# FEATURE SELECTION OPTIMIZATION METHOD OF TRAFFIC CONGESTION CASE DATABASE BASED ON TF-IDF ALGORITHM

Hao ZHANG[1]

*This paper analyzes the key factors of traffic congestion management decision-making in detail and proposes to use Term Frequency-Inverse Document Frequency (TF-IDF) algorithm to optimize the feature attributes of traffic congestion in the decision support model of case-based reasoning. TF-IDF algorithm of text weight ranking has been used to sort the feature attributes. At the same time, in view of the shortcomings of the TF-IDF algorithm, an optimized algorithm is proposed in order to balance the category difference of the distribution of feature attributes. The combination of the probability distribution of feature words between category and information entropy solves the problem of weight shift caused by the uneven distribution of feature words within and among text categories. The traditional TF-IDF, DF algorithm and the optimized TF-IDF algorithm are verified by experiments from three aspects: recall rate, precision rate and F value. The results show that the optimized TF-IDF algorithm has better classification ability and superiority in the selection of traffic congestion feature attributes.*

**Keywords:** case-based reasoning, traffic congestion, TF-IDF algorithm, feature attributes

## 1. Introduction

Case Based Reasoning (CBR) [1], as a new reasoning model, has been introduced in the field of artificial intelligence after Rule Based Reasoning (RBR). It is an important method for knowledge learning and derivation. When faced with certain problems, experts in related fields will analyze the problem firstly, find out the corresponding feature attributes of the problem, further interpret the problem, and recall the process of solving the current problem with solutions to similar problems. It can be seen that the feature selection optimization of the case database in case-based reasoning technology is the key to the whole case-based reasoning process.

In the field of transportation, many scholars have conducted many related research. Wei Zhang [2] proposed a congestion management method under the congestion condition of typical road networks, which can estimate and respond to

[1] Prof., College of Mathematics and Computer Science, Tongling University, China, e-mail: 027510@tlu.edu.cn

different congestion situations; Ji XiaoFeng [3] implanted database technology into the traffic congestion management decision-making system, using rough sets theory to acquire congestion knowledge; Markus Schade [4] gave a CBR allocation model for traffic networks; John L. McLin et al. [5] proposed traffic control system capable of traffic accident monitoring and management based on CBR technology. Xiaoyan Yang et al. [6] verified the position update of the basic particle swarm optimization algorithm for the minimum attribute reduction problem and revised the formula at the same time, which can effectively solve the problem; Xia Xian Zhi [7] introduced the operator of the genetic algorithm into the ant colony algorithm and improved the generation of ants by using the variability of genetic algorithms, so as to improve the convergence speed and the global search ability of the algorithm, thereby improving the retrieval efficiency. Wang Gunnyet al. [8] proposed a method of feature weight optimization of case system based on the combination of genetic algorithm and taboo algorithm; ShenQi [9] used grey correlation in the initial population selection of genetic algorithm to analyze the results, and proposed an optimized case reasoning model; Li Feng Gang et al. [10] used the technology of combining K-nearest neighbors and cross-level to design an optimized selection scheme for the feature attributes of the case database. Glukhikh Igor, Glukhikh Dmitry [11] studied on the decision support for a large-scale complex objects.Wu QiCaiet al. [12] introduced a CBR decision making model in emergency scenarios. Lin Zhang [13] elaborated on the case retrieval in case-based reasoning. Some studies [14, 15] presented an adaptive search model based on genetic algorithm. The another [16, 17] solved the adaptive problem of multi variable in case-based reasoning by using the method of multiple relationship analysis and clustering. Okudan Ozanet al. [18, 19] addressed the case retrieval problem by using fuzzy language to form a formal feature list.

Recent research, adopting case-based reasoning in the traffic safety, mainly focused on rail transit or large network. Because they are more of regularization data, these cases are processed by using rule reasoning and case reasoning. Most of the research related to urban road congestion are in the field of congestion prediction, and there are relatively few research on the timely dredging of the congestion, especially the researches on using case-based reasoning technology to deal with the dredging of urban road traffic congestion and solve unstructured and semi-structured data in complex environment. Therefore, it is a necessary attempt to study the decision support technology of urban traffic congestion relief based on case-based reasoning.

