Analysis of knowledge data discovery and mining by construction of natural language understanding system

QiuFen Wang[1, a], HuiLing Guo[2]

[1]College of computer and information engineering, Nanyang Institute of Technology, Henan Nanyang, 473004, China

[2]School of Computer Science and Technology, Zhoukou Normal University, Henan Zhoukou, 466001, China

[a]wangqiufennylg@163.com

Keywords: Natural language understanding; Data mining; Knowledge discovery; Syntax analysis; Tagging

Abstract. Natural language understanding mainly studies the language communicative process simulation of people with electronic computer, and the computer is able to understand and use the natural language of the human society. This paper analyzes the basic components and functions of knowledge discovery and data mining application system. The paper presents analysis of knowledge data discovery and mining by construction of natural language understanding system. The experimental results show that this method can improve efficiency of knowledge discovery system.

Introduction

Language is used for transmission of information representation, the set of conventions and rules, it consists of sentence composition, each statement and consists of words; the composition statement and language, should follow certain rules of syntax and semantics. If not all spoken and written language, such as English, Chinese, French and German, full and effective communication between human beings would be difficult to imagine. Language is with the development of human society and human itself evolving. Modern language allows any one with normal language ability and others to exchange ideas and technology etc.. To study the natural language understanding, must first have a basic understanding of the structure of natural language.

Natural language understanding is the development of linguistics, logic, physiology, psychology, computer science and mathematics and other related disciplines and with a cross subject forms; it can understand spoken language and written language [1]. Include the following several aspects of the content of language comprehension: (1) to understand the sentence correct word order rules and concepts, and can not understand the sentence containing rules. (2) To know the exact meaning and part of speech of words, form, formation. (3) Understand the semantic classification of words and word polysemy and ambiguity. (4) Specify and indeterminate characteristics and all (membership) characteristics.

Data extraction is a key task in knowledge discovery. Based on information and structure make on data, first of all need to the data source and the extraction principle to accurately define the selection. Then the design of the structure is to store new data and accurate definition of it and the conversion of the source data and the loading mechanism, so as to correctly from each data source to extract the required data [2]. The structure and the conversion of information should be used as metadata (Metadata) is stored. The data extraction process, must fully grasp the structure characteristics of the source data, any negligence may lead to the failure of data extraction. In the process of extraction of multiple heterogeneous data sources, may need to source different data format conversion into an intermediate mode, and then integrates them up.

Natural language understanding system, is not the true meaning of the grammatical analysis, and mainly relies on keyword matching technology to recognize the input sentence meaning. In these systems the designer prior to store a large number of contains certain keywords patterns, each pattern

and one or more explanatory (also called the corresponding response type). The basic word segmentation algorithm is simple without word segmentation method is the most basic add rules or statistical method. It is the foundation of other segmentation methods. The system of the current input sentence with the mode matching, one by one, once successful immediately got the interpretation of the sentence. The paper presents analysis of knowledge data discovery and mining by construction of natural language understanding system.

Construction of natural language understanding system

Natural language understanding, commonly known as the man-machine dialogue, refers to enable the computer to according to this kind of language to express meaning to make the corresponding reaction mechanism [3]. It mainly studies the language communicative process simulation of people with electronic computer, the computer is able to understand and use the natural language of human society such as Chinese, English and so on, to realize the natural language communication between man and machine, to some mental work instead of people, including query data, answering the questions, excerpts from literature, compilation of data and all relevant natural language information processing.

Word segmentation based on statistical method is mainly used in the process of word segmentation ambiguity phenomenon in the disambiguation. Based on the statistical method mainly depends on one or more corpus, this corpus is generally the size of training corpus, although small, but there is a certain representative. The method according to the relevant statistical information from corpora in the obtained data (mainly the word frequency and word adjacency relationship between) to guide segmentation, such as: possible segmentation results according to the statistics of the maximum correct segmentation results.

