Text Classification Retrieval Based on Complex Network and ICA Algorithm

Hongxia Li
Library of Dezhou University, Dezhou, China
Email: lhhongxiadzu@yeah.net

Abstract—With the development of computer science and information technology, the library is developing toward information and network. The library digital process converts the book into digital information. The high-quality preservation and management are achieved by computer technology as well as text classification techniques. It realizes knowledge appreciation. This paper introduces complex network theory in the text classification process and puts forwards the ICA semantic clustering algorithm. It realizes the independent component analysis of complex network text classification. Through the ICA clustering algorithm of independent component, it realizes character words clustering extraction of text classification. The visualization of text retrieval is improved. Finally, we make a comparative analysis of collocation algorithm and ICA clustering algorithm through text classification and keyword search experiment. The paper gives the clustering degree of algorithm and accuracy figure. Through simulation analysis, we find that ICA clustering algorithm increases by 1.2% comparing with text classification clustering degree. Accuracy can be improved by 11.1% at most. It improves the efficiency and accuracy of text classification retrieval. It also provides a theoretical reference for text retrieval classification of eBook.

Index Terms—Complex Networks, Text Classification Retrieval, ICA Clustering Algorithm, Independent Component Analysis, Collocation

I. INTRODUCTION

The modern library is developing toward information and network. The so-called information library using modern information technology to collect, organize, normative process and compress highly value images, text, voice, sound, images, video, software, scientific databases and multimedia information [1,2]. They are converted into digital information. The high-quality preservation and management are achieved by computer technology [3]. It realizes knowledge appreciation. The efficient economical transmission and reception are made by network communication technology. So that the user can get a variety of services from the net at any time and any place. Text classification is one of the key technologies for the library information of text data. Text classification has a very important significance for text information retrieval and identification. Nowadays, the most commonly used text classification is the vector space SVM model. The calculation and controllability of the model is stronger. But the method takes the word as an independent variable. It ignores the links between words, thus affects the efficiency of text classification retrieval. With the widely use of complex network data of technology in recent years, domestic and foreign scholars have carried out research on the complex network of text classification. In which, Zhu used the length of network node average path to extract the keyword. He built text classification network, but there was deficiency in the algorithm and the extraction of keyword was not clustered [4]. Liu used statistical methods of complex networks to build a similarity network classification toward the keywords meaning of China National Knowledge Infrastructure. But the similarity of the classification is not obvious [5]. Huang used the collocations to build the complex network of the words relationship to extract keywords. Peng Zhao used the network nodes and clustering coefficient to extract keywords, but the efficiency and accuracy of the algorithm needs to be improved. These uncertain information need to rely on fuzzy sets and rough sets theory as analysis tool, especially all kinds of complicated data. Rough set and fuzzy set make up non precise information to express two important drawbacks, which are respectively indiscernible and ambiguous. The former is the essential attribute of things; the latter is a classification problem. Fuzzy set theory is proposed by Zadeh in 1965, it has been proved practical in chemistry and other disciplines. In contrast, Pawlak introduced the rough set theory in 1985, although its theory is very popular in many disciplines, it isn’t mentioned in the chemical. This also shows that both the rough set theory and the other set theory have essential difference.

The traditional set theory, such as fuzzy sets, the elements in the collection can be clearly expressed. Using membership function describes common elements and collection property relations, property relations can be with or without. The definition of the membership function does not take into account the elements’ uncertain problems in the collection, in order to deal with uncertainty, fuzzy sets are proposed. The fuzzy set membership function, its value can be from the closed interval of 0 to 1, and allowing the segment. The fuzzy set membership function describes the events on the extent to occurs, rather than whether it occurs.

This paper introduced complex network theory in the text classification process. The classification and extraction of text classification keywords carried out by
network-weighted. The text classification aggregation was build through semantic links between words. At last, we introduced the ICA clustering algorithm. It made up for the deficiencies that the keyword can not achieve clustering [6]. The clustering divided of the keywords network node was carried out. It realized the complex network clustering classification of text classification successfully.