The weight calculation of traffic congestion feature attributes can apply the retrieval idea of web search engines. The traditional method has been abandoned, this study tried to take the text classification as an example, mainly taking the Spatial Vector Model (VSM) [20, 21] as the representation of text.

Firstly, the text is divided into morphemes (word segmentation) [22], and then the selection of eigenvalues and the calculation of the weight of eigenvalues are carried out. Finally, a set of multi-dimensional traffic congestion feature attribute vectors could be formed.

## 2. Traffic Congestion Feature Selection Strategy Based on TF-IDF Algorithm

### 2.1 TF-IDF Algorithm

According to the system characteristics involved in this paper, the values of feature attributes in the urban road congestion management decision-making support system are mostly enumerated data. Since the values of feature attributes are basically presented as text, multi-bit semi-structured or unstructured data, the problem of assigning feature value attribute weights is mainly how to solve the classification and selection of text data. In the CBR-based traffic congestion management decision-making support technology, the weight calculation of the traffic congestion feature attributes can apply the search ideas of web search engines. We have abandoned the traditional principal component analysis method, expected cross entropy and other methods, and took the field of text classification as an example by using the Vector Space Model (VSM) [22] as the text representation. First the morpheme (word segmentation) of the text that needs to be represented is divided. then the value of feature attributes is selected and the feature value weights calculated. Finally, a multi-dimensional traffic congestion feature attribute vector set is formed.

$$W_{it} = TF * IDF = f_{it} * \log \frac{N}{n_t} \tag{1}$$

In this formula, TF is the frequency at which the term *t* appears in document *i*, and in its inverse text frequency IDF = $\log \frac{N}{n_t}$ , *N* represents the total number of documents, and $n_t$ represents the number of documents where the term *t* appears.

### 2.2 Optimization of Different Traffic Congestion Situations within the Category

The introduction of IDF was originally designed to suppress high-frequency useless words appearing in documents and prevent high-frequency words from negatively affecting the entire document. However, a problem is highlighted: high-frequency words are suppressed, and the TF-IDF algorithm does not have the ability of document type identification when feature words that are unevenly distributed in different types of documents.

Feature words that are concentrated distributed in a certain category of documents or in a few categories of documents often have a strong ability to identify document type, which can represent the subject of such documents, and should be given higher weight. Take the traffic congestion event as an example: the initial traffic congestion events in the traffic case database are set to be M, and the traffic congestion events are caused by different reasons, so the management scheme should be divided according to the congestion situation of different reasons. Assuming that congestion events Mare divided into j different categories $C = \{C_1, C_2, \cdots, C_j\}$, and feature word set $T = \{t_1, t_2, \cdots, t_i\}$, the number of occurrences of the two feature words "blizzard weather" and "early peak" in traffic congestion events in three different types of congestion events $C_1$, $C_2$ and $C_3$ are shown in Table 1.

*Table 1*

**Distribution of traffic congestion feature words**

| Congestion event category | Frequency of feature attribute distribution | |
|---|---|---|
| | Blizzard weather | Early peak |
| C1 | 8 | 6 |
| C2 | 1 | 5 |
| C3 | 2 | 7 |

In the traffic congestion event $C_1$, "blizzard weather" appeared 8 times, and "early peaks" appeared 6 times; in the traffic congestion event $C_2$, "blizzard weather" appeared 1 time, and "early peak" appeared 5 times; in the traffic congestion event $C_3$, the two feature words appear 2 times and 7 times respectively. Let the number of documents containing the congestion event be 10, and the weight of these two feature words in the category $C_1$ is calculated according to the traditional IDF:

$$IDF_{blizzard\ weather} = \log(10/8) = 0.0969$$

$$IDF_{early\ peaks} = \log(10/6) = 0.2218$$

$$IDF_{blizzard\ weather} < IDF_{early\ peaks}$$

The result is not consistent with the actual situation.

