Probability model of sensitive similarity measures in information retrieval

Xiaolong Gu¹,² and Jie Zhang¹

Abstract
In today’s Internet age, a lot of data is stored and used, which is very important. In people’s daily life, if these data are sorted, information retrieval technology will be used, and in information retrieval, some information retrieval inaccuracies often appear. Information retrieval model is an important framework and method for fast, complete, and accurate user information retrieval. With the rapid development of information technology, great changes have taken place in people’s production and life. Various information network technologies are widely used in people’s lives. The resulting flow of information shows explosive growth, information retrieval. User requirements are getting higher and higher. How to complete personalized information retrieval in a large amount of mixed information, so that retrieval technology can help us obtain effective retrieval results, has become a realistic problem worth exploring. In this article, the application of probability model based on sensitive similarity measure in information retrieval model is analyzed, and a similarity measure algorithm based on spectral clustering is proposed. By improving the similarity measure, the sensitivity problem of scale parameters is overcome and the retrieval precision is improved. In order to better reflect the superiority of the proposed algorithm, this article compares with ng-jordan-weiss (NJW) and deep sparse subspace clustering (DSSC) algorithms. The experimental results show that the proposed algorithm is superior to NJW and DSSC algorithms for different data sets in different evaluation indicators (Rand and F-measure).

Keywords
Information retrieval, network technology, similarity measure, probability model, spectral clustering algorithm

Date received: 1 November 2019; accepted: 22 December 2019

Introduction
With the rapid development of the information age, data processing technology has been widely used in people’s lives. As a common data processing technology,¹–³ information retrieval technology is the main way for users to query and obtain information and also the method and means to find information. The narrowly defined information retrieval refers only to information retrieval,⁴,⁵ that is, the customer takes a specific method according to the needs and uses the search tool to find out the search process of the required information from the information collection. The generalized information retrieval is a process in which information is processed, organized, and stored in a certain way, and then the relevant information is accurately found according to the specific needs of the information user. Information retrieval is the field of research in library and computer science, with the aim of providing a faster, more accurate and more complete search method. Information retrieval, especially text retrieval, has become

¹ National Research Base of Intelligent Manufacturing Service, Chongqing Technology and Business University, Chongqing, China
² College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China

Corresponding author:
Xiaolong Gu, National Research Base of Intelligent Manufacturing Service, Chongqing Technology and Business University, Chongqing 400067, China.
Email: gxl@ctbu.edu.cn
one of the most influential search tools. On the Internet, it helps people around the world easily access a variety of information at almost no cost. This informational search has provided a powerful fuel for human economic, cultural, and technological development. With the rapid development of the Internet, digital cameras, multimedia, and the popularity of the Internet, people are now increasingly immersed in online search for information, and image queries are becoming indispensable.

Today’s society is rapidly developing in the era of networking and informationization. Computer network technology has become one of the most influential technologies in the world and society. It covers a wide range of areas, covering almost all social and economic fields, and has been well applied in people’s livelihood and military. Therefore, its development and application are crucial and have become synonymous with a long period of time. In countries around the world, “machine substitution” is the mainstream trend in manufacturing. Not only that, but also widely used in education, finance, medical, transportation, security, electricity, and many other fields, reflecting the huge application advantages and market potential. Unfortunately, the robot industry should be a high-end manufacturing industry, but China’s robot industry still has not got rid of the strange circle that can only participate in the low-end field of high-end industries. At the same time, the development of the robot industry lacks top-level design, which undoubtedly affects the development of the robot industry.

Information retrieval is inseparable from network technology and is also based on network technology. However, the data often contain some sensitive elements. If these sensitive elements are directly published or shared, it will cause leakage of user privacy. Therefore, we must consider how to accurately handle sensitive data in a large number of data.

Similarity measure is a method that comprehensively assesses the similarity between two things. The closer the two things are, the greater their measure of similarity and the more alienated the two things, the smaller their measure of similarity. The similarity measure has a wide variety of methods and is generally selected according to actual problems. Commonly used similarities are correlation coefficient (measuring the proximity between variables) and similarity coefficient (measuring the proximity between samples). If the sample gives qualitative data, then measure the proximity between the samples, the matching coefficient, consistency, and so on of the available samples. To quantify things by quantitative methods, we must use quantitative methods to describe the degree of similarity between things. A thing often needs to be characterized by multiple variables. For example, if a group of sample points described by \( p \) variables are classified, each sample point can be regarded as a point in \( p \)-dimensional space. It is natural to use distance to measure the similarity between sample points.

