Advances in Social Science, Education and Humanities Research (ASSEHR), volume 180
2018 International Conference on Education Science and Social Development (ESSD 2018)

Research Status and Development Trend of Chinese Terminology Set Matching

Huimin Hou, Cheng Dong, Yunliang Zhang*
Institute of Scientific and Technical Information of China
Key Laboratory of Organization and Knowledge Service for Rich Media Digital Publishing
Beijing, China
[a]houhm2016@istic.an, [b]dongc@istic.ac.cn, [c]zhangyl@istic.ac.cn

***Abstract*-Term matching plays an important role in natural language processing and knowledge mining. Matching of term sets can enhance the semantic information of terms and improve the precision of term matching, which is crucial for the development of term matching. Therefore, in view of the diversity of the current terminology set matching methods, this paper classifies and compares the commonly used and classical methods in the research area through a survey, and combines the achievements and results of the existing scholars' research, which is beneficial to the future research work.**

Keywords: Term Set; Term Matching; Literature Review

I. INTRODUCTION

Terminology refers to the contractual symbol that expresses or defines a professional concept by voice or text. In our country, it is also called noun or scientific noun (that is different from the noun in Grammar) [1], the main form of terminology is a word or phrase. At the same time, terminology is an expression of the researcher's content of his research, a tool for exchange of ideas and knowledge, representing a specific domain concept, and showing a high degree of relevance within a particular subject [2]. In today's era for rapid development of information network, the efficiency and accuracy of terminology matching are constantly increasing with the increasing demands of users, playing an important role in natural language processing and knowledge mining such as information retrieval and machine translation.

The essence of terminology matching is the calculation of terms similarity, of which terminology set matching is an extension. Generally, the term is extracted from the text content, and the term set includes two parts: the expression and the calculation. The term set expression is the basis and premise of the term set calculation, and the term set calculation is the application and development of the term set expression. The term set expression means that a collection of terms convert text documents into various forms, for example, a vector form or a collection form. Then depending on the expression, you can choose different methods for calculating the set of terms. The research results of the term set matching method are very rich and have been applied to many fields and disciplines [3].

The data in Figure 1 are from Wanfang Data, China National Knowledge Infrastructure and the Web of Science, and the search time is December 21, 2017. Among them, the dissertations, journal articles and conference papers from 1997 to 2017, are covering more than 20 academic fields such as industrial technology, medical hygiene, mathematical sciences and chemistry, involve Harbin Institute of Technology, Zhejiang University and Chinese Academy of Sciences more than 50 institutions, as well as more than 40 experts such as Lin Hongfei, belong to the forefront of hot research areas. Moreover, the articles included in the Web of Science related to terminology matching represent the international research level, and the articles included in CNKI and Wanfang databases represent the research level in China.

As can be seen from Figure 1, in the 20 years from 1998 to 2017, many scholars have published papers on the topic research of "term matching", "terms matching" or "term set matching". As early as the 1990s, abroad began the study of terminology matching, while the domestic research started at the end of 2000, and the related research increased year by year, and the annual growth rate was stable. Since the introduction of the research direction, it has aroused wide attention and has made continuous achievements.