In fact, "blizzard weather" is a relatively special word, as it can represent the theme of traffic congestion events $C_1$, and it should be assigned a larger weight. Therefore, the weight shift caused by the difference in category distribution is a problem that needs to be solved.

First a distribution frequency matrix of traffic congestion feature words is constructed:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} \\ x_{21} & x_{22} & \cdots & x_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} \end{bmatrix} \tag{2}$$

Where $x_{ij}$ represents the frequency of the feature word $t_i$ in the traffic congestion event $C_j$. Then a Boolean matrix of feature words in all traffic congestion event categories is constructed as well. If it appears in a certain category, it is 1. Otherwise, it is 0.

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1j} \\ y_{21} & y_{22} & \cdots & y_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ij} \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \\ 1 & 1 & \vdots & 1 \\ 0 & 1 & \cdots & 1 \end{bmatrix} \tag{3}$$

The distribution frequency of feature words and a Boolean matrix is multiplied:

$$\begin{aligned} X \bullet Y &= \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} \\ x_{21} & x_{22} & \cdots & x_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} \end{bmatrix} \otimes \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1j} \\ y_{21} & y_{22} & \cdots & y_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ij} \end{bmatrix} \\ &= \begin{bmatrix} x_{11}y_{11} & x_{12}y_{12} & \cdots & x_{1j}y_{1j} \\ x_{21}y_{21} & x_{22}y_{22} & \cdots & x_{2j}y_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1}y_{i1} & x_{i2}y_{i2} & \cdots & x_{ij}y_{ij} \end{bmatrix} \end{aligned} \tag{4}$$

The total number of different categories containing feature words is calculated. Let $K_{ij}$ be the proportion of feature words $t_i$ in all feature words in the category $C_j$ :

$$K_{ij} = \frac{x_{ij}}{x_{1j}y_{1j} + x_{2j}y_{2j} + \cdots + x_{ij}y_{ij}} \tag{5}$$

where $K'_{ij}$ is the sum of the weights of the feature words $t_i$ in the other categories except the weights of the $C_j$ category.

Considering the possibility of $K_{ij} < K_{ij}^{'}$, the IDF value will be negative, while the weights of the feature words should all be positive, therefore the IDF is modified. According to the nature of the logarithmic function, if a positive number greater than 0 must be obtained in the value, an influence factor of 1 should be added to rationalize the weight calculation result. The above can be rewritten as:

$$IDF = \log(\frac{K_{ij}}{K_{ij}^{'}} + 1) \tag{6}$$

$$IDF_{blizzard\ weather} = \log(\frac{8/14}{1/6+2/9}+1) = 0.3925$$

$$IDF_{early\ peaks} = \log(\frac{6/14}{5/6+7/9}+1) = 0.1024$$

$$IDF_{blizzard\ weather} > IDF_{early\ peaks}$$

The result is in line with the actual situation.

### 2.3 Optimization for Different Traffic Congestion Situations Between categories

The weight shift caused by the uneven distribution between categories is considered here. According to the actual situation of traffic congestion, the occurrence of feature words is likely to be the same in the same type of traffic congestion events. In other words, if there are two terms with basically the same distribution in the category, the classification effect of the two terms cannot be accurately defined. According to the definition of information entropy [23], the distribution of feature words in the category can be reflected:

Definition If there is a j-document probability distribution for a certain feature word t in a certain type of $C_j$ :

$$P=(p_1, p_2, \cdots, p_j), \quad p_j = \frac{Nd_j}{NC_i} \tag{7}$$

where $Nd_j$ indicates the frequency of the feature word *t* in the document *j* of the category $C_i$, and $NC_i$ indicates the frequency of the feature word ***t*** in all documents of the category $C_i$. Then, the amount of information passed by the feature word ***t*** in $C_j$ is:

$$
\begin{aligned}
I(p) &= -(p_1 * \log(p_1) + p_2 * \log(p_2) + \cdots + p_j \log(p_j)) \\
&= -\sum_{j}^{n} p_j \log(p_j) = -\sum_{j}^{n} \frac{Nd_j}{NC_i} \log \frac{Nd_j}{NC_i}
\end{aligned} \tag{8}
$$

where n is the total number of documents in category $C_i$. It can be reflected from the formula that the more uniform the distribution of the feature word t in the category $C_i$ is, the greater the information entropy within the class becomes, and the better the entry t can reflect the feature information of the category.

We also use the feature words in the case of traffic congestion and select the two feature words "blizzard weather" and "early peak" for comparison. Assuming that there are 5 documents of category $C_1$ in the case congestion event $C_1 =（d_1, d_2, d_3, d_4, d_5）$, all of which contain the above two feature words appearing 10 times in each document of category $C_1$, the distribution of the documents is shown in Table 2:

*Table 2*

**The distribution of feature words in the category C1**

| Feature term<br>Document Name | Blizzard weather | Early peak |
|---|---|---|
| d1 | 2 | 4 |
| d2 | 3 | 1 |
| d3 | 1 | 0 |
| d4 | 2 | 5 |
| d5 | 2 | 0 |

According to the traditional TF-IDF, both "blizzard weather" and "early peak" appear 10 times in all 5 documents, so $IDF_{blizzard\ weather} = IDF_{early\ peak}$ should be assigned the same weight. At this time, it is not possible to show the classification ability of the two words. Obviously, the "blizzard weather" in category $C_1$ should better reflect the theme of traffic congestion, so it should be assigned a larger weight. The probability of two feature words appearing in the information entropy formula is:

$$I_{early\ peak} = I(4/10, 1/10, 0, 5/10, 0) = 1.089$$

$$I_{blizzard\ weather} = I(2/10, 3/10, 1/10, 2/10, 2/10) = 3.619$$

$$I_{blizzard\ weather} > I_{early\ peak}.$$

The result is in line with the actual situation.

The calculation proves that the more uniform the feature word distribution in the documents of the same category, the larger the weight should be assigned between the category.

The author analyzes the experimental results in detail, grabs the eigenvalues of 150 samples according to the improved TF-IDF algorithm, forms a standard case base, and divides the corresponding data types according to the structure of eigenvalues

**1.** Input the sample to be tested.

**2.** Combined with the probability distribution of feature words between classes, the problem of weight shift between classes caused by the difference of class distribution in TF-IDF algorithm is corrected.

**3.** Thirdly, combined with information entropy, the problem of weight shift caused by uneven distribution of feature words in text class is solved.

**4.** The weight of eigenvalues is calculated and sorted according to the weight to end the algorithm.

Finally, the word frequency probability distribution between the category and information entropy are combined to improve the TF-IDF algorithm, so that the feature words can be determined accurately. The TF-IDF algorithm can be rewritten as:

$$TF - IDF = f_{it} \times \log(\frac{K_{ij}}{K_{ij}^{'}} + 1) \times (-\sum_{j}^{n} p_j \lg p_j) \qquad (9)$$

The improved TF-IDF algorithm can effectively solve the problem of the distribution of feature words between and within categories that was not previously considered. Finally, the feature words can be weighted and sorted according to the improved algorithm, and features with higher weights are selected as the feature space vector of the text.

## 3 Experimental Results and Analysis

In this paper, the TF-IDF algorithm is improved to balance the weight shift caused by the distribution of feature words within categories. The superiority of the improved algorithm is verified through experiments.

### 3.1 Experimental Steps

**1.** The representation of the text. In the test, Institute of Computing Technology, Chinese Lexical Analysis System (ICTCLAS [24]), a word segmentation tool used by Computer Research Institute of Chinese Academy of Sciences, is used to segment the test text data set. Some irrelevant words (such as auxiliary function words, etc.), and the terms that almost appear in every document according to the characteristics of the test set are removed. After filtering these terms, a candidate feature set of VSM model is formed.