The information retrieval model (IRM) is the use of mathematical language and tools to translate and abstract information and its processing in information retrieval into a mathematical formula. It is determined in three aspects: (1) the perspective of processing query formulas and documents, (2) the theory of dealing with query formulas and document relationships, and (3) the algorithm between query formulas and documents. The IRM uses mathematical or other language and tools to framework and method for representing and calculating the main elements of information retrieval and the degree of matching between them. Experts and scholars in related fields have been studying a more suitable search model and retrieval method. Since 2000, many experts and scholars have carried out research and research on IRMs. Information retrieval theorists have proposed a large number of IRMs. At present, models such as Boolean vector space and traditional probability are accepted widely.

We have a preliminary understanding of sensitivity metrics and probability-related knowledge. Therefore, based on the above analysis, according to the probability model and the similarity measure of sensitivity degree, a spectral clustering algorithm for improving the similarity measure is proposed, which improves the similarity measure. It overcomes the problem of sensitivity to scale parameters, improves clustering accuracy, and achieves a good clustering effect algorithm to improve the probability theory model and the accuracy of information retrieval. In order to better study the probabilistic model of sensitive similarity measure in information retrieval, this article will introduce the information retrieval, information retrieval probability, and similarity measure algorithm in detail.

**Information retrieval technology**

**Development of information retrieval technology**

Information retrieval technology has experienced the development of early information retrieval technology to computer modern retrieval technology. Before the computer retrieval technology was produced, the information retrieval technology generally experienced the development stage from the complete manual retrieval system → semi-mechanical retrieval system → electromechanical and photoelectric retrieval systems. Before the 1940s, the way of information retrieval was mainly manual retrieval, using some search tools such as books, indexes, abstracts, and so on, which were arranged by literature attributes such as classification, subject words, and authors to find the required documents. More representative such as library catalog cards and some well-known search journals such as CA, BA, SCI, IM and China’s Zhongmu, foreign orders and so on. Although the manual retrieval method is convenient, flexible, and easy to use and master, the retrieval speed is slow, the reliability is poor, the retrieval efficiency is susceptible to external influences, and the multipath and multi-angle search literature cannot be simultaneously performed. Therefore, the quality and quantity of services that are manually retrieved are inefficient. In order to eliminate these limitations, it is necessary
to develop new methods of detection, new retrieval equipment, and establish a more complete system of inspection. In this context, semi-mechanical retrieval methods and electromechanical and photoelectric retrieval methods were gradually developed in the 1950s and 1960s. The semi-mechanical method is represented by the edge perforation card method and then the overlap-to-hole card retrieval method. Their essence is a hand-checking perforation card system, which uses such a retrieval tool. Although a certain degree of multivariate search and the combination of subject concepts can be completed, the search efficiency and retrieval time are improved compared with the full manual search method. The actual retrieval speed is still not high, and the retrieval process is mainly based on manual operation. As technology advances, various mechanical searches have been developed. Things always move forward, and information retrieval methods are constantly evolving as the various movements in the retrieval system change. There is a contradiction between the vast literature and information resources and people’s specific needs. This is the fundamental contradiction in information retrieval. It promotes the development of information retrieval theory and technology methods. On the one hand, it is the “explosion” and “pollution” caused by the huge increase in knowledge and information. On the other hand, people ask for accurate, convenient, and convenient information retrieval technology and led to the information revolution.26 It is in the context of the rapid development of computer technology that computer retrieval technology is ushered in.

The United States successfully developed the world’s first electronic computer ENNIC. In 1949, the United States made the second generation of transistor computers. In 1964, IBM built a third-generation integrated circuit computer. In 1970, the fourth generation of large-scale integrated circuit computers such as IBM-370 came out. In 1971, Intel made the world’s first commercial microcomputer. In 1970, computer networks emerged, and the development of computers became a new category of technology. This was information technology and led to the information revolution.26 It is in the context of the rapid development of computer technology that computer retrieval technology is ushered in.

In the development of information retrieval, we learned that it has experienced multiple stages of development, and the trend of development is becoming more and more intelligent. Today, with the rapid development of science and technology, there are more and more information retrieval objects, including not only text information such as documents and data but also media information such as graphic images, sounds, and videos. These are the categories of information retrieval research. Nowadays, information retrieval has realized the development from network to intelligence. The object of information retrieval has been a long-term improvement from the previous closure to the present, from the previous stability and consistency to the current dynamic and wide distribution. As the Internet becomes more popular, the amount of information resources we need to face is increasing. If you want to get the information you need in the shortest amount of time, it will bring great difficulties to computer information retrieval. But with the development of technology, this is absolutely achievable. Figure 1 shows the framework of the intelligent information retrieval form.

