Model analysis of intelligent data mining based on semantic segmentation technology

Tianshun Huang
Henan Occupation Technical College
Zhengzhou, China
huangtianshunedu@163.com

Zhiqiang Zhang
Henan Occupation Technical College
Zhengzhou, China

Shengyong Yang
Department of Basic
Yellow River Conservancy Technical Insitute

Zhengzhou, China
Hongyun Lian
Henan Occupation Technical College
Zhengzhou, China

Abstract—Intelligent data mining model is designed from the source database mining association rules that satisfy the minimum support and minimum confidence. Semantic segmentation technology is through the analysis to find out the meaning, structure meaning and significance of the combination, so as to determine the true meaning or concept of the expression language. Phrase structural method is a phrase collocation rules. A sentence is a sentence with the word or phrase rules. The paper presents model analysis of intelligent data mining based on semantic segmentation technology. Experiments show that the proposed method is effective.

Keywords- Segmentation; Data mining; Natural language understanding; Association rule; Semantic network

I. INTRODUCTION

Natural language is not immutable and frozen dead language, its development in the social life, the change of mutual influence between different languages and different variants of the same language of the people. A word, a phrase may suddenly pop up overnight; population structure change special could lead to a new language or new language varieties (such as the emergence of dialect). These are required to understand natural language computer program must have the adaptability to external language environment.

Grammar is the language of the organization. The rules of grammar restricts how the morphemes words, words form phrases and sentences. Language is formed in the restrict relationship in this tight. Use of morphemes word rules called word formation rules. Another part is the syntactic grammar [1]. The syntax also can be divided into two parts: the phrase structure method and syntax. Phrase structural method is a phrase collocation rules. A sentence is a sentence with the word or phrase rules.

Database technology has achieved decisive results and has been widely applied. However, database technology as a basic way of information storage and management, still with the on-line transaction processing as the core application, the lack of mechanisms to support advanced features, such as decision analysis, prediction. As is known to all, with the expansion of the size of the database, especially data warehouse and Web model data sources become more and more popular, online analytical processing, decision support and classification, clustering and other complex application becomes inevitable.

The analysis and understanding of the process of language is a hierarchical process. Modern linguists put this process is divided into 3 levels: lexical analysis, syntactic analysis and semantic analysis. If the received voice flow and it is the 3 levels should also join a speech analysis layer. Although this is not between the layers are completely isolated, but this division of hierarchical do indeed help to better reflect the composition of the language itself.

Widespread concern in data mining research interests' appeal to experts and from commercial manufacturers mainly lies in the extensive use of large data system and the conversion of data into useful knowledge urgent need. In order to meet the requirements of electronic information, information technology has been from simple file processing system to effectively change the database system. Research and development of the database system of three main patterns hierarchical, network and relational database has made important progress. The paper presents model analysis of intelligent data mining based on semantic segmentation technology.

II. THE SEGMENTATION TECHNOLOGY BASED ON NATURAL LANGUAGE UNDERSTANDING

Each register ATN consists of two parts: the syntactic features and the syntactic function of register. The characteristics of registers, each dimension feature has a feature name and a set of feature values, and a default value to represent. Function register reflects the relationship between syntactic constituents and functions. Each node of the tree analysis has a register, the upper half part is the characteristics of register, the lower half part is function register.

The main task of lexical analysis is left to have every character of the input string to be scanned, produces a sequence of words, for syntax analysis. The formal ceremony for description (description) of the word structure is very simple and convenient. And put a formal compilation (or conversion) as a NFA and then converted into the corresponding DFA, the NPA or DPA is the

regular expression recognizer for the language represented by sentences.