Analysis of natural language text corpus linguistics research in machine readable collection, storage, retrieval, statistics, grammatical tagging, syntactic and semantic, and the application has the functions of corpus analysis, quantitative in language dictionary compilation, the work style of analysis, natural language understanding and machine translation etc. in the field of. Corpus Linguistics (Corpus Linguistics) began to rise. First, it conforms to the large-scale real text processing needs, put forward to the computer corpus based linguistics and natural language processing of new ideas, as is shown by equation (1) [4].

$$WT_f(a,b) = \left\langle f, \psi_{a,b} \right\rangle \tag{1}$$

Syntax analysis is carried on the structure analysis of sentences and phrases. Many methods of automatic syntactic analysis, a phrase structure grammar, case grammar, functional grammar augmented transition network, etc.. The largest unit parsing is a sentence. Correlation analysis aims to find out the words and phrases in the sentence and the respective roles of, and in a hierarchical structure to express.

The most important factor influencing the learning system design is to provide environmental information system. The knowledge base is stored general principles guiding the implementation of part of the action, but the environment provided to the learning system is a wide variety of information. If the quality of the information is relatively high, and the general principles of the difference is relatively small, is learning the easier part processing. If you provide to the learning system is the specific information haphazard guiding the implementation of specific action, as is shown by Fig (1).

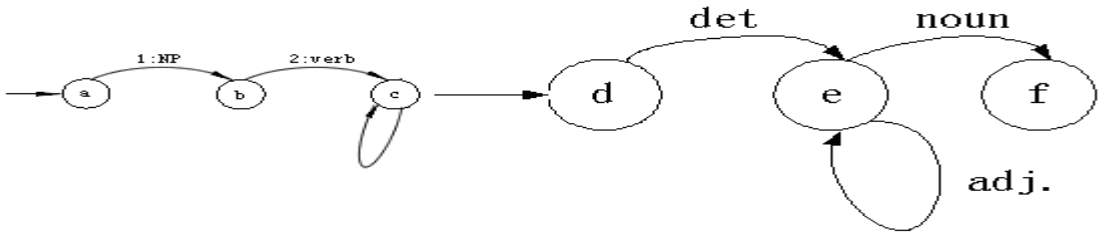

Figure. 1 Transfer network map of natural language understanding

Therefore we need to have a lot of patterns. These patterns can be used to represent the state transition diagram, this is represented by the state transition diagram expression called transfer network, as shown in Figure 1, figure, Q0, q1,... And qT is state, Q0 is the initial state, final state of qT is. Arc is given based on the state transition conditions as well as the direction of transfer. This network can be used to analyze a sentence can also be used to generate sentences.

When the execution part of machine learning system is to solve the problem, the system will remember the problem and its solution [5]. We can put the executive part of learning system abstract as a function, the function of the independent variables in the obtained input values (X1, X2,... , Xn), calculate and output the function value (Y1, Y2,... , Yp). Machine learning in memory simply memory storage of ((X1, X2,... Xn (Y1, Y2),... , Yp)). When you need f (X1, X2,... , Xn), the implementation of part from memory in the (Y1, Y2,... , Yp) simply retrieved instead of re computing it.

ATN is composed of a group of network; each network has a network, each arc on the conditions for extended operating conditions with. This condition and operation using the register method to achieve, placed on the register in each component structure analysis of tree, used to store the syntactic features and syntactic features, and the operating conditions which will be constantly access and set. Mark ATN arc can also be other network tag name.

Knowledge data discovery and mining technology

Data selection can make the back of the data mining and mining focused into task related data subset. Not only improves the mining efficiency, but also to ensure the mining accuracy. We think, the data selection may be made to the target data to be positive to limit or constraint condition, pick and choose those that meet the conditions of the data [6]. We must deeply analyze the application objectives and requirements of the data, to determine the appropriate choice of data or data filtering strategy, in order to ensure the data quality objectives.

R= {I1, I2..... Im} is a group of items set, W is a set of transaction set. In W each transaction T is a group of items, T R. Suppose there is a set of A objects, a transaction T, if A T is called T, transaction support items set A. Association rule is an implication of the following form: A, B, wherein A, B is the two group of goods.