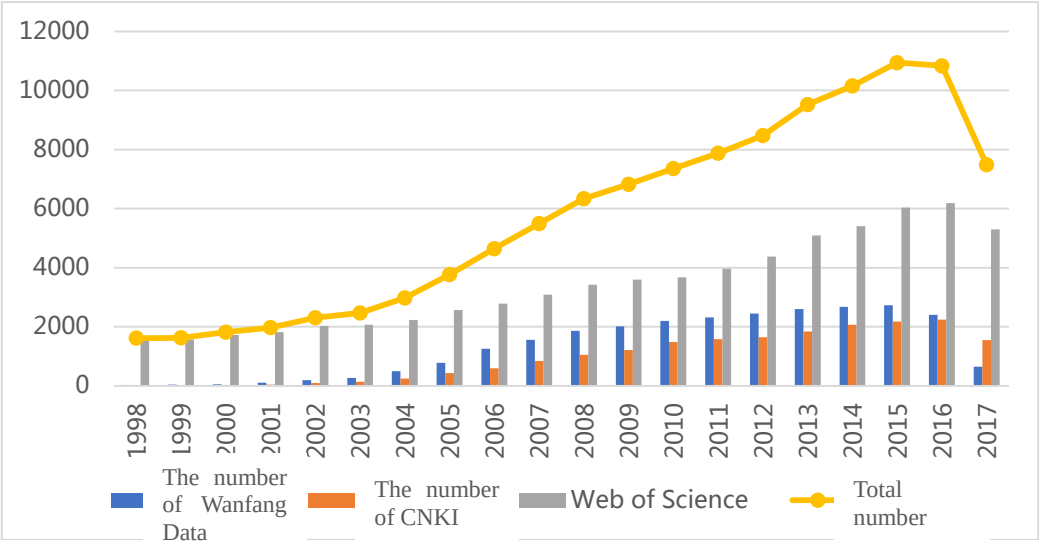

Figure 1. Terminology match the field of volume - time chart

II. TERM SET OF EXPRESSIONS

A. *Vector space model*

The Vector Space Model (VSM) is one of the most widely used mathematical models [4], which has become the dominant textual representation since Salten introduced intelligent system in the 1960s. Multi text information processing system has achieved better results.

The basic idea of the method is to use the bag of words to represent the original content of the text, and to represent the text as a multi-dimensional vector. From a geometric point of view, it is a point in a multi-dimensional space. A set of multiple texts can be represented by one Term-by-document Matrix, a collection of points in space [5].

For any given feature, we also need to calculate the corresponding feature weights (Characteristic Function) and convert the features to numerical representations. In practice, the commonly used methods of characteristic function include Boolean function, square root function, logarithmic function, TF-IDF function, etc. Among them, TF-IDF is most commonly used, using the "cosine" distance or any other distance function to measure the similarity between two documents [6].

B. *Generalized vector space model*

The traditional vector space model does not consider the relationship between the various words in the text, which considers the keywords that make up the document are the simple relationships of mutual orthogonality. However, in practice, there is a certain correlation between keywords, which is not completely orthogonal. Based on this observation, the Generalized Vector Space Model (GVSM) defines the terms through the appearance of patterns in multiple documents. The basic idea is to deduce the correlation relationship between the keywords by inferring the correlation between keywords through the simultaneous occurrence of keywords [7].

Given a set of documents consisting of t keyword, and assuming that the co-occurrence relationship between any keywords is only represented by 0 and 1, that is, 0 means non-simultaneous occurrences and 1 means simultaneous occurrences, the possible co-occurrence relationships between all the keywords can be represented by 2^t minimum:

Min_1= (0, 0, …, 0); Min_2= (1, 0, …, 0); Min_3= (0, 1, …, 0); Min_{2^t}= (1, 1, …, 1)

According to the basic concept of vector, these 2t different items can be clearly expressed by a set of standard orthonormal basis vectors:

M_1= (1, 0, …, 0, 0); M_2= (0, 1, …, 0, 0); M_{2^t}= (0, 0, …, 0, 1)

M_i denotes the i-th minimum, and $M_i M_j$=0 for all i ≠ j. Therefore, the set of vectors intersect each other. We use this set of vectors as the orthonormal basis of the generalized vector space model, and the keywords are related by the vector Mi. With the vector set associated with the smallest term, we can further determine the new keyword vector Ki in that space.