II. RELATED WORK

When extracted Chinese collocation words, Ruifeng Xu introduced the network-perception algorithms. Through network-perception algorithm, it will combine optimization discrimination Eigen values. Different types of collocation will be collocated at different stages respectively [7]. But this collocation relationship was not necessarily linear relationship.

As is shown in Figure 1, the abscissa and ordinate represent the words collocation diagram that we get by using different statistical methods. The hollow point and solid point represent collocation and non-collocation respectively [8]. We can see from the Figure that the collocation words concentrate in the left part. So we can find the boundary of collocation and non-collocation to achieve linear collocation of the words.

The process of text classification needs to do collocation of text classification search keyword. The clustering point is used to form a complex network. Figure 2 show that the cluster has a clear dividing boundary in two-dimensional space. This means that there are two distinct clusters in two-dimensional space. If you simply follow the direction of maximum variance to project the data on longitudinal axis, you cannot separate these clustering data. So that the maximum variance cannot be true response the true semantics of text classification keywords. Therefore, it can show the authenticity of the words semantics through the clustering data rotation, and then projected onto the axis [9]. The paper put forwards ICA algorithm which is the method of rotating space. We don’t need to use the orthogonally of the two-dimensional space and reveal the semantic relationship of keywords.

III. PROPOSED SCHEME

In the massive process of data mining, some of the data are often vague and cannot be classified. These data cannot be counted in a collection and cannot be existed in subset and the complement of a subset, but it can be counted in the boundary of the set. The study of rough set can use the similarity matrix which can be expressed as follows:

\[
\begin{bmatrix}
1 & a_{12} & \cdots & a_{1j} \\
\vdots & \vdots & \ddots & \vdots \\
1 & a_{j1} & \cdots & a_{jj}
\end{bmatrix}
\]  

(1)

In formula (1), \(a_{ji}\) is the value of object \(j\) according to a certain degree of similarity of the object \(i\). If the value of \(a_{ji}\) is larger, the similarity between two objects is smaller.

The use of rough sets can be automatically controlled to extract data. This paper presents a new control strategy FRC--fuzzy-rough control. The basic idea of this control strategy is: We can use a data recording manner to record the state and representative measures in the control process and integrate these data using rough sets. We can summarize the data integration process as follows:

- The rule one IF Condition 1 corresponds to THEN TAKE project 1;
- The rule second IF Condition 2 corresponds to THEN TAKE project 2;
- The rule third IF Condition 3 corresponds to THEN TAKE project 3.

This data processing strategy is called paradigm learning. This approach based on the coarse control and fuzzy control. Rough control has the features of simple, quick and easy. Another feature is data control algorithm comes from the data itself. Its decision-making and reasoning process are easier. It is easier than fuzzy control to inspect and operate and it is applied in a simple algorithm.

The genetic algorithm is an information data processing method based on the principle of binary data and genetic information. GA genetic test functions can be
added in the clustering algorithm for the convergence of clustering algorithm. The GA test function is shown in formula (2) [10].

\[ p(y_j) = 0.02 + \frac{1}{\sum_{j=1}^{i} (y_j - a_j)^2} \]  

\[ (2) \]

In formula (2), \( a \) is judgment matrix. Function has more than one maximum. Generally speaking, if the function value is greater than 1 it is convergence. This test method is fast and robust performance is good.

Assuming that data samples is \( b_k = (b_{k1}, b_{k2}, ..., b_{kn}) \), the distance between \( b_k \) and \( b_i \) can be defined as:

\[ d(b_k, b_j) = \sqrt{\sum_{l=1}^{n} (b_{kl} - b_{jl})^2} \]  

\[ (3) \]

The density of the data samples at the point of \( b_k \) can be expressed in formula (4).

\[ M_k = \frac{1}{\sum_{i=1}^{n} d(b_k, b_i)} \]  

\[ (4) \]

Then, we can select the next best cluster center. First, we can make the sample density distribution around more intensive. The density can be defined in formula (5).

\[ M_{pi} = \exp\left(-\frac{\sum_{j=1}^{n} d(B_j, s_i)}{M_k}\right), j = 1, 2, ..., n, j \neq i \]  

\[ (5) \]

The complex network is complex system that composed by structure distribution of a simple network system. The complex network of language is used to research language search structure of text classification through replication network structure. It increases the efficiency of text search [11]. Soie thought that the search statement of text classification reflected the basic characteristics of complex network. It includes the voice and lexical of text search statement, the syntax and semantics of text search statement.