In order to fully understand the knowledge discovery function of the system, we must analyze the ideal application environment of this kind of system and architecture. We can really master the technology, develop the application system to find suitable for their characteristics of enterprise knowledge, as is shown in equation (2) [7].

$$u_i = \frac{1}{\sqrt{\lambda_i}} A v_i \quad i = 1,2,\cdots,r \tag{2}$$

With the deepening of the study of data mining and knowledge discovery, people on the basic process of knowledge discovery and knowledge representation is very clear pattern, and the accumulation of a number of mining model and algorithm [8]. Therefore it cans the emergence of a number of integrated knowledge discovery auxiliary tools set. The integrated software belongs to the general category of auxiliary tools that can help users to complete the work in different stages of

processing knowledge discovery quickly. Using these tools, users can find expert guidance and participation of the corresponding application development in data mining and knowledge.

It is planned to stage KDD of the project, the need to determine the target selection of mining enterprises, knowledge discovery model, compiling knowledge discovery mode get metadata. Its purpose is to enterprise mining target embedded into the corresponding knowledge model.

Analysis of knowledge data discovery and mining by construction of natural language understanding system

Each level on the natural language contains enormous uncertainty [9]. In speech and text level, a more sound, a sound word problems; in lexical and syntactic level, a word class parts of speech, word boundaries, syntactic structure uncertainty; in semantic and pragmatic level, there are a lot of the connotation, extension, reference, the illocutionary meaning, uncertainty, as is shown by equation(3).

$$\begin{aligned}\mu_{s|a} &= E\{s(k) \mid a(k)\} \\ &= \mathrm{M}^{-1}\{\beta(k)^{\mathrm{T}}\Sigma_{\varepsilon(k)}^{-1}\left(a(k)-\alpha(k)\right)+\frac{s_0(k)}{\sigma_{s(k)}^2}\}\end{aligned} \tag{3}$$

Meet the Markoff model between assumed POS tagging results, and using this auxiliary word segmentation, word segmentation and POS tagging complete. So the idea is for a sentence S, there are many types of participle, however the final output results can have only one, so we choose from which an output [10]. Tend to choose the most satisfied segmentation algorithm of this method, which has the maximum probability with the lexical relationship, as is shown by equation (4).

$$\begin{aligned}\hat{\gamma}^{(\eta)}(m,s) &= E\left\{\gamma(m)\middle|\hat{\gamma}^{(\eta)}(m,s-1), z(m,s)\right\} \\ &= \hat{\gamma}^{(\eta)}(m,s-1)+\overline{K}^{(\eta)}(m,s)\left[z(m,s)-\Psi_w(m,s)\hat{\gamma}^{(\eta)}(m,s-1)\right]\end{aligned} \tag{4}$$

Keyword matching method is the simplest method of natural language understanding. The method is simple to sum up as: specified matching and action of two types of samples in the program. Then build a mapping action by matching to sample to sample. When the input statement with a matching to sample match, to perform the provisions of the corresponding sample action, this seemingly machine truly realized the purpose of user questions can understand. KDD knowledge to interpret and use stage, its purpose is to directly output knowledge or integrated into enterprise knowledge base according to user requirements, as is shown by Fig.2. Syntax analysis tree had put forward many different definitions in many natural languages processing program. To provide real-time guarantees MAC layer and other layers of cooperation.