After obtaining a new vector representing a keyword, we can use this vector to represent the document and then calculate the similarity between the two documents, and make the similarity between the documents calculated as follows:

$$Sim(d_i, d_j) = \frac{\vec{d_i} \cdot \vec{d_j}}{|\vec{d_i}| \cdot |\vec{d_j}|} \quad (1)$$

C. *Latent Semantic Indexing*

Latent Semantic Indexing (LSI) is a new information retrieval model proposed by M.W. Berry and S.T. Dumais in the late 1980s. LSI replaces the conceptual matching between documents in the form of words or between documents and query strings with the concept of statistical learning, and finds that the words have some intrinsic or superficial structure. The hidden relation between word patterns and structures associated with documents is

estimated by the Singular Value Decomposition (SVD) method.

The word frequency matrix is usually very sparse, and the word frequency matrix G and its transpose matrix have certain special meanings. Among them, $G^T * G$ represents the correlation matrix between document and document, while $G * G^T$ represents the word between word matrix. LSI decomposes the word frequency matrix G into $G = U \cdot S \cdot V^T$. Finally, the appropriate k value is selected, and the largest k singular values in S and their corresponding ranks are stored, t and other singular values and their ranks are deleted; then taking U, V front of the k column vector to build U_k and V_k respectively, thereby obtaining $G_k = U_k \cdot S_k \cdot {V_k}^T$.

The new matrix G_k is a k-rank approximation matrix of G, which is the optimal approximation matrix closest to the original matrix in the sense of least squares. G_k contains the main structure information of G, ignoring the noise data of term usage, and approximate terms, such as synonyms, are merged as the dimension decreases, with similar representations in the new smaller k-dimensional space.

In addition, the choice of k has certain requirements: first, k must be large enough to adequately represent the conceptual information of the document; second, k must be small enough to filter out all irrelevant description details. With the new k-dimensional space G_k after dimension reduction, the relation between any two texts can be derived from $G^T * G$ matrix:

$$\begin{aligned} {G^T}_k &= \left(U_k \cdot S_k \cdot {V_k}^T\right)^T \cdot U_k \cdot S_k \cdot {V_k}^T \\ &= V_k \cdot S_k \cdot {U^T}_k \cdot U_k \cdot S_k \cdot {V_k}^T \\ &= V_k \cdot S_k \cdot S_k \cdot {V_k}^T = (V_k \cdot S_k) \cdot (V_k \cdot S_k)^T \end{aligned} \quad (2)$$

With the development of science and technology and the rise of artificial intelligence technology, some scholars have proposed PLSI. PLSI is a generative model of data built on top of the rules of similarity and makes full use of statistical theory to facilitate model fitting, model combination, and complexity control. Furthermore, the factor representation of model can also solve the problem of polysemy.

III. TERM SET CALCULATION METHOD

A. Distance-based term set calculation method

Distance-based term set calculation method determined the similarity between texts by calculating the distance between texts, and the greater the distance, the lower the similarity. Common distance calculation methods are Minkowski distance, Euclidean distance, KL distance and so on.

1) Minkowski distance method

First, we define two texts (di and dj), which are denoted as: $d_i = (w_{1i}, w_{2i}, \ldots, w_{mi})^T$, $d_j = (w_{1j}, w_{2j}, \ldots, w_{mj})^T$. It is the point in the eigenvector space that the term constructs, and the Minkowski distance between the documents is defined as follows:

$$\text{Minkowski}(\text{di}, \text{dj}) = \left(\sum_{k=1}^{m} \left|w_{ki} - w_{kj}\right|^p\right)^{1/p} \quad (3)$$

Minkowski distance, also known as the distance of Mingshi, is the geometric standard unit of measurement. In particular, when p = 1, the Minkowski distance becomes Manhattan distance; when p = 2, a widely used Euclidean distance is obtained.

