Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2015, 7(3):1129-1136

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

A kind of intelligent question-answering system based on sentence similarity calculation model

Shitao Yan[1*] and Wenqiang Tian[2]

[1]Henan Institute of Science and Technology, Xinxiang, China
[2]Xinxiang University, Xinxiang, China

ABSTRACT

In order to improve the effect of query accuracy in intelligent question-answering system, the sentence similarity calculation model has been a research focus in this field. Based on the analysis of overall implementation framework in intelligent question-answering system, this paper proposes a new type of sentence similarity calculation model based on a linear regression algorithm, which combines the advantages to do sentence similarity computing based on keywords characteristics and semantic characteristics, according to the characteristics to calculate the sentence similarity and obtain a high accuracy.

Key words: Intelligent question-answering system; concept dictionary; sentence pretreatment; sentence similarity computing

INTRODUCTION

With the rapid development of computer technology and network technology, network teaching system is increasing, intelligent answering [1] as an important part in teaching activities is increasingly attracted attention. The advantages and disadvantages of answer effect are mainly reflected in the matching accuracy between inputted questions and the questions in answering-question bank, it is a form based on keywords, relying on an exact match of keywords. In practice, accurate keywords matching method will lead to many problems hard to find the answer, or find no answers; it is difficult to achieve a higher recall ratio and precision ratio [2]. In order to solve these problems, the question method based on natural language is proposed, which brings forward the research of intelligent question-answering system based on sentence similarity.

Because of the characteristic of the western languages, the foreign study to intelligent question-answering system is relatively early, the molding products is also more, such as ASK Jeeves for kids of ASK Jeeves company, the START (Syntactic Analysis Using Reversible Transformation) of MIT artificial intelligence laboratory, the FAQFinder developed by the artificial intelligence laboratory at the university of Chicago, etc. Foreign intelligent question-answering system has strong independence relatively, basically which can be regarded as separate module to meet the retrieval of different types of intelligent question-answering material. Inland basically does not have special intelligent question-answering system early, which is usually via E-mail, message boards, bulletin boards to answer questions from the students, basically avoiding the application of artificial intelligence and expert system technology. In recent years, due to the demand of teaching platform and the breakthrough in Chinese language processing areas, many scientific research institutes also increase the research and development of intelligent question-answering system, the developed system is illustrated as AnswerWeb intelligent question-answering system of ShanghaiJiaotongUniversity, the Aslane Smart answer parts in Beijing normal university vclass teaching platform, etc.

The research of current intelligent question-answering system is still in its beginning stage, there are many problems need to be solved on the basis of the careful research practice. And the proposed intelligent question-answering system in this paper, comprehensively summarizes the various means of the intelligent answering questions at present, analyzes their characteristics, and designs a kind of workflow which makes full use of all kinds of intelligent question-answering resources. And aiming to the characteristics of intelligence and openness of the system, this paper analyzes the key technologies to realize the intelligent question-answering system. From the perspective of the use effect of this intelligent question-answering system, which can fully meet the needs of intelligent answering questions in the remote education system, in some sense, intelligent question-answering system can also play the role of a good teaching system.

THE OVERVIEW OF INTELLIGENT QUESTION-ANSWERING SYSTEM

Usually intelligent question-answering system generally includes keywords library and the system knowledge base, in the implementation which includes dictionary building, question sentence pretreatment, sentence similarity calculation [3], and matching display and so on.

Keywords repository is used for user’s questions professional keywords extraction, heuristic questions (i.e., key words extension), sentence similarity calculation, etc., at building time which needs to determine the weights of words according to the experience or the related algorithm [4], in this way the built keywords library can better supply service for the next phase of sentence similarity computing.

Usually the weights calculation can adopt the following process:First, in generally the general keywords weight isaccording to the part of speech to determine the value, here it is defined as the weight W , the value range is (0, 1);Second, the weight of professional keywords is defined as W' .