**2.** Feature item selection. Compared with traditional TF-IDF algorithm, DF algorithm and improved algorithm including word frequency probability between the category and information entropy are used to calculate the weight of candidate feature items in the test document, which are arranged in descending order according to the weight. The first M items with larger weights are selected to form a feature vector space with dimension M so as to form a feature vector set.

**3.** The choice of classifier. There are many algorithms for text classification. Commonly used algorithms include K-nearst neighbor (KNN), ListTree, and Support Vector Machines (SVM) algorithms. The purpose of the experiment is to test the document collection with a general performance evaluation method. It is reflected in the comparison of precision rate P (Precision), recall rate R (Recall), and $F\ value$ .

KNN algorithm was selected for document testing in experiments. The KNN algorithm is actually used to find the K samples that are closest to the unclassified sample X among the known class samples. Since the experiment is to classify a known text sample, we selected 5 different types of text data at this time, so this is K = 5, and in text classification, the distance between text data is calculated. Here, the cosine function can be used to replace the traditional Euclidean distance, and the angle between the document feature vectors is measured. The smaller the angle, the higher the similarity. The formula is as follows:

$$sim(X,Y) = \frac{(X,Y)}{|X| \bullet |Y|} = \frac{\sum_{i=1}^{t} w_{Xi} * w_{Yi}}{\sqrt{\sum_{i=1}^{t} (w_{Xi})^2 \sum_{i=1}^{t} (w_{Yi})^2}} \tag{10}$$

KNN predicts unknown samples based on the characteristics of samples of known categories. A simple prediction is that the unknown sample category contains the largest sample category among the nearest samples of the known sample categories. The experiment uses VC ++ 6.0 to implement the algorithm.

### 3.2Experimental Results and Analysis

In the experiment, the traffic congestion processing documents collected by the author in a city's traffic police are used. First, the traffic congestion events collected are classified into traffic accidents, special services, bad weather, normal congestion, and public security incidents according to different situations. 30 of each of the five types of traffic congestion events, 150 congestion events in total, are randomly selected to form a text set to be classified in the test classification.

The classification based on the KNN algorithm is mainly used to compare with that of the traditional TF-IDF, document frequency, and optimized TF-IDF algorithm in the accuracy rate P, recalls rate R and F values. The test results are shown in the following Table 3:

*Table3*

**Classification results of TF-IDF algorithm**

| Events<br>Algorithm | Special Service | Security Events | Traffic Accident | Bad Weather | Normal Congestion |
|---|---|---|---|---|---|
| TF-IDF | 22 | 22 | 21 | 21 | 24 |
| DF | 21 | 21 | 19 | 20 | 23 |
| improved TF-IDF | 24 | 23 | 23 | 23 | 26 |

*Table 4*

**Comparative analysis of classification effects**

| Algorithm<br>Events | TF-IDF | | | DF | | | Improved TF-IDF | | |
|---|---|---|---|---|---|---|---|---|---|
| | P% | R% | F% | P% | R% | F% | P% | R% | F% |
| Special Service | 84.61 | 73.33 | 78.54 | 79.31 | 76.67 | 77.97 | 85.71 | 80.00 | 82.75 |
| Security Events | 81.48 | 73.33 | 77.19 | 77.78 | 70.00 | 73.69 | 82.14 | 76.67 | 79.31 |
| Traffic accident | 77.77 | 70.00 | 73.72 | 73.07 | 63.33 | 67.87 | 79.31 | 76.67 | 77.96 |
| Bad weather | 80.76 | 70.00 | 77.07 | 80.00 | 66.67 | 72.76 | 82.14 | 76.67 | 79.30 |
| Normal congestion | 88.89 | 80.00 | 84.22 | 88.46 | 76.67 | 82.14 | 89.65 | 86.67 | 88.13 |