According to the definition of Mingshi distance, the distance between two documents is not between (0, 1). In addition, the distance is inversely proportional to the degree of matching, and the greater the distance between two texts, the lower the degree of matching. Therefore, the Ming-distance-based term set matching can be implemented using the following formula:

$$\text{Sim} = \left(d_i, d_j\right) = e^{-Minkowski(d_i, d_j)^2} \quad (4)$$

2) Kullback-Leibler distance

The Kullback-Leibler (KL) distance, which is the relative entropy, measures the difference between two probability distributions P and Q, defined as follows:

$$\text{KL}(\text{P}, \text{Q}) = \sum_X P(X) log \frac{P(X)}{Q(X)} \quad (5)$$

Among them, the agreement is $0 \log \frac{0}{q} = 0$, $\text{plog} \frac{p}{0} = \infty, \forall p > 0$. It should be noted that KL (P, Q) ≠ KL (Q, P). That is, the KL distance is asymmetric, therefore, KL distance is not a true distance.

In particular, in information theory, D (P||Q) represents the loss of information produced when a real distribution P is fitted to a probability distribution Q, where P represents the true distribution and Q represents the fitted distribution of P. Furthermore, KL distance has important applications in the field of information retrieval and statistical natural language.

3) Hamming distance method

Text similarity calculation based on Hamming distance was proposed by Zhang Huanjiong et al. In 2001 [8], Hamming distance was a basic concept in information theory. It described the distance between two n-long codewords $\text{x} = (x_1, x_2, \ldots, x_n)$ and $\text{y} = (y_1, y_2, \ldots, y_n)$ as follows:

$$\text{Hamming}(\text{x}, \text{y}) = \sum_{k=1}^{n} x_k \oplus y_k \quad (6)$$

Here, $\oplus$ means modulo-2 addition. When $x_k = y_k$, $x_k \oplus y_k = 0$, otherwise, $x_k \oplus y_k = 1$. Therefore, the Hamming distance represents the number of identical positions but different symbols between two equal-length codewords, reflecting the differences between the codewords. Hamming distance is a metric in vector space of codeword length, which satisfies nonnegative, uniqueness, symmetry and trigonometric inequality.

When Hamming distance applied to the similarity calculation of text, we first need to represent the text as a sequence of code words, for example, the text d = (w1, w2, w3, w4, ...) is represented as (1, 0, 0, 1, ...) where 0 and 1 represent the occurrence of the corresponding term's position, 0 indicates that the term of the text does not appear at the current position, and 1 indicates the occurrence. After the sequence of code words corresponding to the text is obtained, the similarity calculation based on the Hamming distance of term set or text can be performed. The calculation formula is as follows:

$$\text{Sim}\left(d_i, d_j\right) = 1 - \frac{1}{n} \sum_{k=1}^{n} x_{ki} \oplus x_{kj} \quad (7)$$

Here, x_{ki} and x_{kj} denote the codeword corresponding to the kth position in the text d_i and d_j respectively, and the value is 0 or 1. It can be seen from the above formula

that the operation result of the method is between 0 and 1, and the result is 1 when the two texts are completely similar, and the result is 0 when they are completely different. The similarity function can quantitatively reflect the differences between the texts. In addition, the method only needs modulo-2 addition, and it has the advantages of convenient use and fast speed. Moreover, this method is essentially the same as the vector space model, but it needs to convert the text into a specific code sequence. This pretreatment takes a certain amount of time and affects the overall efficiency of the algorithm.

Distance-based terminology set matching methods are more widely used, which can be summarized as direct application and indirect application:

(1) Indirect application mainly refers to the use of classification and clustering algorithm. Aiming at the inaccurate judgment of cluster number and slow clustering in the consistent clustering algorithm, a new clustering algorithm based on Minkowski distance is proposed by Xu Degang, which uses Minkowski distance to characterize the similarity between samples. According to the random walk strategy, combined with different data eigenvalue distribution analysis method for clustering, the number of clusters can be automatically identified [9]; Guo Yanhui and others put forward a local binary improved algorithm for image classification for the time-consuming and large footprint LBP clustering feature extraction process. This method replaces the Euclidean distance with the Hamming distance and realizes the local binary description of multi-scale images with high accuracy and fast running speed [10].