The importance of professional keywords has such a characteristics: usually the number it appears in some sections is more, to the professional keywords that less frequently appears in the whole subject knowledge, the word includes more amount of information, it is more important. Therefore such professional keywords are given higher weights.For example: if in a question about math there are "limit" and "double integral" these two major keywordsat the same time, "double integral" can be easily positioned to "integration" this chapter; and "limit" is hard to place its position, because "limit" almost can be found in each chapter of mathematics.Thus theprofessional keywords more often occurs in the small scale, the more is important, the term should be given a higher weight.

Supposing Q is the calculated weight of professional keywords according to the characteristics of above professional keywords, in order to ensure the calculatedweight of professional keywords is bigger than the weight of general keywords, the calculation expression of professional keywords weight is adjusted as: $W'=Q+1$; the calculation method of Q is as follows, here three kinds of circumstances are listed:

The first case, when a professional keywords appears in many chaptersof course, the chapter number it appearing is noted as n_1 , and $Q = \frac{n_1}{total\ chapter\ number}$ .The second case, when a professional keywords only appears in the multiple sectionsof a chapter in course, the appearing section number is noted as n_2 , and $Q = \frac{n_2}{the\ section\ number\ of\ chapter}$ .The third case, when a professional keywords only appears in the multiple paragraphs of a section in a chapter, the appearing paragraph number is noted as n_3 ,and $Q = \frac{n_3}{the\ \text{paragraph number of a section in a chapter}}$.

At the same time, in these three cases, according to the characteristics of professional keywords from the previous analysis, there is: the Q value of the third case > the Q value of the second case > the Q value of the first case.

In Chinese, the questions include interrogative sentences, and rhetorical question, the purpose of interrogative sentences is mainly used to ask opposite side, this is also the most common form in the answer system. Usually in processing it needs to determine the question type, do word segmentation for the question, keywords extraction and extension analysis, etc. What needs special attention is that, for the production of relevant problem sets, association rule is mainly used to do mining to the user question logs, the question logsare basically consistedby the user’s questions and therelevant answers in intelligent answer library, using association rules can mine thecorrelation degree between the user's question and the answer in the intelligent library, so that an effective heuristic questions can be bring forward for the user's question, which can represent the system intelligence.

The keywords of questions are obtained after preprocessing, according to the weight distribution [5], generally can represent the whole question, at this time needs to find matching in the system knowledge base, so as to find the answer of the knowledge base and display. This process can be the important link of the whole question-answering system, directly affects the overall performance of system. This article focuses on analyzing the commonly used algorithm of sentence similarity calculation model, the principle and implementation method of each algorithm, which combines with the advantages of two kinds of commonly used algorithms, puts forward a new kind of sentence similarity plane calculation model, and applies it in the intelligent question-answering system.

THE IMPROVED SENTENCE SIMILARITY CALCULATION MODEL

Sentence similarity (also called statement similarity) calculation in the field of Chinese natural language processing has a very wide application background, which refers to the similar degree of two statements. Similarity reaches to a certain threshold value; the two statements are regarded as similar. This section based on the analysis of two kinds of sentence similarity calculation model commonly used in the intelligent question-answering system, proposes the new sentence similarity computing model based on the implementation of linear regression algorithm [6], which makes the objective function to be optimal within certain scope, greatly improves the accuracy.

THE COMPUTING MODEL BASED ON MORPHOLOGICAL, WORD ORDER AND WORD SIMILARITY

According to the sentence similarity, from the view of morphology similarity, statement length similarity, word order similarity [7,8], sentence similarity computing is analyzed, and based on this three characteristics the new sentence similarity calculation model is proposed. Conjugated similarity plays a main role, and the statement length similarity plays a secondary role, and the role of word order similarity is minimal.