**Principles of information retrieval probability model**

Its application is based on four related principles: the principle of related mind independence, the independence of words, the relevance of literature, and the principle of probability ordering. Based on probability theory, the model builds a probability model for documents and queries and calculates the similarity between documents and queries based on the model. The probability model is based on the distribution of question keywords in related and unrelated documents and is represented by the weight of the keywords. The query results are sorted according to the sum of the weights of the keywords that meet the question. The probability model is a model that is simple to implement and works well. It is assumed that both the document $D$ and the user query $Q$ can be represented by a binary term vector $x = (x_1, x_2, \cdots x_n)$. If the term $T_i \in D$, then $x_i = 1$, otherwise $x_i = 0$, while assuming two mutually exclusive events, such as $W_1$: The document is related to the user query and $W_2$: The document is not relevant to the user query. By calculating $P(W_1/x)$ or $P(W_2/x)$ of the document, it is possible to determine the relevance of the document to the user query. For discrete distributions, you can use the Bass formula and simplify it to get the function between the document and the user query

$$\text{sim}(D, Q) = \sum \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

where $p_i = r_i/r$ and $q_i = (f_i - r_i)/(f - r_i)$, $f$ denotes the total number of documents in the training document set. $r$ represents the number of documents in the training document set related to the user query. $F_i$ represents the number of documents containing the term $T_i$ in the training document set. $R_i$ denotes the number of documents containing the term $T_i$ in the $r$ related documents. In order to improve the description probability of the ideal result set, the system needs to interact with the user.

**Similarity measurement algorithm**

Similarity metrics use quantitative methods to classify things, and quantitative methods must be used to describe the degree of similarity between things. A thing often needs to be characterized by multiple variables. If a group of sample points described by $p$ variables are classified, each sample point can be regarded as a point in the $p$-dimensional space. One analytical method often used for similarity metrics is cluster analysis. Cluster analysis is a method of group analysis using the principle of “objects are clustered” and is a far-reaching statistical analysis method for classifying samples and indicators. The traditional
spectral clustering algorithm usually uses the Gaussian kernel function as the similarity function. Because the algorithm is very sensitive to the kernel parameters, it is difficult to determine a suitable scale parameter. In order to solve this problem, the spectral clustering algorithm is given by improving the similarity function. Spectral clustering algorithm is a new type of clustering algorithm proposed in recent years. Different from the traditional clustering algorithm, the spectral clustering algorithm obtains the optimal result by solving the optimal partition of the graph. The advantage is that it can be applied to sample space of arbitrary shape and can converge to global optimal solution. The spectral clustering algorithm is widely used in image processing, computer vision, text mining, machine learning, and other fields. Spectral clustering algorithm is also a hot spot in the field of machine learning research. The similarity function is the focus of current research on spectral clustering improvement.

The spectral clustering algorithm is based on the theory of spectral partitioning, and the data clustering is regarded as the graph partitioning problem. The essence of the graph partitioning problem is the approximation of the graph partitioning criterion. The optimal solution of graph partitioning is an non-deterministic polynomial (NP)-hard problem. Think of all the data samples as fixed vertices \( V \) in the undirected weighted graph \( G = (V, E) \) space, which can be connected by edges. The weighted edge \( E = [W_{ij}] \) is represented by the similarity between the \( i \)th vertex and the \( j \)th vertex. The similarity matrix is defined as: if \( i \neq j \), \( W_{ij} = \exp(-d(x_i, x_j)^2/\sigma^2) \); otherwise \( W_{ij} = 0 \). The similarity matrix \( W \) contains all the information needed for clustering. By segmenting the graphs composed of all the data points, the weights of the different subgraphs after the graph segmentation are as low as possible, and the edge weights in the subgraphs are as high as possible, so as to achieve the purpose of clustering. The clustering problem is solved by the multipath segmentation problem solving the undirected graph, and the original problem is transformed into the spectral decomposition of the similar matrix or Laplacian matrix.

**Improve similarity measure**

Let \( G = (V, E) \) be an undirected graph with a vertex set of \( V \) and an edge set of \( E \). Think of the data point as the vertex on the graph \( G \), giving the definition of the manifold distance between the two vertices as \( R_{ij} \).

