solves the problem of Chinese language logic but also realizes text mining. If this algorithm is used well, I think it should help us in all aspects. Data mining is a new solution in text and data mining. Chinese text mining is a complex text information system, and it is an art data mining, is the core of data mining, and is the foundation and structure of data mining. In a study on data banking, our data control mining technology is not suitable for text analysis. Data control mining technology is suitable for text analysis, which is an effective text retrieval method. In the popular research of WWW, it is a new network information retrieval method. War in network text is the key to text retrieval. The effect and running efficiency of the intelligent detection algorithm of Chinese language article logic realized by web text mining will only be better.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding this work.

References

[1] Y. Hao, Y. Huang, and Y. Feng, "Research on the application of big data technology in information statistics research system," *Journal of Physics: Conference Series*, vol. 1865, no. 4, Article ID 042112, 2021.

[2] P. Cui, *Research on the Intelligent Recognition Method of Academic Document Content Based on Text Mining*, Beijing Jiaotong University, Beijing, China, 2019.

[3] W. Xue and Y. Lu, "Research on text mining technology," *Journal of Beijing Union University*, vol. 19, no. 4, pp. 59–63, 2005.

[4] G. Yang, "Natural language understanding," *Foreign Language Teaching and Research: Foreign Languages Bimonthly*, vol. 3, pp. 26–29, 1987.

[5] X. Liu, "Summary of natural language understanding," *Statistics and Information Forum*, vol. 2, pp. 5–12, 2007.

[6] J. C. Wang, R. Xiao, Z. X. Sun, and F. Y. Zhang, "Web information retrieval research progress," *Computer Research and Development*, vol. 38, 2001.

[7] K. Liu and B. Guo, *Natural Language Processing*, Science Press, Beijing, China, 1991.

[8] X. Li and M. Zhuang, "Free text information extraction technology," *Information Science*, vol. 7, pp. 48–54, 2004.
[9] J. Wang, J. Pan, and F. Zhang, "Research on web text mining technology," *Computer Research and Development*, vol. 37, no. 5, pp. 513–520, 2000.
[10] B. Liao, J. Yu, H. Sun, and M. Nian, "Energy-saving algorithm of distributed storage system based on storage structure reconfiguration," *Computer Research and Development*, vol. 1, pp. 5–20, 2013.
[11] X. Xiao and Y. Gao, "Web text mining," *Computer Knowledge and Technology*, vol. 9, pp. 822-823, 2007.
[12] M. Liu, X. Wang, and Y. Huang, "Data preprocessing in data mining," *Computer Science*, vol. 4, pp. 56–59, 2000.
[13] F. Jiang, G. Li, and X. Yue, "Research method of document feature extraction based on semantics," *Computer Science*, vol. 43, no. 2, pp. 254–258, 2016.
[14] J.-G. Sun, J. Liu, and L. Zhao, "Clustering algorithms research," *Journal of Software*, vol. 19, no. 1, pp. 48–61, 2008.
[15] F. Chen and L. Hu, "Research and implementation of a text data integration method," *Journal of Northeast Normal University*, vol. 48, no. 1, pp. 78–83, 2016.
[16] X. Xie, Y. Fu, H. Jin, Y. L. Zhao, W. Z. Cao, and H. Jin, "A novel text mining approach for scholar information extraction from web content in Chinese," *Future Generation Computer Systems*, vol. 111, pp. 859–872, 2020.
[17] J. Schedlbauer, G. Raptis, and B. Ludwig, "Medical informatics labor market analysis using web crawling, web scraping, and text mining," *International Journal of Medical Informatics*, vol. 150, Article ID 104453, 2021.
[18] X. Wu, Z. Wu, and Y. Feng, "A text category detection and information extraction algorithm with deep learning," *Journal of Physics: Conference Series*, vol. 1982, no. 1, Article ID 012047, 2021.
[19] A. Karim and M. A. Yaqin, "Implementasi vector space model untuk meningkatkan kualitas pencarian dan penentuan derajat hadits pada kitab-kitab hadits," *Jurnal Telekomunikasi dan Komputer*, vol. 10, no. 1, pp. 1–10, 2020.
[20] C. Ke, Z. Jiang, H. Zhang, Y. Wang, and S. Zhu, "An intelligent design for remanufacturing method based on vector space model and case-based reasoning," *Journal of Cleaner Production*, vol. 277, Article ID 123269, 2020.
[21] H. Azarbonyad, M. Dehghani, M. Marx, and J. Kamps, "Learning to rank for multi-label text classification: combining different sources of information," *Natural Language Engineering*, vol. 27, no. 1, pp. 1–23, 2020.
[22] B. B. Bogomolov, V. S. Boldyrev, A. M. Zubarev, V. P. Meshalkin, and V. V. Men'shikov, "Intelligent logical information algorithm for choosing energy- and resource-efficient chemical technologies," *Theoretical Foundations of Chemical Engineering*, vol. 53, no. 5, pp. 709–718, 2019.
[23] S. Wu, F. Liu, and K. Zhang, "Short text similarity calculation based on jaccard and semantic mixture," in *Bio-Inspired Computing: Theories and Applications. BIC-TA 2020. Communications in Computer and Information Science*, L. Pan, S. Pang, T. Song, and F. Gong, Eds., vol. 1363, pp. 37–45, Springer, Berlin, Germany, 2021.
[24] E. Y. Puspaningrum, B. Nugroho, A. Setiawan, and N. Hariyanti, "Detection of text similarity for indication plagiarism using winnowing algorithm based K-gram and jaccard coefficient," *Journal of Physics: Conference Series*, vol. 1569, no. 2, pp. 1–6, 2020.
[25] A. Ahmad and S. S. Khan, "initKmix-A novel initial partition generation algorithm for clustering mixed data using *k*-means-based clustering," *Expert Systems with Applications*, vol. 167, no. 2, Article ID 114149, 2020.