A statement $S(Sentence)$ in language L (for example, the Chinese characters) is an ordered set of the single word and special symbol (hereinafter referred to as the single word) in L. The length of S is the number of single words in S, here it is expressed by $Lenght(S)$. $SameWC(S_1,S_2)$ expresses the number of same single words in S_1, S_2,and Y, when the occurrences number is different in S_1, S_2, the less occurrence number will be counted. The similarity degree $WordSimilar(X,Y)$ in the statements S_1, S_2 is decided by the following formula: $WordSimilar(X,Y) = SameWC(X,Y)Max(Lenght(X),Lenght(Y))$. It is easy to obtain $WordSimilar(X,Y) \in [0,1]$, its meaning is that the more are the same words in two statements, the more similar are the two statements. $Lenght(S_1)$ and $Lenght(S_2)$ respectively represent the lengths of the statement S_1 and the statement S_2, that is, the number of single words in two statements. Statement length similarity $LenghtSimilar(S_1,S_2)$ is determined by the following formula, $LengthSimilar(S_1,S_2) = 1 - \frac{Abs(Length(S_1) - Length(S_2))}{Length(S_1) + Length(S_2)}$.It is easy to know $LengthSimilar(S_1,S_2) \in [0,1]$, which means the more close are the lengths between the two statements, the more similar are the two statements.

$OrderOccur(S_1,S_2)$ represents the single word set that appears only once in S_1, S_2. $PFirst(S_1,S_2)$ represents the vector consisting of the position serial number of single words in S_1, $PSecond(S_1,S_2)$ represents the vector generated by the component in $PFirst(S_1,S_2)$ according to the corresponding word in S_2 according to the sequence.

In intelligent question-answering system, after user inputs a problem, at first problem matching is preceded in the FAQ library [9], if matching degree reaches the expected value, the corresponding answers will output, intelligent answering questions will be over. If there is no matching, the problem of matching can be regarded as an application in similarity calculation of two statements. Supposing the input question is $Query$, problem set is $Qset$ in the intelligent question-answering library, q is a question statement in the library, $q \in Qset$, and the process of problem matching can be described in the following formula: $Query' = MaxSimilar(Query, q)(q \in Qset)$. Therefore, $Query'$ represents the most similar statement found in the intelligent answering questions library with the inputted question. According to above formula, finding the most similar statement with $Query$, it is needed to compute all the similarity between $Query$ and intelligent answering questions library, the biggest on is selected. If the traversing method is adopted [10], many statements with very low similarity or 0 similarity with $Query$ will participate in the calculation, the algorithm efficiency is low, and it is also affected by intelligent question-answering library.

THE SENTENCE SIMILARITY CALCULATION MODEL BASED ON KEYWORDS FEATURE AND THE SEMANTIC DISTANCE

Sentence similarity calculation based on the keywords characteristics, it is through the use of all effective words in two sentences (remove stop words) to form the vector space [11], then calculate the vector of two sentences, and use the included angle cosine between two vectors as sentence similarity. There are two sentences S_1 and S_2, their vector space is $V = (X_1, X_2, \cdots, X_n)$ which is consisted of all effective words, hereinto, X_n is an effective word. There is a vector $V_1 = \{\omega_1, \omega_2, \cdots, \omega_n\}$ in sentence S_1, in the expression, ω_n is the appearance number of effective word X_n in sentence S_1. In sentence S_2, there is a vector $V_2 = \{\psi_1, \psi_2, \cdots, \psi_n\}$, in the expression, ψ_n is the appearance number of effective word X_n in sentence S_2.The similarity of two sentences is as follows:

$$sim(S_1, S_2) = \overrightarrow{V_1} * \overrightarrow{V_2} = \frac{\sum_{i=1}^{n} \omega_i \psi_i}{\sqrt{\sum_{i=1}^{n} \omega_i^2} * \sqrt{\sum_{i=1}^{n} \psi_i^2}}$$

This method is simply using the word surface information, the effect is better for some corpus with less content correlation. But this method does not consider the meaning of the word itself and the syntax information, thus it has certain limitations.

Sentence similarity calculation based on the semantic distance, needs certain knowledge resources of word meaning as the foundation. Computing the semantic similarity among sentences needs to make sure what is the word meaning expressed by the word in this sentence.