**Definition 1.** Any two vertices \( p_0 \) and \( p_1 \) on the graph \( G \), there is a vertex sequence \( r = (p_0, p_1, \cdots, p_l) \) indicating a path of length \( l \) connecting \( p_0 \) and \( p_1 \) on the graph \( G \), where \( p_k \in V \ (0 \leq k \leq l) \) and \( (p_k, p_{k+1}) \in E \). Let \( R_{ij} \) denote the set of all reachable paths connecting the two data points \( p_i \) and \( p_j \) on the graph \( G \). The manifold distance between the vertices \( p_i \) and \( p_j \) is shown below.

![Figure 1. The framework of the intelligent information retrieval form.](image-url)
\[
LD(p_i, p_j) = \min_{p \in R_k} \sum_{k=1}^{n-1} |l_i - l_j|^{-1}(e^{\rho \text{dist}^2(p_k, p_{k+1})} - 1)
\]
where \(\text{dist}(p_k, p_{k+1})\) represents the Euclidean distance between data points \(p_k\) and \(p_{k+1}\). The scaling factor \(\rho\) (\(\rho > 1\)) is a tunable parameter.

**Definition 2.** In order to standardize the similarity value between 0 and 1, to improve the density-sensitive distance in the DSSC algorithm, the improved manifold distance measurement function is defined as follows:

\[
LS(p_i, p_j) = \frac{1}{\min_{p \in R_k} \sum_{k=1}^{n-1}(e^{\rho \text{dist}^2(p_k, p_{k+1})} - 1)} + 1
\]

**Definition 3.** Weighting factor for each data point:

\[
\omega(p_i) = \frac{\sum_{p_j \in \Delta} \sum_{p_j \in \Delta} LS_{ij}}{\max_{p_j \in \Delta} \sum_{p_j \in \Delta} LS_{ij}}
\]

**Definition 4.** Improved manifold distance similarity function:

\[
LS(p_i, p_j) = \omega(p_i) \omega(p_j) LS_{ij}
\]

Equation (5) satisfies the characteristics of nonnegative, reflexive, symmetrical, and triangular inequalities and satisfies the global consistency clustering hypothesis and the local consistency clustering hypothesis.

**Algorithm steps**

The time complexity of the algorithm is \(O(n^3)\) \((n\) is the number of data points in the data set). The algorithm steps are as follows:

- **Input:** Data set \(X\), number of clusters \(k\), number of neighbors \(k'\)
- **Output:** Division of the data set: \(C = \{C_1, C_2, \ldots, C_k\}\)

1. **Step 1:** Calculate the Euclidean distance for any two points \(x_i, x_j\) in the data set \(d_{ij} = (||x_i - x_j||^2)^{1/2}\)
2. **Step 2:** Construct a Laplacian matrix, where the diagonal matrix is \(D\) and \(LS_i\) is calculated by the formula (3). When \(i = j\), \(LS_i = 0\).
3. **Step 3:** Feature decomposition: calculating the feature vectors \(v_1, v_2, \ldots, v_k\) corresponding to the \(k\) largest eigenvalues of the matrix \(L\), and constructing the matrix \([v_1, v_2, \ldots, v_k] \in R^n \times k\).
4. **Step 4:** Normalization processing: unitize the row vector of \(V\) to get the matrix \(Y\), where \(Y_j = \frac{V_j}{\sqrt{\sum V_j^2}}\).
5. **Step 5:** Consider each line \(y_i\) of the matrix \(Y\) as a point of the \(R_k\) space, and use the k-means algorithm or another algorithm to calculate, and obtain \(k\) clustering results: \(C = \{C_1, C_2, \ldots, C_k\}\).
6. **Step 6:** If the \(i\)-th row of \(Y\) belongs to the \(j\)-th class, the original data point \(x_i\) is also classified into the \(j\)-th class.

**Experimental results and analysis**

In order to verify the effectiveness of the proposed algorithm, the NJW and DSSC algorithms were compared with the artificial data set and University of California Irvine (UCI) data set, respectively. In order to evaluate the performance of each clustering algorithm, this article uses the evaluation indicators to measure.

**Evaluation indicators**

Due to the data set in the UCI database, the number of clusters and the correct classification of each data point are known. Therefore, it is only necessary to use clustering indicators of external metrics to evaluate the effectiveness of the clustering results for these data sets. The two clustering indicators implemented in this article belong to the external measurement method. After implementing these two effectiveness indicators, the clustering results of a certain data set in the UCI database can be measured to match the preknown structure. Evaluate the pros and cons of different clustering algorithms for clustering results of the same data set. In the clustering performance evaluation method, the validity index can find the partition with the best number of clusters. In order to evaluate the correctness of the clustering results, this article gives a comparison of the Rand index and the F-measure index. These two statistics can be used to calculate how similar the two clustering results are to the expected results.