### A. Independent Component Analysis

\[ \text{Complex Networks System B} \]

\[ \text{S} \xrightarrow{=} \text{P} \]

Figure 3. The data transmission diagram of complex network system

The independent component analysis of complex network analyzes the input text classification information search source \( P \) and e-book database system \( B \) under the circumstance of no priori known. Only complex network system output data \( S \) to estimate \( P \) and \( B \) as shown in Figure 3 and Figure 4.

Assuming that \( P \) is the m-dimension vector, and then \( P \) can be expressed as \( P = [P_1, P_2, ..., P_n]^T \).

Assuming that there is a relatively independent relationship between each component, it is shown in formula (6).

\[ Q(P) = \prod_{j=1}^{n} Q(P_j) \]  

\[ (6) \]

Text classification information search source \( P \) and complex network system \( B \) combine into n-dimension visualization vector \( S = [S_1, S_2, ..., S_n]^T \). It can be written in the form of independent component, as is shown in formula (7).

\[ S_j = \sum_{j=1}^{n} \delta_{ji} P_j, j = 1, 2, ..., m \]  

\[ (7) \]

It can be written in the form of vector, as is shown in formula (8).

\[ S = Pb \]  

\[ (8) \]

ICA algorithm is based on the vector \( S \). By solving complex network search matrix \( A \), it can represent each search component of text classification information search source \( P \) by \( z \). The estimation process is shown in Figure 4.

![Figure 4. The solution of complex network systems A](image)

In which, \( A \) is mixing matrix of m*n. \( z \) is the system output variable of search results. Generally speaking, the solution of complex network system is composed of two parts. One part is spheroidizing decomposition matrix \( C \), the other is orthogonal decomposition matrix \( D \). That is to say \( A = CD \), it is shown in Figure 5.

![Figure 5. The solution of complex network matrix decomposition](image)

### B. Text Classification Algorithm

We can see from the independent component analysis of complex networks. As for text classification retrieval of complex network, we can combine with network nodes. The nodal connections are established between the keywords of different text [11]. Thereby, the keywords and information are able to reflect the related semantics and context characteristics. Through the combination of ICA classification algorithm, this paper constructs the classification algorithm steps of network text. As is shown in the follows:

Step 1: Enter the text to be classified.

Step 2: The text classification is processed. It includes sorting out the overlapping words of text classification; the removing of auxiliary verb, the removing of conjunctions words. The identification of the sentence is based on the criteria of full stop of each sentence.
Step 3: The weight of semantically related words are created. The weight can be defined in formula (9).

\[ w_j = \log(1 + \frac{\text{Tre}(e_i, e_j)}{\text{Tre}(e_i) + \text{Tre}(e_j) - \text{Tre}(e_i, e_j)}) \] (9)

Among them, \(\text{Tre}(e_i, e_j)\) represents the co-occurrence frequency of characteristic word \((e_i, e_j)\) in the span of less than 2; \(\text{Tre}(e_i)\) and \(\text{Tre}(e_j)\) represent the total frequency text classification; \(\text{Tre}(e_i, e_j)\) represents the high frequency of classification words of associated words. So, the higher value of \(\text{Tre}(e_i)\) and \(\text{Tre}(e_j)\) indicate the greater probability of classification word alone.

Step 4: The language-related search words of complex network text classification is created. We merge the characteristic words of text classification retrieval that the frequency span is less than 2.

