Algorithmic Models for MicroBlog Information Mining System

Jing Leng¹, ², Wei Xu¹, ²*

¹ Department of Information Technology, Hubei University of Police, Wuhan, Hubei, 430034, China;
² Electronic Forensics and Trusted Applications Hubei Collaborative Innovation Center, Wuhan, Hubei, 430034 China;
*Corresponding author’s e-mail: daleng0127@sina.com

Abstract. Based on the “MicroBlog Community Massive Information Mining and Monitoring System”, this paper introduced the algorithm and model construction of microblog information word processing and mining conversion core functions of microblog information subsystem, in order to obtain the information objects needed in each link of the process of microblog information mining, which provides a basis for the realization of monitoring and early warning in the MicroBlog monitoring platform.

1. Introduction
The MicroBlog information mining subsystem is the core function module of the “MicroBlog community massive information mining and monitoring system” [1]. The module carries out word processing, classification, clustering and information transformation of MicroBlog information in business-related databases to obtain MicroBlog feature word information, event information, classification information and monitoring value information, so as to provide the basis of research and judgment for the realization of accurate monitoring and real-time warning. The flow chart of MicroBlog information mining subsystem is shown in Fig.1. Its function is based on the application of algorithm and the construction of model, mainly including automatic word segmentation algorithm, automatic feature word algorithm, similar clustering model and information quantization algorithm.

This paper will mainly introduce the application and model construction of the above several algorithms in the system.

2. Automatic Word Segmentation Algorithm
The function of automatic word segmentation for the whole sentence or the whole text string of MicroBlog is the basic function of micro-blog information mining subsystem, and the automatic word segmentation algorithm is also the basis of many algorithms of the system. Designing an automatic word segmentation algorithm suitable for the characteristics of MicroBlog text information is the premise to ensure the scientificity of automatic word segmentation function. [2]

2.1 Analysis of Word Segmentation Algorithm
At present, Chinese word segmentation technology is basically divided into three categories: mechanical algorithm (also known as word segmentation algorithm based on string matching), understanding-based word segmentation algorithm and statistics-based word segmentation algorithm.
2.1.1 Mechanical Word Segmentation Algorithm
According to certain strategies, the Chinese character string to be analyzed is matched with the entries in a “sufficiently large” machine dictionary (Chinese related dictionary). If a string is found in the dictionary, a word is recognized.

2.1.2 Understanding-based Word Segmentation Algorithm
Understanding-based word segmentation algorithm achieves the effect of recognizing words by letting computer simulate people’s understanding of sentences. Its basic idea is to carry out syntactic and semantic analysis at the same time of word segmentation, and use syntactic and semantic information to deal with ambiguity.

2.1.3 Statistics-based Word Segmentation Algorithm
Statistics-based word segmentation algorithm performs word segmentation by counting the frequency of combinations of adjacent co-occurrence words.

By analyzing the characteristics of MicroBlog and three kinds of word segmentation algorithms, mechanical word segmentation is more suitable for this system. Because of the short language and various forms of expression of MicroBlog text information, the frequency of MicroBlog text information is not distinguishable, so the statistics-based word segmentation algorithm is not

Figure 1. Flow chart of micro-blog information mining subsystem

Figure 2. Automatic Word Segmentation Model
applicable, while the understanding-based word segmentation algorithm is still in the experimental stage, and the accuracy is not high enough.

The micro-blog information mining subsystem chooses the mechanical word segmentation algorithm which is more mature and suitable for the characteristics of micro-blog text information as the basic algorithm. However, since word segmentation is an intelligent decision-making process, mechanical word segmentation method can not completely solve the two basic problems of word segmentation stage: ambiguity segmentation and unknown word recognition. Therefore, the segmentation accuracy of mechanical word segmentation algorithm can not fully meet the needs of MicroBlog information mining subsystem. We use mechanical word segmentation as a preliminary segmentation method of MicroBlog information, and then use other ways to further improve the accuracy of segmentation.