Chinese word segmentation is the Chinese analysis and computer processing Chinese difficult, the cause of Chinese word segmentation accuracy is not high in general are: words (or Chinese analysis basic unit) definition, scope, word segmentation dictionary because the algorithm generated ambiguity problem. Causes of ambiguity in the process can be attributed to the following three categories: (1) the ambiguity caused by the two semantics of natural language, known as the first category ambiguity [2]. The two segmentation forms both in syntax and semantics are correct, is artificial word segmentation will also cause ambiguity, only the combination of context will give the correct segmentation. (2) Special ambiguity produced by machine automatic segmentation, known as the second types of ambiguity, as is shown by equation (1).

$$\Sigma_{\varepsilon(k)} = \mathrm{diag}[\sigma^2_{\varepsilon_1(k)}, \sigma^2_{\varepsilon_2(k)}, \mathrm{K}\ , \sigma^2_{\varepsilon_q(k)}] \tag{1}$$

Linguistics to these uncertainties is called "ambiguity". Ambiguous language unit of ambiguity itself gets resolved through generally cannot, and must rely on the larger linguistic units and non language environmental background factors and common sense to solve. Human beings have to rely on the overall elimination capacity of the local uncertainty and common sense reasoning ability is very strong, reflected in language is the use of disambiguating context information and knowledge ability. The ability to obtain the same powerful computer is engaged in the natural language understanding scholar dream goal.

The k- dimension tree is a representation, it is a kind of decision tree, the tree, (a) may answer set by points, and one point is likely to the nearest point. (b) Each test required neutral zone a coordinate, a threshold and a threshold around the sub point of. (c) Each test according to which side of each point in the threshold and the collection of points are divided into two groups.

Many methods of automatic syntactic analysis, a phrase structure grammar, case grammar and functional grammar augmented transition network, etc.. The largest unit parsing is a sentence. Correlation analysis aims to find out the words, phrases and sentences respectively in the role, and in a hierarchical structure to express. This hierarchy can be reflected directly subordinate relationship, the relationship between components, but also the relationship between grammatical function.

LFG description of the sentence is divided into two parts: the direct component structure (Constituent Structure, referred to as C-Structure) and function structure (Functional Structure, referred to as F-structure), the C-structure is generated by a context free grammar surface analysis results. On this basis and it is through a series of algebraic transformations to generate F-structure. LFG uses two kinds of rules: add a subscript context free grammar rules and lexical rules is equation (2) [3].

$$\overline{w}(m) := \left[\overline{w}^T(m,1), \overline{w}^T(m,2), \Lambda\ , \overline{w}^T(m,M)\ \right]^T \tag{2}$$

If a node n has at least one it except his own sons, and marked A, A positive in VN; if the node n (marked A) direct descendants, from left to right is the order of the node N1, N2, n3,... And NK, the marked A1, A2, A3,... So, Ak, A1 A2, A3 A 61664,... , Ak must be a production in R.

According to the phonological rules, from the speech stream distinguish individual phonemes, then according to the morphological rules to find a phonemic syllable and the corresponding morpheme or word. Lexical analysis, analysis aims to find each morpheme words, obtain the linguistic information from. Syntactic analysis, structure analysis of sentences and phrases, the purpose is to find out the mutual relationship between words and phrases and their role in the sentence. The semantic analysis, the purpose is to find out the meaning of a word, structure analysis and its significance in combination with meaning, so as to determine the true meaning or concept of the expression language.

Research on concept learning to follow two different routes, namely, there are two different points of view. One is the concept of project learning method based on the learning mechanism, it starts from the possible (regardless of whether the mechanism exists in the life within the organization), try to test and determine the engineering method of concept learning. The other is a concept learning based on cognitive modeling, to develop computational theory of human concept learning. But only from the engineering point of view to study the concept is based learning [4]. The first task is to construct the type definition of concept learning.

LFG description of the sentence is divided into two parts: the direct component structure (Constituent Structure, referred to as C-Structure) and function structure (Functional Structure, referred to as F-structure), the C-structure is generated by a context free grammar surface analysis results. It is on this basis through a series of algebraic transformations to generate F-structure. LFG uses two kinds of rules: add a subscript context free grammar rules and lexical rules, as is shown by equation (3).

$$x_i \approx \sum_{j=1}^{m} c_j^{'} u_j \quad i = 1, 2, \Lambda\ , M \tag{3}$$

Semantic analysis, just according to the part of speech information to the analysis of a sentence is grammatical structure, its correctness can not be guaranteed, this is because the grammar structure of some sentences, need the help of semantic information to determine, that is to carry out the semantic analysis. A simple method of semantic analysis is to use semantic grammar. The so-called semantic grammar is based on the traditional phrase structure grammar, N (noun), V (verb) the concept of grammatical categories, with the discussion of specialized classes to replace the field.