From Table 3 and Table 4, it can be seen that the accuracy rate P, the recall rate R, and the F value are improved based on the improved TF-IDF algorithm. The curve comparison charts of the accuracy rate, the recall rate and F value are simulated respectively, shown in Fig. 1:

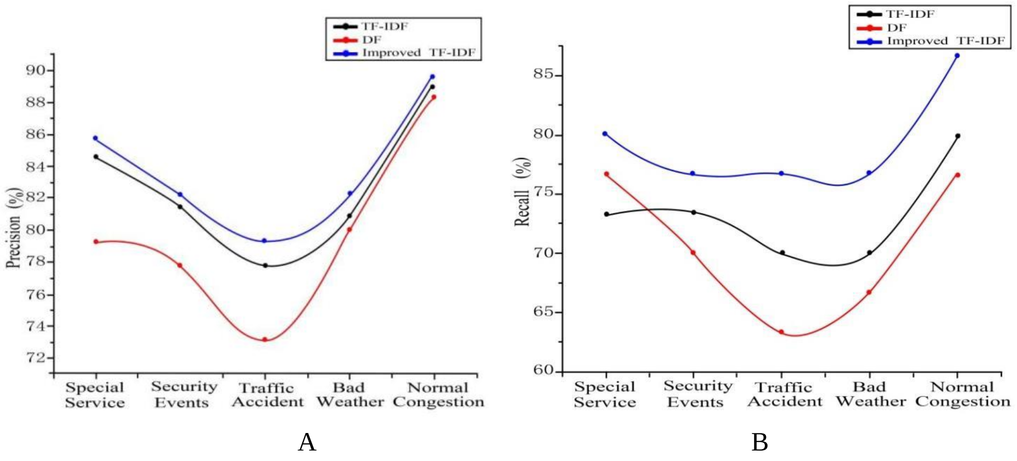
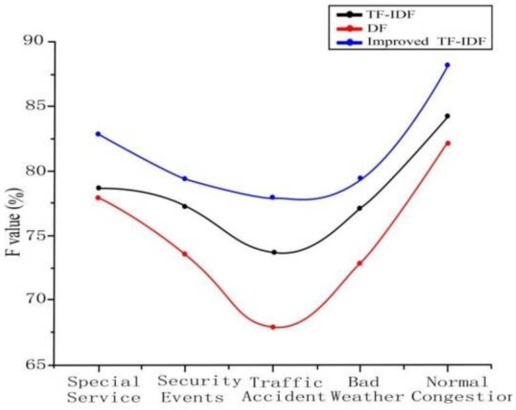


Fig. 1. A) B) Comparison of precision ratio and recall rate and F values



C

Fig. 1. C) Comparison of precision ratio and recall rate and f values

It can be reflected from the curve that the document frequency performance comparison of the accuracy rate, recall rate, and F value is the weakest based on the classification results of the improved TF-IDF algorithm, the traditional TF-IDF algorithm and DF algorithm. However, the improved TF-IDF algorithm performs better than the TF-IDF algorithm, which indicates that the improved TF-IDF algorithm has better feature selection advantage.

**1.** Since traffic congestion cases are presented as unstructured data, feature vectors are used to represent case knowledge. The improved TF-IDF algorithm is used to extract feature values of 150 groups of traffic jam cases. A more reasonable division of the weight setting and importance between different classes is carried out, and some useless words that appear repeatedly in too many schemes are removed. In this way, the weight calculation of the feature values attributes is more reasonable. According to formula 11, the weight of traffic congestion cases is calculated. The results are as follows shown in Table 5:

*Table 5*

**Weight calculation results**

| number of feature value | index | weight |
|---|---|---|
| 1 | morning rush hour | 0.78 |
| 2 | morning rush hour | 0.76 |
| …… | …… | …… |
| 6 | normal traffic congestion | 6.84 |
| 7 | emergencies | 6.79 |
| …… | …… | …… |
| 60 | visibility >200m | 0.67 |
| 61 | visibility between 100 and 200m | 0.66 |
| …… | …… | …… |

Thus, the weight vector of the congestion feature attributes is obtained:

Y=(y1,y2,……,y270),T=(0.78,0.76,……,0.66,…)T .