The specific method is as follows:

Supposing there are two sentences M and N, M contains the words M_1、M_2、$\cdots M_m$, N contains the words N_1、N_2、$\cdots N_n$. Then the similarity between words $M_i (1 \le i \le m)$ and $N_i (1 \le j \le n)$ can be expressed in $Similar(M_i, N_j)$. Thus the similarity of two arbitrary words in two sentences is obtained, and the semantic similarity $Similar(M, N)$ between M and N is

$$Similar(M, N) = (\frac{\sum_{i=1}^{m} a_i}{m} + \frac{\sum_{i=1}^{n} b_i}{n}) / 2,$$

In the expression, $a_i = \max(Similar(M_i, N_1), Similar(M_i, N_2), \cdots, Similar(M_i, N_n))$

$$b_i = \max(Similar(N_i, M_1), Similar(N_i, M_2), \cdots, Similar(N_i, M_n))$$

In similarity calculation, this method fully considers the deep layer information of each word in the sentence, the words that surface meaning is different and the deep layer meaning is same are excavated [12], and the similarity calculation is not recognized based on keywords feature. But as a result that the dictionary is not comprehensive and some unregistered word meaning codes are missed, all of these bring forward the calculation error.

IMPROVED SENTENCE SIMILARITY PLANE CALCULATION MODEL

An object can be described from the views of linear and plane, etc., which are corresponding to the space description as one-dimension, two-dimension, etc. Of course the higher is the dimensional number; the description information of the object is more full and accurate. If the sentence is described in accordance with the word sequence, it is one-dimensional linear space; if each vector in the sentence is expressed according to word meaning, it is equivalent to two-dimensional space. Comparing two kinds of forms, the description of sentences from the view of two-dimension is similar to the hologram, can make the information that sentences contain more accurate and more comprehensive.

The method based on keywords feature embodies the surface information of the sentences; the similarity method based on semantic distance embodies the semantic information of every word of a sentence in deep layer. Therefore the keywords feature and semantic feature of sentences foster strengths and circumvent weaknesses, complement with each other, and describe a sentence together, thus according to these characteristics calculating the similarity among sentences can obtain the higher accuracy. Here involves how to determine the weights of these traits, in natural language processing, a lot of problems are determined according to the experience value. But this paper sets up a mathematical model according to the problem, introduces a linear regression algorithm to simply solve, so that the objective function in a certain range can reach the optimal. The linear regression algorithm is as follows:

Through the above discussion, from two sides the information contained in a sentence can be described, i.e., the keywords characteristics (KW), semantic characteristics (SE) [13, 14], and these characteristics can be combined together to do the calculation of sentence similarity, so as to get a more accurate method of similarity calculation. Thus the sentence similarity computation formula of plane calculation model is as follows:

Formula one is two similarities multiplying and doing a square root, $Similar(S_1,S_2)=\sqrt{Similar1*Similar2}$. Formula two is respectively two similarity multiplying coefficient and adding together($\alpha+\beta=1$), $Similar(S_1,S_2)=\alpha*Similar1+\beta*Similar2$.

In the expression, $Similar1$ represents the similarity value of S_1 and S_2 based on the words characteristics, $Similar2$ represents similarity value of S_1 and S_2 based on semantic features, the objective function is to find a set of possible parameter combination $\{\alpha,\beta\}$ to make the similarity calculation more accurate, hereinto $\alpha+\beta=1$. In order to calculate α and β, at first the scope of parameter selection of α and β is limited as $(0,1)$, so how to get the value of α and β? Here regression analysis (its main use is predicting, that is, giving some values of the independent variables, to get the corresponding point estimation and interval estimation) is used to obtain the value of α and β, the specific calculation process is as follows:

Supposing $Similar1$ and $Similar2$ are two common variables, $Similar$ is an random variable, and $X=Similar1$、$Y=Similar2$、$Z=Similar$,to any given set of values $(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)$ of X and Y to do experiment, the observed values $Z_1,Z_2,\cdots,Z_n$ corresponding to random variable Z are obtained, thus n pairs of data can be obtained, $(X_1,Y_1,Z_1),(X_2,Y_2,Z_2),\cdots,(X_n,Y_n,Z_n)$ referred to as a set of samples with capacity n, n pairs of data are distributed in the space, which is called a scatterplot [15]. A scatter diagram visually presents the trend of n given points. For the relationship between automatic grading and human raters, the binary linear regression can be used, assuming to each value of X, there is $Z=\alpha X+\beta Y$.