Step 5: The weighted degree of each node is calculated. The weighting coefficient of the nodes and clustering coefficient of nodes can take weighted normalized treatment. The construction of evaluation right function is shown in formula (10).

\[ W = \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3 \] (10)

In which, \(\alpha_i (i=1,2,3)\) is adjustable parameter, it represents weight coefficient, there is \(a_1 + a_2 + a_3 = 1\).

### TABLE I. A PROTEIN COMPONENT OF DECISION TABLE

<table>
<thead>
<tr>
<th>U</th>
<th>a1</th>
<th>a2</th>
<th>a3</th>
<th>a4</th>
<th>a5</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>X0</td>
<td>0.23</td>
<td>0.34</td>
<td>251.2</td>
<td>2.215</td>
<td>0.02</td>
<td>8.3</td>
</tr>
<tr>
<td>X1</td>
<td>0.48</td>
<td>0.51</td>
<td>396.5</td>
<td>2.966</td>
<td>0.03</td>
<td>8.2</td>
</tr>
<tr>
<td>X2</td>
<td>0.61</td>
<td>1.20</td>
<td>287.9</td>
<td>2.994</td>
<td>1.08</td>
<td>8.6</td>
</tr>
<tr>
<td>X3</td>
<td>0.45</td>
<td>1.40</td>
<td>282.9</td>
<td>2.933</td>
<td>0.11</td>
<td>10.3</td>
</tr>
<tr>
<td>X4</td>
<td>0.11</td>
<td>0.29</td>
<td>335.0</td>
<td>3.458</td>
<td>1.19</td>
<td>6.5</td>
</tr>
<tr>
<td>X5</td>
<td>0.51</td>
<td>0.76</td>
<td>311.6</td>
<td>3.243</td>
<td>1.43</td>
<td>8.8</td>
</tr>
<tr>
<td>X6</td>
<td>0.00</td>
<td>0.19</td>
<td>313.6</td>
<td>2.932</td>
<td>1.03</td>
<td>7.1</td>
</tr>
<tr>
<td>X7</td>
<td>0.15</td>
<td>0.25</td>
<td>337.2</td>
<td>3.856</td>
<td>1.06</td>
<td>7.9</td>
</tr>
<tr>
<td>X8</td>
<td>1.20</td>
<td>2.10</td>
<td>322.6</td>
<td>3.530</td>
<td>0.04</td>
<td>9.9</td>
</tr>
<tr>
<td>X9</td>
<td>1.28</td>
<td>2.00</td>
<td>324.0</td>
<td>3.518</td>
<td>0.12</td>
<td>8.7</td>
</tr>
</tbody>
</table>

Rough set theory is applied to the composition of quantitative analysis problem, to model data source that is a protein component, wherein the amino acid of 10 coded has 5 attributes: \(a_1 = \text{PIE}\) is expressed as side-chain fat-soluble; \(a_2 = \text{DGR} = \Delta G\) is said that the protein convert to water content; \(a_3 = \text{SAC}\) is surface area; \(a_4 = \text{MR}\) is molecular refraction index; \(a_5 = \text{LAM}\) is side-chain polarity.

First of all, to establish decision table, then the quantitative attributes of \(\{a_1, a_2, a_3, a_4, a_5\}\) and table decision attribute \(\{d\}\) constitute a decision table, which is shown in Table I.

The condition attributes are encoded as 3 quantitative intervals, such as using 1, 2 and 3 respectively expresses low, medium and high, all the attributes use natural number coding to quantitative interval, as shown in Table II.

### TABLE II. THE PROPERTIES OF QUANTITATIVE ANALYSIS

<table>
<thead>
<tr>
<th>Property</th>
<th>Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.115</td>
</tr>
<tr>
<td>2</td>
<td>0.115</td>
</tr>
<tr>
<td>3</td>
<td>0.54</td>
</tr>
<tr>
<td>4</td>
<td>1.195</td>
</tr>
</tbody>
</table>

All the attribute set of each molecule is upper and lower approximation, to calculate all approximate accuracy that approximately equal to 100%. Therefore, the quality of classification is also approximately 100%.