2.2 Automatic Word Segmentation Model
The automatic word segmentation model designed by the MicroBlog information mining subsystem takes the mechanical word segmentation algorithm as the basic word segmentation algorithm and makes corresponding improvements according to the characteristics of MicroBlog text information. The model for building automatic word segmentation is shown in Fig.2.

The model is divided into three parts. In the first part, the text information of MicroBlog to be analyzed is matched with the entries of Chinese dictionary and English dictionary in the word segmentation lexicon. If a string is found in the dictionary, the matching is successful. The initial segmentation strings are formed by the initial segmentation of MicroBlog word information by mechanical word segmentation algorithm.

In the second part of the model, the first initial result is scanned by using the priority recognition technology, and some words with obvious features (Words other than Chinese dictionary and English dictionary in word segmentation lexicon) are recognized and segmented twice in the initial character string. With these words as breakpoints, the second matching calculation of the string is carried out to reduce the matching error rate and improve the accuracy of segmentation.

Finally, the final word segmentation information can be obtained by comprehensive analysis of the results formed by mechanical word segmentation and priority recognition technology.

The automatic word segmentation model first classifies the text by mechanical word segmentation, then uses priority recognition technology to make up for the inaccuracy of mechanical word segmentation, and finally analyzes and integrates to form the results of word segmentation, which has high accuracy and efficiency.

2.3 Flow Chart of Automatic Word Segmentation Algorithm
The operation flow of the automatic word segmentation algorithm is shown in Fig.3.

2.3.1 Preliminarily Classify the Meanings of MicroBlog Words
The most commonly used positive maximum matching method in mechanical word segmentation is used to preliminarily classify the meanings of MicroBlog words. The steps are as follows:

1) Constructing a word segmentation lexicon with enough scope;
2) Extracting text information S from the text of MicroBlog to be processed;
3) Test whether S.substring (0,k) matches words in the lexicon. If so, save the length max=k. Otherwise, after k plus 1, the matching is continued in the lexicon. Repeat this step (k is the maximum length of words in the lexicon).
4) Get the largest word. The length is max.
5) Return to (3) and continue processing S=S.substring(max), that is, the remaining characters, to get the initial segmentation results of semantics of MicroBlog information.

2.3.2 The Second Matching Calculation
On the basis of mechanical word segmentation, a feature scanning comparison of the previous initial segmentation results is carried out by using priority recognition technology. Second recognition and
segmentation is carried out to get words with obvious features as breakpoints for the second matching calculation of strings.

2.3.3 Comprehensive Analysis
The final word segmentation information is obtained by comprehensive analysis of the above two results. The error rate of matching should be minimized and the accuracy of MicroBlog text string segmentation should be improved.

3. Automatic Feature Word Algorithms

3.1 Analysis of Automatic Feature Word Algorithm
In order to extract, classify and analyze the MicroBlog information correctly, the MicroBlog information mining system needs to extract the feature words from the MicroBlog information with automatic word segmentation using the automatic feature word algorithm as the semantic representation of the MicroBlog text information.

The existing feature word extraction algorithm based on word frequency statistics is not suitable for the characteristics of MicroBlog text information with short language and various expressions. According to the particularity of MicroBlog text, the sub-system of MicroBlog information mining has designed a suitable algorithm.
3.2 Automatic Feature Word Algorithm Model
MicroBlog text information generally contain only a few dozen words or sentences. Therefore, the extraction of feature words for MicroBlog text information is actually equivalent to feature word extraction for several sentences.

The idea of building an automatic feature word algorithm model is based on the weighting operation of each word, and then by comparing the weights, the first words are extracted as feature words. In the design of automatic feature word model, the MicroBlog information mining subsystem is mainly realized through the following two steps:

1. According to the characteristics of Chinese sentences, the part of speech, word frequency and position are taken into account to extract the information feature words.
2. Comparing with the existing feature lexicon to extract the specified feature words.