(2) Direct application refers to the application of the distance-based term set matching method directly in the actual scenario, and the problem is often targeted. Aiming at the limitation of distance metric between intuitionistic fuzzy sets, Shen Xiaoyong proposed a measure of dissimilarity based on weighted Minkowski distance, which overcome the shortcomings of existing IFS distance measurement and solves several special measurement of distance between intuitionistic fuzzy sets [11]. Aiming at the problem that traditional particle swarm optimization is not suitable for solving discrete problems, an improved particle swarm optimization algorithm based on Hamming distance was proposed by Qiao Shen. This method retains the basic idea and flow of particle swarm optimization algorithm. Based on Hamming distance, it defines a new type of velocity representation and improves the ability for searching of the algorithm and the global search ability in the global solution space [12].

B. Set theory based on the term set calculation method

The text itself can be regarded as a collection of several features (for example, the words can be treated as features). Therefore, based on the theory of sets, similarities between different texts can be achieved by comparing the number of intersection of the elements in the feature sets of the two texts Degree measurement. Generally, the more the same number of features between two texts, the higher the similarity between them. Conversely, the less the same number of features, the lower the similarity. The common similarity calculation method based on set theory is Dice coefficient method and Jaccard coefficient method.

1) Dice coefficient method

Taking the comparison between two texts as an example, the Dice coefficient method is defined as follows:

$$Sim(d_i, d_j) = \frac{2\times\sum_{k=1}^{m} w_{ki}\cdot w_{kj}}{\sum_{k=1}^{m} w_{ki}^2+\sum_{k=1}^{m} w_{kj}^2} \quad (8)$$

When choosing a binary weight scheme, the Dice factor can be further reduced to:

$$Sim(d_i, d_j) = \frac{2\times|d_i\cap d_j|}{|d_i|+|d_j|} \quad (9)$$

Further, let T be the number of the same items contained in vectors of di and dj, ni represents the number of non-zero entries in di, nj represents the number of non-zero entries in dj, then:

$$Sim(d_i, d_j) = \frac{2\times T}{n_i+n_j} \quad (10)$$

In the Dice formula, the denominator serves as a normalization function for limiting the measure of similarity between (0-1).

2) Jaccard coefficient method

The Jaccard coefficient method is based on set theory, and uses the set calculation method to carry out the term set matching, divided into narrow Jaccard and Jaccard generalized computations. Most of the constraints based on set similarity functions can be transformed into constraints based on set intersections. The principle of Jaccard distance is that the ratio of the size of the intersection of A and B with the size of the union of A and B is the similarity value of Jaccard distance, and the size of union is the size of the two sets minus the size of their intersection. The specific formula is defined as follows:

$$J(A,B) = \frac{|A\cap B|}{|A\cup B|} = \frac{|A\cap B|}{|A|+|B|-|A\cap B|} \quad (11)$$

When A and B are both empty, J (A, B) is defined as 1.

Generalized Jaccard similarity, the value of the element can be real. Also known as Tanimoto coefficient, it is expressed by EJ and calculated as follows:

$$Ej(A,B) = \frac{A*B}{\|A\|^2+\|B\|^2-A*B} \quad (12)$$

Where A and B are represented as two vectors respectively, and each element in the set is represented as one dimension in the vector. In each dimension, the value is usually between (0, 1), and A*B represents a vector product, ||A||^2 represents the modulus of the vector. The main disadvantage of this approach is that when using multiple hash functions in combination with the Minhash algorithm, computational complexity can be high.