**Rand indicator.** The Rand indicator is a commonly used evaluation indicator for clustering results. It is used to measure the degree of agreement between clustering results and external standard classes of data. Each sample is either divided into the same class or different classes. Among the clustering results, the data originally belonging to the same class still belong to one class in our clustering results. Data that do not belong to the same class still do not belong to one class in our clustering results. The accuracy is equal to the ratio of the correct matching logarithm to the total matching logarithm, that is, \(RI = \frac{r + s}{q + s + r + t}\)

Let \(C\) denote the actual category information; \(K\) indicate the clustering result, where \(r\) stands for the logarithm of the same class in both \(C\) and \(K\); and \(s\) denotes the logarithm of the different classes in \(C\) and \(K\). The sum of \(r\) and \(s\) is the number of elements that are correctly divided. \(q\) denotes the logarithm of elements belonging to the same category in \(C\) but not belonging to the same category in \(K\), and \(t\) denotes the logarithm of elements belonging to the same category in \(C\) but belonging to the same category in \(K\). \(g + s + r + t\) stands for the logarithm of the elements that can be formed in the data set. The Rand index value is between 0 and 1.
The larger the Rand index value, the greater the similarity of the data in the cluster and the higher the degree of agreement between the two partitions. A value of 1 for the Rand indicator indicates that the two divisions are identical.

The Rand index, that is, a cluster structure of the data set is $C = \{C_1, C_2, \cdots, C_m\}$, and the data set is known to be divided into $P = \{P_1, P_2, \cdots, P_p\}$. Let $a$ denote if the two points belong to the same cluster in $C$, and the number of the same group in $P$. $b$ indicates whether two points belong to the same cluster in $c$ but the number of different groups in $P$. $c$ represents that if two points do not belong to the same cluster in $C$, then $P$ belongs to the same group. $d$ stands for that if two points do not belong to the same cluster in $C$ and the number of different groups in $P$, then $a + b + c + d = M$ is the maximum number of all pairs in the data set, that is, $M = N(N - 1)/2$, where $N$ is the total number of points in the data set. The degree of similarity between $C$ and $P$ can be defined by the following validity index as Rand index $R = (a + d)/M$.

**F-measure indicator.** F-measure is a combination of two indicators: Precision and Recall. In order to accurately describe the evaluation index, the number of data points in different cases is represented by variables (take the classification of Iris data sets in the UCI database as an example), as shown in Table 1.

| Actual category | Setosa | Versicolor | Virginica |
|-----------------|--------|------------|-----------|
| Setosa          | A      | B          | C         |
| Versicolor      | D      | E          | F         |
| Virginica       | G      | H          | J         |

The total F-measure is $F = \frac{1}{i} \sum |i| \times F(i)$, where $|i|$ is the number of all objects in the category $i$ (actual category).

The clustering results of the affinity propagation (AP) algorithm to the Iris data set in the UCI database are shown in Table 2.

As can be seen from Table 2, the clustering results of the AP algorithm on the Iris data set, the 50 data points actually belonging to the Setosa class, are correctly clustered into the Setosa class. There are actually 50 data points in the versicolor class, 45 are correctly clustered into Versicolor, and 5 points are incorrectly clustered into the Virginica class. There are 50 data points actually belonging to the Virginica class, 43 points are correctly clustered into the Virginica class, and 7 points are incorrectly clustered into Versicolor.

The calculation of the F-measure indicator should first calculate the accuracy and recall rates as $P(ij) = \frac{n_{ij}}{m_i}$ and $R(ij) = \frac{n_{ij}}{n_j}$, respectively. Then calculate the F-measure indicator, the indicator is $F(ij) = \frac{2 \times P(ij) \times R(ij)}{P(ij) + R(ij)}$.

In the formula, $n_i$ is the number of data samples contained in cluster $i$, and $n_j$ is the number of data samples that should belong to cluster $j$ but are incorrectly divided into cluster $i$. The larger the F-measure indicator value, the better the clustering performance.

**Artificial data set**

The experiment was carried out on the artificial data set, and then the experiment was evaluated for the Rand index. The $k$-means algorithm in all spectral clustering algorithms takes the best results from 100 iterations. It is suggested in the previous experiment that the optimal solution can be obtained in (1, 60). Therefore, the parameters of DSSC and the parameter $p$ of the algorithm in the following experimental comparisons are in the range of [2, 60] which are compared with other algorithms. The data set information is shown in Table 3. In order to make the comparison of the algorithm more obvious, the Rand evaluation index comparison chart of the three algorithms is listed below (see Figure 2), and the F-measure evaluation index is shown in Figure 3.