The automatic feature word algorithm model is shown in Fig.4.

3.2.1 Modeling Factors of Automatic Feature Word Algorithm
The factors that affect the weight size of a word are the main factors of the automatic feature word algorithm model construction, as follows:

1. Part of speech. In the MicroBlog text, words that can identify the characteristics of the MicroBlog text is mostly the notional words in the MicroBlog text, such as nouns, verbs, adjectives and so on. However, some function words in the text, such as interjections, prepositions, conjunctions and so on, have not contributed to the identification of the characteristics of MicroBlog text. Therefore, when extracting the feature words of MicroBlog text, we should first remove these useless function words. Among the notional words, nouns and verbs have the strongest expressive power for the characteristics of MicroBlog text, so the weights of nouns and verbs in MicroBlog text should be the highest, followed by adjectives.

2. Word frequency. It is undoubtedly very important that there are words with high word frequency in the short tens of words of MicroBlog text. It also has a strong ability to distinguish MicroBlog text. When designing the algorithm, the word frequency should be positively correlated with the weight of words.

3. Syntactic structure. The MicroBlog text generally consists of several sentences or a compound sentence. According to the attributes of the sentence, declarative sentences have strong distinctive ability in MicroBlog text, while interrogative sentences and exclamatory sentences are not representative in content. Therefore, when designing the algorithm, the words contained in the statement should be higher in weight.

4. Word length. Long words in Chinese words often reflect more specific and subordinate concepts, while short words often represent relatively abstract and superordinate concepts. Short words have higher frequency and more meanings and are function-oriented, while long words have lower frequency and are content-oriented. Increasing the weight of long words is conducive to truly reflecting the importance of long words in micro-blog text.

5. Initial appearance position. Through simple statistics, it can be found that keywords generally appear earlier in the article, so the weight of the candidate words which appear in the front position should be increased.

6. Existing feature lexicon. If a word has appeared as a feature word, it shows that the word has attracted the attention of the system. The representative meaning of the word is relatively strong, and the weight of the word should be increased accordingly when designing the algorithm.

3.2.2 Formula of Automatic Feature Word Algorithm
The calculation formula of automatic feature word algorithm is designed as follows:

If the relevant factors information of a word i are as follows:

- Part of speech X_i: If it is a noun and verb, X_i=2, if it is a adjective, X_i=1, if it is other words, X_i=0;
- Word frequency P_i: P_i is equal the occurrence number of word i;
- Word length C_i: C_i is equal to the letter number contained in word i;
3.2.3 Extraction of Automatic Feature Words

After calculating Qi of all the words in the MicroBlog text, ten words with Qi value in the top ten are selected as the feature words of the MicroBlog text.

4. Similar Clustering Model

4.1 Model Overview

Similar clustering model realizes the similar clustering function of MicroBlog information mining subsystem, that is, to cluster similar MicroBlog information under the same event. Due to the characteristics of short text and various expressions, the similarity between MicroBlog texts is very low, which results in that the traditional similar clustering algorithm based on text similarity comparison can not meet the accuracy requirements of MicroBlog information mining subsystem.

Therefore, the MicroBlog information mining subsystem uses the method of comparing the semantic similarity of the MicroBlog text to complete the clustering. From the above, it can be seen that the feature words are the representative of the MicroBlog text information semantics. Therefore, the MicroBlog information mining subsystem uses the method of comparing the similarity of the feature words to complete the similar clustering operation. [3]

4.2 Model Design

There are two attributes worth considering in the similar clustering algorithm of feature words, one is the meaning of the word, and the other is the weight of the feature words, that is Qi in formula (1). Among them, the meaning of a word is the attribute of the feature word itself, while the weight is the extension attribute.

If the feature words of some paragraphs of MicroBlog text have similar meanings and weights, then these paragraphs are most likely to express the same event.

4.2.1 Semantic Similarity Algorithms

In the design of semantic similarity, the similarity algorithm of feature words under the framework of KDML (Knowledge Database Mark-up Language) is used to calculate the similarity of feature words of several paragraphs.