The weight threshold is set to 0.5. After the weight calculation, the feature attributes with a weight less than 0.5 are discarded, and finally a text feature attribute data set containing 70 feature attributes is formed.

**2.** Finally, a case library F containing 7 feature attribute sets is formed, and the structure of the traffic congestion management case library formed by congested vehicles $F=(S_a,S_b,S_c,S_d,S_e,S_f,S_g,S_h)$, in which the attributes represented by each element are as follows: $S_a$ represents the time period of road congestion. According to the time period of traffic congestion, it can be roughly divided into the third stage, peak period, flat peak period and low valley period; $S_b$ represents the location of congestion, that is, the type of road section at that time. According to the primary and secondary points of urban roads, the congestion locations are divided into: main road, secondary trunk road, general road section, branch road, level intersection and grade intersection; $S_c$ indicates the cause of congestion when the case occurred on the road section, including normal congestion, emergencies, group incidents, road repairs and other incidents, etc. $S_d$ indicates the type of congestion that occurred in the case. Generally, it can be divided into two types according to the road congestion: initial congestion and subsequent congestion; $S_e$ represents the congestion range of vehicles, which is generally divided into three situations: point, line, and area, and linear congestion and area congestion can be expressed by the corresponding congestion length and area; $S_f$ represents the weather conditions of congestion, including sunny days, rainy days, haze days, sandstorms, snowy days, etc.; $S_g$ indicates the level of congestion, which can be divided into four situations: mild, moderate, severe, and deadlock. The feature attributes and values of the case library construction are shown in Table 6.

*Table 6*

**Case features and values of case library**

| Case attributes | Classificationoffeature attributes | Types of feature attribute |
|---|---|---|
| Sa | congestion time | enumeration |
| Sb | congestion location | enumeration |
| Sc | congestion causes | enumeration, numerical type |
| Sd | congestion types | enumeration |

| | | |
|---|---|---|
| Se | congestion range | enumeration, numerical type |
| Sf | weather condition | numerical interval type |
| Sg | congestion level | enumeration |

## 4. Conclusion

This paper places emphasis on the optimization selection of feature attributes in case-based reasoning decision systems. This study focuses on the optimization selection of eigenvalue attribute in case-based reasoning decision system. Combined with the data characteristics of feature attributes in the decision support system for urban road traffic congestion relief, this paper tries to use TF-IDF algorithm of text weight ranking to rank the weight of feature attributes. Aiming at the shortcomings of TF-IDF algorithm, an optimization algorithm is proposed to balance the class difference degree of feature attribute distribution.

By analyzing the text of traffic congestion cases, the TF-IDF algorithm is optimized to make the extraction of feature attributes more accurate. The past cases, better matching the current congestion situation, can be found in the case retrieval process, which provides a quick reference and decision support for the traffic management department to make plans.

## Acknowledgement

This work was supported by Project No. KJ2021ZD0126.

## REFERENCES

[1]. *Tian Jun, Xin Hong, Cheng Shaochuan* Progress in the Research of Decision Support Systems in China. Science andTechnologyGuide.2005, Vol.23 No.7: 71-75

[2]. *Zhang Wei, He Rui Chun*. Traffic dispersion aid decision method based on CBR.COMPUTERENGINEERING ANDDESIGN.2014, Vol.35 No.10:3621-3625.