From the comparison diagrams of Figures 2 and 3, it can be found that the performance of DSSC is slightly worse on the three circles of 20 data sets, and the performance of NJW and the algorithm of this article are good. This shows that all the spectral clustering algorithms have obvious clustering effect on the convex data set, and whether the similarity measure method has no direct influence on the clustering effect. The method of this article is slightly better than the NJW algorithm on the Size5 data set. The DSSC performs best on the squared data set. It can be seen that the DSSC is not sensitive to parameter changes and the algorithm is unstable. The DSSC results on the squared data set are better than NJW and the algorithm in this article. The above comprehensive analysis shows that the proposed algorithm is better than DSSC. This is because the algorithm of this article can fully exploit the global characteristics of the data and can well handle outliers such as noise points, which can be better applied to the data sets of convex and manifold. Therefore, the algorithm of this article is better than NJW algorithm and DSSC algorithm.

**UCI data set**

In order to further verify the effectiveness of the proposed algorithm, the UCI data sets Glass, Wine, Iris, and Vehicle are selected. NJW, DSSC algorithm, and the algorithm of this article are compared. These data sets have category
Table 2. Chaotic matrix of AP algorithm for clustering Iris data sets.

| Actual category | The cluster center point is subscripted as 8 (actual Setosa) | The cluster center point is subscripted as 79 (actually Versicolor) | The cluster center point is subscripted as 81 (actually Versicolor) | The cluster center point is subscripted as 106 (actually Virginica) | The cluster center point is subscripted as 148 (actually Virginica) |
|----------------|-------------------------------------------------------------|---------------------------------------------------------------|---------------------------------------------------------------|---------------------------------------------------------------|---------------------------------------------------------------|
| Setosa         | 50                                                          | 0                                                            | 0                                                            | 0                                                            | 0                                                            |
| Versicolor     | 0                                                           | 27                                                           | 18                                                           | 0                                                            | 5                                                            |
| Virginica      | 0                                                           | 6                                                             | 1                                                             | 11                                                           | 32                                                           |

Table 3. Artificial data set.

| Data set     | Number of samples | Number of attributes | Number of categories |
|--------------|-------------------|----------------------|----------------------|
| Twenty       | 299               | 2                    | 3                    |
| Square1      | 1000              | 2                    | 20                   |
| Size5        | 1000              | 2                    | 4                    |
| Square4      | 1000              | 2                    | 4                    |

Table 4. UCI data set.

| Data set     | Total number of samples | Dimension | Number of categories |
|--------------|-------------------------|-----------|----------------------|
| Glass        | 214                     | 9         | 6                    |
| Wine         | 178                     | 13        | 3                    |
| Iris         | 150                     | 4         | 3                    |
| Vehicle      | 100                     | 18        | 4                    |

Figure 2. Comparison of artificial data set Rand evaluation indicators.

Figure 3. Comparison of F-measure evaluation indicators of artificial data sets.

Figure 4. UCI data set Rand evaluation indicators comparison chart.

Figure 5. Comparison of F-measure evaluation indicators for the three algorithms. Figure 5 shows a comparison of the F-measure evaluation indicators of the three algorithms. The comparison shows the best clustering effect on the data set.

It can be seen from Figures 4 and 5 that the DSSC algorithm has the largest F-measure evaluation value on the Glass data set, but the Rand evaluation value is lower than the algorithm in this article. The DSSC algorithm performance on the three data sets Iris, Wine, and Vehicle is not as good as the algorithm in this article. From the overall situation of the UCI data set, because the algorithm of this article improves the manifold distance measure, it fully exploits the various intrinsic links between data points. Therefore, the algorithm of this article can find the optimal solution under two kinds of evaluation indicators, and the algorithm is relatively stable, which is better than NJW and DSSC algorithms.

Through the comparison of the experimental results of the above artificial data set and UCI data set, the improved label experiments, and the clustering effect is more clearly contrasted with the expected effect. Table 4 lists the basic information of these four UCI data sets. Figure 4 shows a
manifold distance spectral clustering algorithm proposed in this article has achieved good clustering effect. Considering the global and local consistency, it fully reflects the data. The spatial characteristics, good robustness, stable algorithm, and good processing of outliers such as noise points make the similarity of theoretical calculations more consistent with the real situation and better clustering performance.