Syntax analysis method, analysis method, from left to right: always from left to right to identify the input string of symbols, first identifying string of symbols in the left most symbol, a symbol to identify the right. Analysis method from right to left: always from left to right to identify the input string of symbols, first identifying symbol string in the rightmost symbol, a symbol and

recognition on the left. Top down approach: also known as top-down analysis method, analysis method of object oriented.

The main idea of natural language understanding: 1 the cognitive view, regard human beings as a kind of advanced information processing system, emphasizing the research for human intelligence activities and in computer simulation and implementation [5]. 2 the pragmatic perspective, language as a medium of communication between people and people, any understanding a discourse or it cannot be separated from the discourse context and discourse exist before and after the user's psychological background, as is shown by equation(4).

$$\tilde{V}_j = \tilde{V}_J \bigoplus_{k=0}^{J-j-1} \bigoplus_{\lambda=1}^{3} W_{J-k}^{\lambda} \tag{4}$$

Node A said that the whole structure, it consists of two parts: a node B (wedge) and node C (brick). The two arched structure description. Except the top object type (a top brick, the other is a wedge) different outside, other descriptions are the same. But should pay attention to two support relationship, except for the left and right chain, also used is not connected with the chain that is not connected with the two supporting body. If two objects have a surface connection, and a common boundary, says they are connected. Phase relationship is the key to define the arch, as is shown by Fig .1.

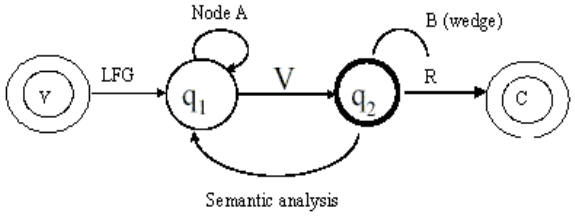

Figure 1. Segmentation technology based on natural language understanding.

Analysis of the sentence procedure is as follows: (1) using LFG syntax analysis to obtain C-structure with context free grammar, do not consider the subscript grammar; the C-structure is a direct component tree; (2) each non leaf node is defined as a variable, according to the subscript vocabulary and grammar rules in the establishment, function description (a set of equations); (3) as algebraic transformations on equations, calculated each variable, gain of function structure of F-structure.

The establishment of syntactic structure is language understanding one step further in the model, is required to obtain the expression of the meaning of language. The first step is to determine the expression of each word in the sentence meaning, which involves the ambiguity of meaning and syntactic structure, such as the English word goes can have more than 50 kinds of meaning. But even a word a lot, in the context of certain conditions, in the phrase, its meaning is often the only. This is due to the constraints of the reason. The constraint relationship can be used to represent a logical form, by the logic form to obtain the meaning and the meaning of a sentence.

Semantic network with 4 partitions: S0 partition contains some general concepts, such as the dollar, exchange and bolts; S1 partition containing and purchase bolt special entities concerned; S2 partition contains and the pump is fixed on the bench this special operation entities concerned; S3 partition contains a special entity associated with the same fixed operation etc.. Using partitioned semantic network, using its relevance in some level of the partition, you can better deal with the focus problem. When a partition is the focus, is a high-rise partition elements becomes observable.

The difference between the establishments from the syntax tree can be a good way to understand the two kinds of methods. The top-down approach is to start from the grammar symbols, to turn it into a syntax tree roots downward, gradually establish a syntax tree, the end node symbol syntax tree series is just the input string of symbols; a bottom-up approach is to begin from the input string of symbols, to do it the end node symbols for syntax tree series, from the bottom to the structure the syntax tree.

III. INTELLIGENT DATA MINING TECHNOLOGY

Data mining system to mining knowledge in relational database, the transaction database, spatial database, data warehouse, data such as text, WEB data organization form, then the knowledge discovery in database is just one aspect of data mining. This is the early. This is the early popular view, in many literature can see this argument. Therefore, from this sense, data mining is the mining process of useful knowledge from database, data warehouse and other data storage mode.