KDML is a new system of knowledge description norms. It is a common knowledge base which takes the concepts represented by Chinese and English words as descriptive objects to reveal the basic contents between concepts and between the attributes of concepts, containing rich lexical semantic knowledge and world knowledge. In KDML, the description of lexical semantics is defined as semantic item (concepts), and each word can be expressed as several semantic items. Semantic items are sometimes described by a knowledge representation language. The words used in this knowledge representation language are called sememes. The semantic tree of KDML organizes the sememes describing lexical semantics into tree-like structures. According to the attribute relationship between the sememes, the sememes are divided into several sememe trees. There are some relationships between trees, forming the network knowledge structure of KDML.
Let the set of feature words of two paragraphs of MicroBlog texts be $S_1 = \{N_1, V_1, A_1\}$ and $S_2 = \{N_2, V_2, A_2\}$ respectively, where $N$ is the noun set, $V$ is the verb set, $A$ is the adjective set, and $N$ is taken as an example to demonstrate the calculation process of similarity.

$$N_1 = \{w_{11}, w_{12}, \ldots, w_{1n}\}$$

$$N_2 = \{w_{21}, w_{22}, \ldots, w_{2n}\}$$

$wij$ is the feature word of the set, which is the basic element. The formula for calculating the similarity of two words is as follows:

$$\text{sim}_W(w_1, w_2) = \max_{i=1, \ldots, m} \text{sim}_{WS}(s_i, s_j)$$

Let $N_{12}$ be the characteristic matrix of the similarity of $N_1$ and $N_2$ set,

$$N_{12} = N_1 \times N_2^T = \begin{bmatrix} w_{11}w_{21} & \cdots & w_{1n}w_{21} \\
\vdots & \ddots & \vdots \\
w_{11}w_{2n} & \cdots & w_{1n}w_{2n} \end{bmatrix}$$

Among them, $w_{ij, w_{ij}} = \text{sim}_W \left( w_{ij}, w_{ij} \right) ;$

When calculating, we first traverse the similarity feature matrix, extract the word combination with the greatest similarity, then delete its rows and columns from the similarity matrix, and continue to select the combination with the greatest similarity in the remaining matrix until the element in the matrix is zero. At this point, the similarity sequence of the word combination with the greatest similarity can be got.

$$\max L = \{\text{sim}_{W_{\text{max}}}, \text{sim}_{W_{\text{max}}}, \ldots, \text{sim}_{W_{\text{max}}}\}$$

The similarity degree of the noun sets $N_1$ and $N_2$ of $S_1$ and $S_2$ are as follows:

$$\text{sim}_S (N_1, N_2) = \sum_{i=1}^{n} \text{sim}_{WS}(s_i, s_j)$$

Similarly, the similarity degree of verb and adjective sets can be obtained. The similarity simS(S1,S2) of the whole set can be obtained by weighted average of similarity of each set.

$$\text{sim}_S (S_1, S_2) = \sum_{i=1}^{\beta} \beta_i \text{sim}_S (S_i, S_j)$$

Among them, $\beta_i$ is a weighted coefficient, and its value is selected according to linguistic knowledge.

4.2.2 Weight Similarity Algorithms

The weighted similarity algorithm calculates the similarity degree of the weights of feature words according to the maximum semantic similarity feature word combination $L$ formed by the operation of the formula (4) in the semantic similarity algorithm. Let maximum semantic similarity feature word combination $L$ be the following:

$$L = \{w_{i1}w_{21}, w_{i2}w_{22}, \ldots, w_{in}w_{2n}\}$$

The similarity degree $\text{sim}_Q$ of the weights of feature words is:

$$\text{sim}_Q = |Q_i - Q_j|$$

The weight similarity of set $S_1$ and set $S_2$ is:

$$\text{sim}_Q (S_1, S_2) = \frac{1}{n} \sum_{i=1}^{n} \text{sim}_Q$$
4.2.3 Comprehensive Formula

After obtaining the semantic similarity and weight similarity of the feature words, the overall similarity of the two sets can be obtained by synthesizing the two sets. The formula for synthesizing semantic similarity and weight similarity are as follows:

\[
sim_Z = A \cdot \text{sim}_S(S_1, S_2) + B \cdot \text{sim}_Q(S_1, S_2)
\]  

(9)

A and B represent the weights of semantic similarity and weight similarity in the overall similarity.