Conclusion
Because the similarity measure is very important for the clustering effect of spectral clustering, the traditional clustering parameter sensitivity and multi-scale problem cannot get good clustering effect. The existing DSSC algorithm is not stable enough. The spectral clustering algorithm of similarity metrics overcomes the problem of sensitivity to scale parameters by improving the similarity measure, improves the clustering accuracy, and achieves a better clustering algorithm than the DSSC algorithm. It can not only handle the data set of the convex distribution but also the data set of the manifold distribution, with good robustness and better performance. The time complexity of the algorithm in this article is $O(n^3)$, which is the same order of magnitude as NJW. How to reduce the computational complexity and better application to big data is the focus of the next phase of research.

As the amount of information in various fields continues to increase, the demand for information retrieval is increasing. Traditional information retrieval methods are gradually being replaced by intelligent information retrieval systems. Intelligent information retrieval satisfies people’s needs for information diversification and is conducive to improving the efficiency of information retrieval. The intelligent information retrieval technology based on the Semantic Web enhances the ability of computers to recognize natural language and accelerates the realization of knowledge representation and acquisition. However, in many computer information retrieval processes, due to the use of natural language indexing and retrieval, inaccurate queries may occur. Especially in the era of Internet information, search demand is gradually difficult to meet people’s growing demand for information retrieval. There are still the following problems. (1) Content problem: At present, network information resources are becoming more and more abundant, whether the retrieved content is accurate, and whether the network information resources of the query can be displayed, which is a problem. When we search for information, it is common to search for content that does not meet our requirements. Therefore, in order to increase the amount of retrieval and ensure the singularity of the query method, a lot of work needs to be done. (2) Object problems: In the process of information retrieval, the information retrieval needs of different people are different. How to classify these requirements to personalize the user’s use and also ensure accuracy, these are the objects that need improvement.

In response to the above problems, we propose corresponding countermeasures. (1) Language intelligence: The so-called “smart intelligence,” that is, when we input keywords into the information retrieval system through natural language, we can search processing and ambiguity analysis and assist the query at the knowledge level or concept level. Through the system to give us some intelligent tips, we can help us get the best results. (2) Content specific: In an information retrieval system, the ability to analyze content needs to be improved. In this process, information that is not related to the search content should be screened out. This not only makes the title and the full text a search point but also searches by sound, image, and the like.
Technology intelligence: Nowadays, some intelligent retrieval technologies have emerged in China, including not only automatic indexing, automatic summarization technology but also intelligent technology such as automatic tracking and automatic roaming. These search techniques are gradually being improved and optimized. In recent years, concepts such as "smart browsers" and "knowledge sharing agents" have been proposed. With the in-depth study of the IRM, we find that each retrieval model has its own characteristics, advantages, and deficiencies. Their development is not synchronous but complementary. In addition, many models are in the active stage of exploration and experimentation, and the development of each model is not the same due to the different scope of application. The general development trend of modern network information retrieval technology is to develop in the direction of multifunctionalization and intelligence, to adapt to the transformation of information organization from structural to unstructured, so as to meet the requirements of people's information acquisition and utilization to the utmost extent. Although search technology has developed rapidly in all aspects, there are still many problems in information retrieval technology in the network environment. For example, the object feature is automatically drawn and taken. Based on multi-similar feature indexing, query, retrieval, and other issues, the ontology theory derived from the field of knowledge engineering and artificial intelligence can well handle natural language understanding problems and language inference mechanisms. It is the hot issue of information retrieval in the current web environment. As information service personnel, we should constantly track and master the latest developments in modern information technology and should have a strong sense of technology promotion, make full use of modern information technology to carry out work, and make information services for the whole society.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Natural Science Foundation of China under grant number 51605061 and the Science and Technology Research Program of Chongqing Municipal Education Commission under grant numbers KJ1706172 and KJ1706162.