ID3 algorithm is the mutual information in information theory as a measure of the borrowing capacity of single attribute, the heuristic function is not optimal, and the main problems are: there are many characteristics and the calculation of mutual information depends on the attribute values, and this property is not necessarily the best; ID3 is the non incremental learning algorithm; the anti noise worse, the positive and negative training examples are more difficult to control, as is shown by equation (5).

$$I = -\sum_{t=0}^{T} q(t) \log q(t) - (-\sum_{t=0}^{T} p(t) \log p(t)) \tag{5}$$

Data mining is based on the mining object can be divided into: Mining relational database, object-oriented database, spatial database, temporal database, text data source, multimedia database, heterogeneous database, heritage database and web object [6]. Based on the mining methods can be divided into: machine learning method, statistical method, the method of clustering analysis, exploratory analysis, and neural network.

In order to improve the efficiency of generating frequent itemsets by layer, an important property called Apriori properties used to compress the search space. Properties of Apriori: all nonempty frequent item set must also be frequent. The Apriori property is based on the following observation. By definition, if the set I does not meet the minimum support threshold min_sup, and it is not frequent, namely P (I) <min_sup. If A is added to the set I, then the result of item sets (I A) may not occur more frequently than I. Therefore, I A is not frequent, namely P (I A) <min_sup.

The data source must be real, large, noisy; found is interested knowledge discovery; knowledge to be acceptable, understandable, can use; not required to discover universal knowledge, only specific problems found support [7]. Form data, information is also knowledge, but there is more to the concept, rule, model, rules and constraints such as knowledge. People take the data as the formation of the source of knowledge, as if from ore mining or gold as. Raw data can be structured, such as data in a relational database; it can also be a semi-structured, such as text, graphics and image data; and even heterogeneous data distribution in the network.

$$x = \left(\prod_{i=1}^{n} x_i^{w_i} \right)^{(1/\sum w_i)} \qquad x_i > 0 \tag{6}$$

With S% in W transaction also support items set A and B, S% called support association rules from A to B. Support the degree of probability description of the A and B of the two items set union C in all affairs appeared to have much. If one day a total of 1000 customers to the mall to buy items, of which there are 100 customers at the same time to buy a hammer and nails, so support association rules above is 10%.

The first one scan data set, each a calculating the support of itemsets, according to a given minimum support degree min value, get a frequent set L1.

Then through the connection operation, get two candidates, again scan data set for each candidate set, obtains each candidate set of support, and then compared with the minimum support degree. Two frequent set L2. And so on, until can't connect to generate new candidate set so far.

Data mining system must be faithful to the source data is for large capacity database to store data set. So, after the expansion of the sample set can be difficult to accurately and effectively realize the concept of "covers all the positive samples but does not cover any negative samples" sums up [8]. According to the probability statistics method, get the concepts in the test portion of the positive samples or negative sample case description. Therefore, one of the key problems to maximize the use of induction is the sample must be solved. Second, the data mining system, the positive samples from the source database, and the negative samples is not likely to be directly stored in the source database, the concept but lack of contrast class information induction is not reliable, as is shown by equation(7).

$$X_2 = \left[a(\theta_1)e^{j\phi_1}, \quad \Lambda, \quad a(\theta_N)e^{j\phi_N} \right] S + N_2 = A\Phi S + N_2 \tag{7}$$

Adopting the only support and it will not consider all attributes of different important degree in. In real life, some affairs happen very frequently, and some affairs is very sparse, so to mining is the problem: if the minimum support threshold set too high, although the speed, but cover less data, meaningful rules may not be found; if the minimum support threshold setting too low, so a lot of nonsense rules will fill in the whole data mining process, greatly reduces the availability of the mining efficiency and rules. The formulation of this will affect or even mislead decision.

The formation of concepts at different levels is by these rules in the process of data mining abstract. The concept of hierarchical structure should be determined by specific background knowledge, by field experts and knowledge engineers organized into a suitable form (such as concept tree, the queue or rules) and input into the model library. General knowledge of data mining system in the concept hierarchy according to the layered structure of automatic on specific summed up from the database corresponding.