The next step of rough set analysis is to build the minimal subsets of independent attributes, to ensure the quality of classification and collection has the same effect. There are four \(d\) reductions that are given as follows:

Set \#1 = \(\{a_1, a_2\}\)

Set \#2 = \(\{a_2, a_7\}\)

Set \#3 = \(\{a_1, a_2, a_4, a_5\}\)

Set \#4 = \(\{a_2, a_3, a_4\}\)

The intersection of all \(d\) reduction is the core of the property \(d\). In this example, the \(d\) core is \(\{a_2\}\), it means the property is the most important basis for classification. In order to ensure the classification quality is not reduced, this property can’t be reduction. The base of reduction can be 2 or 3, the excess is not necessary, to remove the classification quality will not be affected, minimal reduction set is \#1 and \#2. Therefore, the initial coding of information system attribute can be reduced from 7 to 2.

This shows that based on the \(d\) core and \(d\) reduction, the relevant attributes can be further reduced. After the reduction, the information system can be viewed as a decision table. The classification accuracy of reduction set \#1 is shown in Table III.

### TABLE III. THE CLASSIFICATION ACCURACY SET \(\{a_2, a_4\}\)

<table>
<thead>
<tr>
<th>Category</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a_2)</td>
<td>1.000</td>
</tr>
<tr>
<td>(a_4)</td>
<td>0.361</td>
</tr>
</tbody>
</table>

It can be seen from the Table III, the minimum reduction set \#1 = \(\{a_1, a_2\}\) can ensure the classification accuracy that is 100%, while using separately \(a_1\) or \(a_4\) can’t guarantee the classification accuracy, and it shows that this algorithm is effective and correct.

### IV. EXPERIMENTAL RESULTS

In order to verify ICA clustering algorithm of complex network text classification that proposed in this paper, we download data from the Google Web site. And the text is taken as the raw materials of classification. The text materials includes 10 different texts, they are...
entertainment, publicity, advertising design, finance, sports, education, military, science and technology, legal, and news. And we take 600 articles from each text. In experiment, the 6000 articles are divided into 10 parts by the method of cross-over experiment. At first, we use the principle of independent component analysis to extract the characteristic words of the text independently. Then, the mathematical statistics of the text characteristic words of clustering degree is based on the principle of clustering algorithm. Finally, we get the statistical classification results through the key search word. The specific procedures are shown in Figure 6.

It can be seen from Figure 6, text classification process is divided into three steps:

Step1: The classification documents are sorted. The weighted complex networks of document Eigen values are constructed. The overlapping portions of characteristics words node are brought forward.

Step2: The algorithm classifier can be used to classify the classification documents. The real-time feedback of classification results are made in the classification process.

Step3: In order to achieve optimal results, we adjust the classification parameters.

Figure 6. Txt classification process diagram of complex network

This paper choose 6000 article in the Google website to do text classification.

The characteristics of characteristic words are extracted from the results. The ICA algorithm is used to do text classification clustering. The pseudo-code of ICA algorithm is shown below:

get the search term Linear discriminate of a set of search term location extracted online.
function get Search Term (initial Search Term location)
search Engineering Village using initial Search Term location, get top 10 received results.
identify all the Search Term entities in the received results using Lingpipe.
extract a few of Search Term of the location entities online, \( F = \{F_1, F_2, \ldots, F_n\} \in K^2 \)
given: a Search Term frequency \( j: K^2 \times K^2 \rightarrow K \)
a function for computing the mean \( l: f(k) \rightarrow K \)
select a initial centers \( n_1, n_2, \ldots, n_s \)
while stopping criterion is not true do

for all Term entities \( b_i \) do

\( b_i = \{f_i \in f(f_i, \{n_i\}) \leq f(f_i, \{n_i\})\} \)
end for

for all means \( h_i \) do

\( n_i = l(b_i) \)
end for

return \( h_1, h_2, \ldots, h_s \)
end function

As for the reliability of the algorithm, we should detect clustering degree of text classification characteristic words first. The test result of clustering quality indicator is shown in Figure 7.