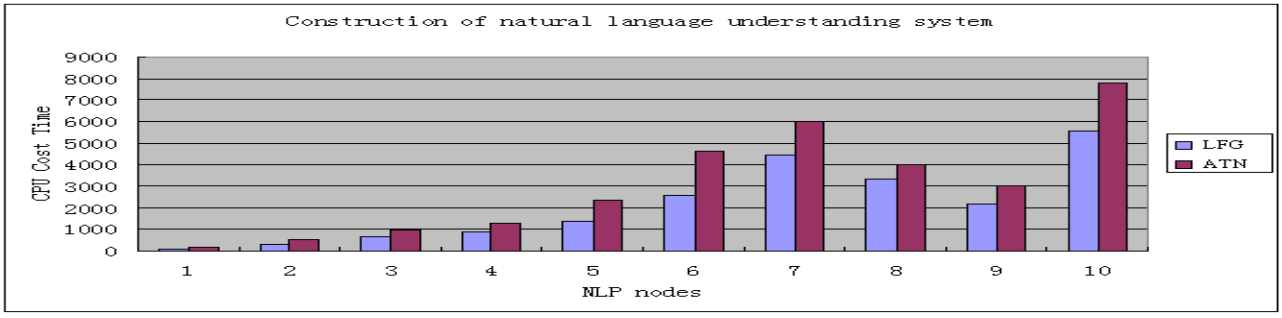

Figure. 2 Analysis of knowledge data discovery and mining by construction of natural language understanding system

LFG uses a structure to express the characteristic, function, and composition of the sequence of vocabulary. ATN grammar and transformational grammar is all directional, ATN syntax and the operating condition requires the use of grammar is a direction, because the register can be accessed

only after being set. An important work of LFG is through the multi-layer description not in conflict with each other to eliminate the order restriction. Before the real call mining algorithm, must be on the enterprise decision-making mechanism and process of a full investigation; understand the urgent problems need to determine the target mining and deliverable system index accurately. Knowledge discovery project data preparation is a time-consuming and laborious work.

Summary

Natural language understanding system requirements developed can handle large-scale real texts, but not as the previous research system that can handle only a few entries, and typical sentence. Only in this way, the developed system has real practical value. In view of the fact that the real understanding of natural language is very difficult, the system does not require deep understanding of natural language text, but to be able to extract useful information. The paper presents analysis of knowledge data discovery and mining by construction of natural language understanding system. KDD data mining phase, users specify data mining algorithm to obtain the corresponding knowledge.

References

[1] Pakawan Pugsee, Wanchai Rivepiboon, "Extracting Serial Verb Denotation for Interpretation Based on Analyzing Related Words", IJIPM, Vol. 4, No. 3, pp. 199 ~ 207, 2013

[2] Yanfeng Jin, Bin Gu, Yongping Wang, Qi Wang, "Knowledge Discovery Scheme based on Certainty Factor Model", JCIT, Vol. 7, No. 22, pp. 280 ~ 287, 2012

[3] Chen Qian, "Research on Set Pair Analysis Model and Data Mining", JDCTA, Vol. 7, No. 6, pp. 454 ~ 461, 2013.

[4] Liu, Tao, Mya Arnold, "Data mining system based on network client", IJACT, Vol. 5, No. 9, pp. 130 ~ 138, 2013

[5] Noor Diana Ahmad Tarmizi, Farha Jamaluddin, Azuraliza Abu Bakar, Zulaiha Ali Othman, Abdul Razak Hamdan, "Classification of Dengue Outbreak Using Data Mining Models", RNIS, Volume 12, pp. 71 ~ 75 , 2013

[6] Lu Jiang, "The Application of Data Mining Technology in the Customer Relationship Management Based on the Customer Loyalty Strategy", IJACT, Vol. 5, No. 7, pp. 89 ~ 96, 2013

[7] Liao Jianping, "Evaluation and Prediction of Credit Assessment Task Based on Data Mining and Neural Networks Approach", JDCTA, Vol. 7, No. 7, pp. 191 ~ 198, 2013

[8] Liu Feng, Wang Jing, "The Strategy Design of the Adaptive PSO Algorithm in Data Mining based on Cloud Computing", JCIT, Vol. 8, No. 5, pp. 840 ~ 848, 2013

[9] Kun Gao, Chen Dong, Yunpeng Liu, "LTKDC: Lattice Theory Based Knowledge Discovery Cloud", JDCTA, Vol. 6, No. 9, pp. 188 ~ 194, 2012

[10] Shu-Meng Huang, Ping-Yu Hsu, Hwynh Nguyen Nhat Lam, "An Attribute-Oriented Induction Approach for Knowledge Discovery from Relational Databases", AISS, Vol. 5, No. 3, pp. 511 ~ 519, 2013