[3]. *JiXiao Feng*. Congestion Management Methods based on Traffic Information Extractionin RegionalRoadNetwork.SouthwestJiaotongUniversity.2009

[4]. *Markus Schade*. Using case-based reasoning to control traffic consumption. Germany: VDM Verlag2007

[5]. *John L McLin, William T Scherer*. Development and evaluation fa-control system for regional traffic management. Advances in Civil Engineering, 2011:1-11

[6]. *Yang Xiaoyan, Chen Guolong*. Minimum Attribute Reduction algorithm based on Particle Swarm Optimization. Journal of Fuzhou University (Natural Science). 2010, Vol.38 No.2:193-197

[7]. *Xia Xianzhi, Du Xinyu, Zheng Yangfei*. Attribute Reduction based on Ant Colony Genetic Algorithm. Computerand Modernization. 2013, Vol.34 No.1:25-28

[8]. *Wang Guanyu, GuoYong*. Simulation Research on Case system Feature Weight Optimization Algorithm. Computer Engineering and Applications.2013, Vol.49 No.1:261-264

[9]. *Shen Qi*. Using Genetic Algorithm to Further Optimize CBR case-based reasoning Model. Computer and Modernization.2013, Vol.34 No.2:147-149

[10]. *Li Fen gang, Ni Zhiwei, Yang Shanlin*. Attribute Reduction and its Performance Evaluation in case-based reasoning. Journal of Tsinghua University (NATURAL SCIENCE EDITION). 2006, Vol.46(S1):1025-1029

[11]. *Glukhikh Igor, Glukhikh Dmitry*.Case-Based Reasoning with an Artificial Neural Network for Decision Support in Situations at Complex Technological Objects of Urban Infrastructure. Applied System Innovation. Volume 4, Issue 4. 2021. PP 73-78.

[12]. *Wu Qicai, Yuan Haiwen, Yuan Haibin*. Development of Ground Special Vehicle PHM with Case-Based Reason Model.Applied SciencesVolume 11, Issue 10. 2021. PP 4494-4499.

[13]. *Lin Zhang*.Research on case reasoning method based on TF-IDF. International Journal of System Assurance Engineering and Management.2021. PP 1-8.

[14]. *Bi Xin, Nie Haojie, Zhang Xiyu et al*. Unrestricted multi-hop reasoning network for interpretable question answering over knowledge graph. Knowledge-Based Systems, 2022, 243-247.

[15]. *Stanovov Vladimir, Akhmedova Shakhnaz, Semenkin Eugene* The automatic design of parameter adaptation techniques for differential evolution with genetic programming. Knowledge-Based Systems, 2022, 239.

[16]. *Lin Mingchi,He Dubo, Sun Shengxiang*. Multivariable Case Adaptation Method of Case-Based Reasoning Based on Multi-Case Clusters and Multi-Output Support Vector Machine for Equipment Maintenance Cost Prediction. IEEE Access Volume 9, 2021.

[17]. *Liu Hao, Zhou Shuwang, Chen Changfang et al.* Dynamic knowledge graph reasoning based on deep reinforcement learning. Knowledge-Based Systems, 2022, 241-249.

[18]. *Okudan Ozan, Budayan Cenk, Dikmen Irem*A knowledge-based risk management tool for construction projects using case-based reasoning. Expert Systems with Applications, 2021, 173-185.

[19]. *H Wu, G Salton.* A Comparison of Search Term Weighting: Term Relevance vs. Inverse Document Frequency. Cornell University, 1981

[20]. *Coyle L, Cunningham P*. Improving recommendation ranking by learning personal feature weights[C]. In: Proceedings of the 7th international conference on case-based reasoning, 2004:560-572

[21]. *Hsu C-I, Chiu C, HsuP-L*Predating information systems outsourcing success using a hierarchical design of case-based reasoning. ExpertSyst Apple 2004, Vol.26:435–441

[22]. *KhattakA, KanafaniA*.Case-based reasoning: planning tool for intelligent transportation systems. Transp Res-PartCEmerge Techno 2014, Vol.5:267–288

[23]. *Kohavi R, Langley P, Yun Y* (1997) Theutility of feature weighting in nearest neighbor algorithms. Posterpaperof the European conference onmachine learning, Prague

[24]. *Kulick J, Lieck R, ToussaintM* (2014) Active learning of hyper parameters: an expeted crosse stroppy criterion for active model selection. Eprint Arxiv, 1099–1125.