ORCID iD
Xiaolong Gu https://orcid.org/0000-0002-6225-3535

References
1. Dinh TTA, Liu R, Zhang M, et al. Untangling blockchain: a data processing view of blockchain systems. IEEE T Knowl Data Eng 2018; 30(7): 1366–1385.
2. Forsberg EM, Huan T, Rinehart D, et al. Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS online. Nat Protoc 2018; 13(4): 633.
3. Manogaran G and Lopez D. A gaussian process based big data processing framework in cluster computing environment. Clust Comput 2018; 21(1): 189–204.
4. El Mahdaouy A, El Alaoui SO, and Gaussier E. Improving Arabic information retrieval using word embedding similarities. Int J Speech Technol 2018; 21(1): 121–136.
5. Mitra B and Craswell N. An introduction to neural information retrieval. Found Trends Informat Retriev 2018; 13(1): 1–126.
6. Wang Q and Lu P. Research on application of artificial intelligence in computer network technology. Int J Pattern Recognit Artif Intell 2019; 33(05): 1959015.
7. Prakash M and Singaravel G. Haphazard, enhanced haphazard and personalised anonymisation for privacy preserving data mining on sensitive data sources. Int J Bus Intell Data Mining 2018; 13(4): 456–474.
8. Hatwood C, Alexander S, and Imsand E. Nurse researchers move to the cloud: protecting sensitive data in cloud-based storage environments. Clin Nurse Specialist 2019; 33(4): 164–166.
9. Peng X and Dai J. Approaches to single-valued Neutrosophic MADM based on MABAC, TOPSIS and new similarity measure with score function. Neural Comput Appl 2018; 29(10): 939–954.
10. Fei L, Wang H, Chen L, et al. A new vector valued similarity measure for intuitionistic fuzzy sets based on OWA operators. Iranian J Fuzzy Syst 2019; 16(3): 113–126.
11. Ly A, Marsman M, and Wagenmakers EJ. Analytic posteriors for Pearson's correlation coefficient. Stat Neerl 2018; 72(1): 4–13.
12. van Doorn J, Ly A, Marsman M, et al. Bayesian inference for Kendall's rank correlation coefficient. Am Stat 2018; 72(4): 303–308.
13. Tu B, Zhang X, Kang X, et al. Hyperspectral image classification via fusing correlation coefficient and joint sparse representation. IEEE Geosci Remote Sens Lett 2018; 15(3): 340–344.
14. Cardenas CE, McCarroll RE, Court LE, et al. Deep learning algorithm for auto-delineation of high-risk oropharyngeal clinical target volumes with built-in dice similarity coefficient parameter optimization function. Int J Radiat Oncol Biol Phys 2018; 101(2): 468–478.
15. Orlando G, Raimondi D, and Vranken WF. Auto-encoding NMR chemical shifts from their native vector space to a residue-level biophysical index. Nat Commun 2019; 10(1): 2511.
16. Yang J and Li S. An efficient information hiding method based on motion vector space encoding for HEVC. Multimedia Tools Appl 2018; 77(10): 11979–12001.
17. Paul PM and Ravi R. A collaborative reputation-based vector space model for email spam filtering. *J Comput Theoret Nanosci* 2018; 15(2): 474–479.
18. Trillos NG and Slepčev D. A variational approach to the consistency of spectral clustering. *Appl Comput Harmonic Anal* 2018; 45(2): 239–281.
19. Li Z, Nie F, Chang X, et al. Rank-constrained spectral clustering with flexible embedding. *IEEE Trans Neural Net Learn Syst* 2018; 29(12): 6073–6082.
20. Li Y, He K, Kloster K, et al. Local spectral clustering for overlapping community detection. *ACM Trans Knowl Discov Data (TKDD)* 2018; 12(2): 17.
21. Borlund P. A study of the use of simulated work task situations in interactive information retrieval evaluations: a meta-evaluation. *J Document* 2016; 72(3): 394–413.
22. Scholer F, Kelly D, and Carterette B. Information retrieval evaluation using test collections. *Informat Retriev J* 2016; 19(3): 225–229.
23. Cobărzan C, Schoeffmann K, Bailer W, et al. Interactive video search tools: a detailed analysis of the video browser showdown 2015. *Multimedia Tools Appl* 2017; 76(4): 5539–5571.
24. Bouramoul A. Contextualisation of information retrieval process and document ranking task in web search tools. *Int J Space-Based Situated Computing* 2016; 6(2): 74–89.
25. Ng KA, Greenwald E, Xu YP, et al. Implantable neurotechnologies: a review of integrated circuit neural amplifiers. *Med Biolog Eng Comput* 2016; 54(1): 45–62.
26. Cummings AS. Of sorcerers and thought leaders: marketing the information revolution in the 1960s. *Sixties* 2016; 9(1): 1–25.
27. Wukich C and Mergel I. Reusing social media information in government. *Govern Informat Quart* 2016; 33(2): 305–312.
28. Ge R, Feng J, Gu B, et al. Predicting and deterring default with social media information in peer-to-peer lending. *Manage Informat Syst* 2017; 34(2): 401–424.
29. Speicher R. Free probability theory. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 2017; 119(1): 3–30.
30. Costello F and Watts P. Explaining high conjunction fallacy rates: the probability theory plus noise account. *J Behav Dec Making* 2017; 30(2): 304–321.