The term set matching method based on set theory is mainly used for the research of evidence conflict metrics and the application of recommended algorithms:

(1) In terms of evidence conflict measurement: in order to avoid the evidence of highly conflict when using the theory of uncertainty reasoning (Dempster-Shafer evidence), Ji Weihua proposed a weighted combination method based on generalized Jaccard coefficients. This method can not only improve the support of the evidence body of the similar recognition results, but also further reduce the impact of conflict evidence, so as to provide realistic conclusions. Aiming at the problem of conflict

misjudgment of traditional conflict measure and Jousselme distance measure [13], Meng Chen proposed a measure method of evidence conflict based on similarity coefficient and Jaccard coefficient matrix. This method effectively represents the degree of conflict between evidences and comprehensively and accurately measures conflict between evidence [14].

(2) The application of recommended algorithm: aiming at the problem that the traditional sparseness and similarity measure based on the neighborhood-based collaborative filtering algorithm can only use the common scoring of users, a symbiosis based on the Papaniculau and Jaccard Coefficients Filter Algorithm (CFBJ) was proposed by Yang Jiahui et al. This algorithm selects the nearest neighbor by improving the accuracy of item similarity, optimizes the preference prediction and personalized recommendation of target users, effectively alleviates the problems caused by the sparseness of user rating data and improves the recommendation system Prediction accuracy [15]. There is Jccard coefficient method to determine the similarity of the project category proposed by Li Xiaohui, so as to personalize recommendations, and achieve good results [16].

IV. CHINESE TERMINOLOGY SET MATCHING METHOD COMPARATIVE ANALYSIS

TABLE 1. CHINESE TERM SET MEANS COMPARATIVE ANALYSIS

Method Name	Principle	Advantages	Disadvantages
Vector Space Model	The bag of words is used to represent text content. Each term is regarded as an independent one-dimensional space in the feature space. The term set can be seen as a vector in the feature space.	①simple and intuitive ②fast processing	①low accuracy ②high number of dimensions semantic ③characteristics are not obvious ④did not consider the length of the text question
Generalized Vector Space Model	Incorporate the relationship between terms, the terms are related by orthogonal basis, get new term set vector	①Including the co-occurrence of words and phrases ②including certain semantic information	①high cost ②practical value small
Latent Semantic Index Model	Compresses the VSM's eigenvectors using linear projections to map document and term event matrices to potential representations	Map documents and queries to the lower-dimensional space associated with the concept	①Time and space complexity too high, difficult to parallelize ②lack of strong statistical theory

As can be seen from Table 1, the expression of the term set based on the form of a vector to expand multiple models. These models all match the term set based on the vector calculation principle. Although these methods have some limitations, to some extent, they have solved the corresponding problems through their own advantages. Many scholarly scholars have been working on relevant research, continuously improving the representation of term set matching, improving the effectiveness of term set matching for term matching calculations and matching items, and improving the system of term set matching methods.

TABLE 2. COMPARISON OF CHINESE TERMS SET CALCULATION METHOD

Method Name	Principle	Advantages	Disadvantages
Distance-based term set calculation method	Determine the degree of similarity between term sets by distance. The greater the distance, the lower the similarity.	Facilitate understanding	Having some limitations and cannot fully characterize textual information.
Set theory based term set calculation method	By comparing the number of intersections of elements in a feature set between term sets to similarity measures for different texts	①Simple implementation ②Speed fast	①lack of semantics ②low accuracy

It can be seen from Table 2 that the methods of calculating for the terms set mainly include the distance-based method and the set-based method, each of which has advantages and disadvantages and each has its own purpose. The distance-based term set matching method is widely used in the classification and clustering algorithm, more practical problems in the application of the solution; based on set theory terminology set matching method is generally in evidence conflict research and algorithm recommended aspects of application. Two kinds of methods are complementary to each other.