IV. MODEL ANALYSIS OF INTELLIGENT DATA MINING BASED ON SEMANTIC SEGMENTATION TECHNOLOGY

The original data is not being mining, mining and refining to obtain the needed rules useful for commercial purposes of the knowledge. This is the origin of the name of the data mining. So, from a business point of view, data mining is established according to the enterprise's business objectives, rules of in-depth analysis of many data to reveal the hidden, unknown and its model, so as to support the business decision activities.

The first step of the Apriori algorithm is a simple statistical all items containing one element sets the frequency of the occurrence, to determine the largest collection of one-dimensional project. In step k, divided into two stages, first with a function sc_candidate (Hou Xuan), the article (k-1) support the largest project step generates a set of Lk-1 to generate the candidate item sets Ck. and database searching candidate itemsets Ck calculation.

Linguistics in general will "word" is defined as "the smallest grammar be applied independently, a unit of meaning". Sentences in natural language is composed by words, but the computer to understanding and natural language processing is from the words of the first step. Chinese is different from west, in a Chinese sentence; between the words have no obvious delimiters (spaces). In addition, Chinese lexical restriction is not standardized, and the ever-changing, gave the Chinese word segmentation has brought a lot of trouble.

So ATN is a recurrent neural network. In ATN there is an air arc jump, it does not correspond to a syntactic constituents also does not correspond to an input vocabulary [9]. Each register ATN consists of two parts: the syntactic features and the syntactic function of register register. The characteristics of registers, each dimension feature has a feature name and a set of feature values, and a default value to represent. Such as "feature dimensional number" two "and" singular eigenvalue "complex", the default value may be null.

To satisfy the minimum support degree is k-itemset, then known as the high frequency k- project group (Frequent k-itemset), usually expressed as Large K or Frequent K. Algorithm and from the Large K project group then produce Large k+1, until it can no longer find high-frequency project so far longer group.

Data mining is a data abstraction into an important component of knowledge. It is always based on the model and algorithm specific, focused on the target data normalization, complete knowledge of refining work. In general, it should be repeated using knowledge and user interaction has been obtained, reached the final formation

of customer satisfaction model of knowledge. For mining a multi strategy system, should be designed or choose to include features such as description, association, classification, clustering analysis and evolution and deviation analysis, data mining tools. Intermediate excavated or final knowledge stored in the knowledge base.

In the connection part, Lk-1 and Lk-1 connected to generate a possible candidate (step 1-4). Pruning part (step 5-7) using the Apriori property is deleted non frequent subsets of candidate. Non frequent subsets of test are in the process has_infrequent_subse. With the frequent itemsets can be strong association rules by the following method, for each frequent item sets L, all non empty L generating set, the L of each nonempty set s, if Support_count (L) /support_count (s) is more than or equal to min_conf, then the output rule "s - (L-s)". Where min_conf is the minimum confidence threshold as is shown by equation (8).

$$U_S = \begin{bmatrix} U_{S1} \\ U_{S2} \end{bmatrix} = \begin{bmatrix} AT \\ A\Phi T \end{bmatrix} \tag{8}$$

ATN the lexical and syntax analysis belongs to an enhanced context free grammar, its basic idea is the constituent structure to continue using the context free grammar to describe the sentence; but production adds some functions on individual grammar, is mainly about some necessary grammatical constraints and the establishment of the deep structure of the sentence.

The establishment of syntactic structure is language understanding one step further in the model, is required to obtain the expression of the meaning of language. The first step is to determine the expression of each word in the sentence meaning, which involves the ambiguity of semantic and syntactic structures; the second step is to determine the semantic based on background knowledge. The logic of the form of expression is the structure of a frame; it is the expression of a specific form of case and a series of additional facts.