If \( \text{sim}_Z \) meets the criterion of clustering, it judges that \( S_1 \) and \( S_2 \) belong to the same event, and classifies several paragraphs of MicroBlog text into the same event, and takes the subject term of the feature word in the similar clustering lexicon as the feature word of the event.

4.3 Model Flow Chart

The flow chart of similar clustering model is shown in Fig.5.
5. Information Quantification Model

5.1 Design of Information Quantization Algorithms

Information quantization is the transformation of MicroBlog information. It changes the MicroBlog information from the character carrying language meaning to the value carrying quantitative information, and provides a scientific basis for monitoring and early warning of the system. The quantitative value is the monitoring value, which indicates the attention of quantitative MicroBlog events.

The degree of attention to a MicroBlog event depends on the number of fans of the original author and the forwarding microbloggers and the number of comments forwarded. Among them, the number of comments forwarded is a real concern, which is the hard attention, and the number of fans represents how many people may see this microblog, which is the soft attention. [4-5]

By synthesizing the differences between hard and soft concerns, the significance of forwarding and commenting numbers and the correlation between MicroBlog events and public security work, the quantitative formula of information is deduced, which can be expressed as follows:

Let a MicroBlog event be \( k \), the MicroBlog text contained in this event is \( j \), the forwarding commentator of \( j \) is \( i \), the number of forwarded comments of \( i \) for \( j \) is \( Z_{ij} \), the number of fans of \( i \) is \( F_{ij} \), the monitoring value of \( k \) is \( C_k \), and the correlation degree of \( k \) is \( X_k \);

If \( Z_{ij}=5 \) is taken as a degree of concern, then

\[
C_k = \left( \frac{Z_{ij}}{5} + \frac{F_{ij}}{10} \right) / X_k
\]  

(10)
5.2 Flow Chart of Information Quantization Algorithm
The flow chart of the information quantization algorithm is shown in Fig.6.

6. Conclusions
This paper mainly introduces the relevant algorithms and models in the process of information mining and processing of MicroBlog information mining system.

The system constructs an automatic word segmentation model to semantically divide MicroBlog text strings and analyze them for many times, which lays the foundation of system information processing, extracts the feature words from the MicroBlog text information with automatic word segmentation by using the automatic feature word algorithm to realize the correct extraction and classification of micro-blog information, and constructs a similar clustering model based on the comparative semantic similarity of MicroBlog text to realize the similar clustering function of MicroBlog information mining subsystem.

Finally, the information quantification model is deduced, and the MicroBlog information is quantified to provide a scientific basis for the monitoring and early warning of the system.

Acknowledgments
This work had supported by the Educational Commission of Hubei Province of China (B2019205), the 2020 Research Project of Hubei University Of Police.

References
[1] Leng, J., Zhuge,R.X.,Chen, P.M. (2013) Technical Document of "Massive Information Mining and Monitoring in Microblog Community",136:41-44
[2] Baidu Wikipedia.Chinese word segmentation.(2019) . http://baike.baidu.com/view/19109.htm
[3] Jin B.,Shi Y.J,Teng H.F.(2005) Similarity algorithm of text based on semantic understanding. Journal of Dalian University of Technology,45(2):291-297.
[4] Zhao Y.(2009) Brief Analysis of public sentiment supervision technology in America. Police Technology, 25(6):19-21.
[5] Cheng,J.S.(2011)Talking about the influence of microblog technology on mass incidents.National Defense, 24(5):59-61.