V.SUMMARY OF PROSPECTS

Terminology set matching has a very wide range of applications, and is also one of more important research topics. It plays an important impact for the development of certain areas. This article introduces many approaches to Chinese term sets and analyzes the advantages and disadvantages of the commonly used term set matching methods. It can be seen from the above analysis that the related methods of term set matching have been relatively mature and the system is relatively perfect. However, there are still many new research directions and research hot spots, which are summarized as follows:

(1) The integration of knowledge maps, ontology or domain word systems and other processes from the knowledge organization system to the term set matching improves the semantics of term sets, and further enhances the connotation of terms, so that the higher accuracy of matching is an important study of term set matching. direction;

(2) In the process of term set matching, the advantage of machine learning is better played, especially the effective use of deep learning and artificial neural network algorithm models, improving the efficiency of

term set matching is another important direction of term set matching.

ACKNOWLEDGMENT

This work is partially supported by ISTIC Key Project Program (Grant No.: ZD2018-07), CKCEST Project Program (Grant No.: CKCEST-2018-1-26) and National Digital Composite Publishing System Project (Grant No.: XWCB-ZDGC-FHCB/28). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Feng Zhiwei. Introduction to Modern Termology [M]. Commercial Press, 2011. (In Chinese)

[2] Nenadic G, Spasic I, AnaniadouS. Automatic discovery of term similarities using pattern mining [C]. Taipei: International Conference On Computational Linguistics, 2002: 1-7.

[3] Li M, Chen X, Xin M L, et al The Similarity Metric [C]. In: IEEE Transactions on Information Theory. 2003. 86 3-872.

[4] Salton, G., & Buckley, C. Term weighting approaches in automatic text retrieval. Information Processing & Management, 1988, 24 (5) 513-523.

[5] Salton Gerard. Automatic text processing the transformation, analysis, and retrieval of information by computer [M]. Addsion-Wesley Longman Publishing Co., Inc., 1989.

[6] Zobel, J., & Moffat, A. Exploring the similarity space. ACM SIGIR Forum, 1998, 32 (1) 18-34.

[7] Wongn Skm, W Ziarko, Pcn Wong. Generalized vector spaces model in information retrieval [C]. 1985. 18-25.

[8] Zhang Huanjiong, Wang Guosheng, Zhong Yixin. Calculation of Text Similarity Based on Hamming Distance [J]. Computer Engineering and Applications, 2001, 37 (19): 21-22. (In Chinese)

[9] Xu Degang, Xu Xiyang, Chen Xiao et al. Improved consensus-clustering algorithm based on Minkowski distance and its application [J]. Journal of Hunan University (Naturnal Science), 2016, 43 (4): 133-140. (In Chinese)

[10] Guo Yanhui, Yin Xijie, Zhang Hong. A local binary improved algorithm for image classification [J]. Chinese Journal of Light Industry, 2017, (3): 73-77. (In Chinese)

[11] Shen Xiaoyong, Lei Yingjie, Cai Ru et al. Measurement method for dissimilarity of IFS based on weighted Minkowski distance [J]. System Engineering and Electronics, 2009, 31 (6): 1358-1361.

[12] Qiao Shen, Lv Zhimin, Zhang Nan et al. Improved Traveling Salesman Problem Based on Hamming Distance Improved Particle Swarm Optimization [J]. Computer Applications, 2017, 37 (10): 2767-2772. (In Chinese)

[13] Ji Weihua, Lv Guofang. Method of Handling Conflict Evidence Based on Generalized Jaccard Coefficients [J]. Control Engineering, 2015, (1): 98-101. (In Chinese)

[14] Meng Chenchen, Xiao Jianyu, Luo Lan.Evidence Conflict Measurement Method Based on Similarity Coefficient and Jaccard Coefficient Matrix [J]. Journal of Chongqing University of Posts and Telecommunications, 2017, (3): 421-426. (In Chinese)

[15] Yang Jiahui, Liu Fangai.Collaborative filtering algorithm based on Bahman coefficient and Jaccard coefficient [J]. Computer Applications, 2016, (7): 2006-2010. (In Chinese)

[16] Li Xiaohui. Research on personalized recommendation algorithm based on similarity of Jaccard items [D]. Central South University, 2010. (In Chinese)