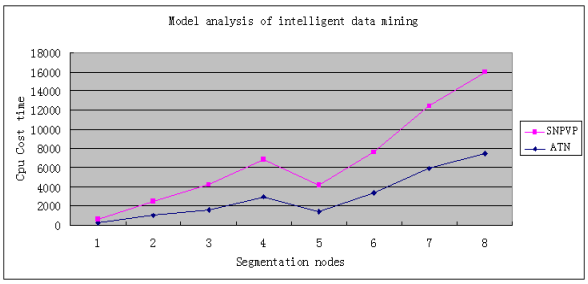

Figure 2. Model analysis of intelligent data mining based on semantic segmentation technology.

The above process if it can get more than a group of solutions, sentence is identifiable and gets more analytical results. Analysis of multiple solutions is obtained indicating the presence of ambiguity in the original sentence, no solution that cannot identify. The analysis process is shown in Fig .2. A girl handed her baby the sentences in toys. The establishment of the equation is as long as the arrow to be replaced by the parent node variables, down with the current node to replace it. The subscript S, NP rules of VP there are two groups: one is (arrow Subject) = down, replace get (x1 V Subject) =x2; another is the up / down x1=x3 =. Equation (x1 V Subject) the meaning of =x2 is "subject of X1 is x2", therefore, the above two equations directly available equations transform x1=x3=[Subject=x2].

In the representation of syntactic structure of each node additional complex marker, not like the phrase structure grammar originally just tagging single marker for a class of syntactic category, not only effectively solved only through the original some language phenomena can be resolved, but also creates conditions for the organic combination of syntactic and semantic analysis.

V. SUMMARY

The paper presents model analysis of intelligent data mining based on semantic segmentation technology. Longitudinal data mining solutions are proposed. The core of this method is application specific, providing complete data mining and knowledge discovery solution. Because the business logic and specific combination, therefore, data mining technology designed to solve some specific problems were used, become a part of enterprise application system. The segmentation method based on rules which is to join the lexical rules, in the segmentation process grammar rules or even semantic rules to improve segmentation quality. Only on the basis of part of speech information to analyze a sentence grammar structure, is its correctness cannot be guaranteed, this is because the grammar structure of some sentences, need the help of the semantic information to determine, that is to carry out the semantic analysis.

REFERENCES

[1] Aliyu Isah Agaie, Masrah Azrifah Azmi Murad, Nurfadhlina Mohd Sharef, Aida Mustapha, "A Proposed Framework for the Development of an Interactive Natural Language Interface to Ontologies", JCIT, Vol. 9, No. 5, pp. 70 ~ 80, 2014

[2] Nurfadhlina Mohd Sharef, Shahrul Azman Mohd Noah, Masrah Azrifah Azmi Murad, "Issues and Challenges in Semantic Question Answering through Natural Language Interface", JNIT, Vol. 4, No. 7, pp. 50 ~ 60, 2013.

[3] Zhijuan Deng, Shaojun Zhong, "A Kind of Text Classification Design on the Basis of Natural Language Processing", IJACT, Vol. 5, No. 1, pp. 668 ~ 677, 2013.

[4] Somboon Anekritmongkol, Kulthon Kasemsan, "SQL Model in Language Encapsulation and Compression Technique for Association Rules Mining", IJIPM, Vol. 4, No. 1, pp. 65 ~ 75, 2013.

[5] JIANG Fei, "Research on Association Rule Mining of Adaptive Genetic Simulated Annealing algorithm", JCIT, Vol. 8, No. 5, pp. 876 ~ 883, 2013.

[6] Nurfadhlina Mohd Sharef, Shahrul Azman Noah, "Natural Language Query Translation for Semantic Search", JDCTA, Vol. 7, No. 13, pp. 53 ~ 63, 2013.

[7] Hu Shu-jie, Shi Zhen-gang, "A New Approach for Color Text Segmentation based on Rough-Set Theory", JCIT, Vol. 8, No. 4, pp. 166 ~ 172, 2013.

[8] Pei Yin, Hongwei Wang, Wei Wang, "Extracting Features for Sentiment Classification: in the Perspective of Statistical Natural Language Processing", AISS, Vol. 4, No. 15, pp. 33 ~ 41, 2012.

[9] Zhang GuoYin, Li Heng, "Research on Semantic Networks Based Online Learning Knowledge Management", JDCTA, Vol. 6, No. 8, pp. 126 ~ 134, 2012