Figure 7. The results of clustering quality indicators

In Figure 7, we can obtain the statistical results of Iq quality indicators by statistical of clustering quality indicators. In which, the compact degree of clustering can be seen from the right part of the figure. The longer of deep color part represents that the compact degree of clustering is higher. The degree of similarity between the classifications can be seen from the left part of the diagram. In which, the longer of the dark part represents that the degree of similarity between classifications is higher and the separability is better.

When the number of clustering is less, the compact degree of clustering is higher and the distinction between
classifications is better. But with the increase in the number of clustering, the clustering will become worse and worse and the similarity degree between classifications is growing. It is staggered between elements. The distinction of clustering effect is worse. It can be seen in Figure 8.

Figure 9. Clustering degree diagram of different algorithm

Figure 8 shows the clustering degree diagram in the case of different algorithm. In order to minimize the impact between the text classifications, the paper will make cross treatment of ten texts after classification. In which, five groups will use collocation algorithm, and the other five groups will use ICA algorithm as shown in Figure 9. The clustering degree of the ten groups’ text classification is quite different, in which the effect of ICA algorithm is better than collocation. In order to comprehensively verify the reliability of ICA clustering algorithm, it gives the accuracy figure of algorithm as shown in Figure 10.

Figure 10. The accuracy figure of different algorithms

We can see from Figure 10, the different type of algorithms have different accuracy. With the increase in the number of characteristic words, the accuracy rate is significantly decreased. ICA algorithm is obviously better than collocation algorithm. In order to display the comparison results of the two algorithms in visual, the data is tabulated in the form of table as shown in Table IV.

<table>
<thead>
<tr>
<th>Algorithm number</th>
<th>Collocation clustering (%)</th>
<th>Collocation accuracy (%)</th>
<th>ICA algorithm clustering (%)</th>
<th>ICA algorithm accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97.2</td>
<td>62.1</td>
<td>98.2</td>
<td>68.1</td>
</tr>
<tr>
<td>2</td>
<td>98.1</td>
<td>63.2</td>
<td>99.1</td>
<td>70.2</td>
</tr>
<tr>
<td>3</td>
<td>97.3</td>
<td>64.4</td>
<td>99.3</td>
<td>75.3</td>
</tr>
<tr>
<td>4</td>
<td>98.3</td>
<td>65.3</td>
<td>99.5</td>
<td>76.5</td>
</tr>
<tr>
<td>5</td>
<td>97.2</td>
<td>66.1</td>
<td>99.3</td>
<td>77.2</td>
</tr>
</tbody>
</table>

We can see from Table IV, ICA algorithm text classification search word clustering and accuracy are obviously better than collocation algorithm. In which, the highest collocation algorithm clustering and accuracy was 98.3% and 66.1%, respectively. The highest ICA algorithm clustering and accuracy was 99.5% and 77.2%, respectively. From the results of clustering and accuracy, ICA algorithm was higher than collocation algorithm. It verified the validity and reliability of ICA algorithm programming.

V. CONCLUSIONS

In this paper, we can combine the basic theory of complex network algorithms and conducted the depth analysis research of text classification retrieval theory. The experiment research of text classification effect is made by using the collocation and ICA algorithms. Firstly, this paper describes the principle of collocation algorithm and ICA algorithm in detail. And then it gives the construction of ICA method independent component and complex network in detail by the way of mathematical modeling. It gives the pseudo-code of the ICA algorithm. Finally, we obtain the comparison results of text classification algorithm clustering and accuracy through the algorithm experiment. In which, the highest collocation algorithm clustering and accuracy are 98.3% and 66.1%, respectively. The highest ICA algorithm clustering and accuracy are 99.5% and 77.2%, respectively. It is higher than collocation algorithm by 1.2% and 11.1%. ICA algorithm is higher than collocation algorithm. It verifies the validity and reliability of ICA algorithm programming.

REFERENCES


**Hongxia Li** (1976.5-), female, Han nationality, Dezhou, Shandong Province, librarian, Working at the Library of Dezhou University, 6 professional papers have been published.