Among them, constants α and β have nothing to do with X,Y, $Z=\alpha X+\beta Y$ is known as the regression equation; α and β are regression coefficients. The purpose is to use the sample to estimate the value of α and β ,the estimated values are obtained as α' and β'. $Z'=\alpha' X+\beta' Y$ is called experience regression equation. Using this regression equation can do prediction. The least square method [16] is used to solve the regression

equation. According to a set of sample values $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$, the least square method is used to solve the value of α and β.
The first formula is suitable for integrating the two "and" relationship factors, and the second formula is more suitable for integrating the two "or" relationship factors. Here the latter is selected, because two similarities are complementary relationship, are relatively independent. As long as one of the sentences has high similarity, two sentence similarities can be regarded as high, rather than both sentences have high similarity. And choosing the latter can change the proportion of two similarity that α and β dynamically adjust. Of course, each similarity has its own advantages and disadvantages, only proper combination can fully play their respective advantages to get the best system performance.

CONTRAST TEST

The calculation model based on morphological, word order and word length similarity, the sentence similarity calculation model based on keywords characteristics and word meaning distance, and the improved sentence similarity calculation plane model are used in comparative experiments, 100 testing sentences are selected , the similarity values are greater than or equal to 0.7, the test results are shown in table 1.

TAB. 1:Experimental comparison results.

Method	Test the sentences	The sentences with correct result	The correct rate（%）
The similarity based on the morphological, word order and word length	100	63	63
The similarity based on the keywords characteristic and word meaning distance	100	80	80
The similarity based on improved plane calculation model	100	86	86

The test results show that the similarity of plane calculation model compares with other two kinds of calculation method, the accuracy of the query result is the highest, the model can play an important role in calculating the user questions and answers matching application in library, which can make the system accuracy improve greatly.

APPLIATION MODEL TO INTELLIGENT ANSWERING MODULE

According to the improved sentence similarity calculation model mentioned in section 3, sentence similarity can be computed combining with the problem vector $QuestionVector$, and set up the threshold ξ as a certain value, if the sentence similarity is in scope of ξ between the problem and intelligent answering library, then the answer corresponding to the sentence in the intelligent answering questions library will be returned to the user, otherwise the problem will become more creative or in solution repository there is no answer to meet the requirements, the system will automatically transfer the problem to experts distribution function module, automatic store into the intelligent question-answering log repository depending on the type of problems, after the expert gives an answer, this question will serve as a new subject stored in the system problem library, at the same time be sent to the students who bring a question. If the teacher thinks it is necessary, the question can also be massed and sent to the appropriate students.

Frequently asked question set (FAQ) can be used as a component part of the intelligent answering questions. It keeps the questions and the answers that users often ask. To the problems that user inputs, the answer can be first in the FAQ. If you can find the corresponding problem, the corresponding answer of the problem can be returned to the user, and without any question understanding, the sentence similarity calculation, and many other complex process, to improve the efficiency.

Using the frequently asked question library to answer the user problem, the basic calculation procedure is shown in the figure below:

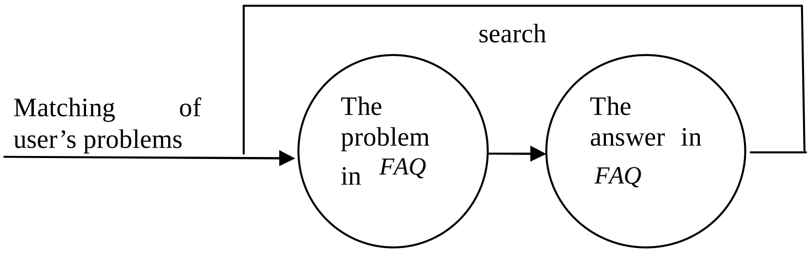

FIG. 1: FAQ Library problem solving process.

When an intelligent question-answering system is designed, there are 6 steps to consist the following main frame. The problem input: refers to input the user's question of natural language that needs to deal with into the answering library or questions bank. The problem analysis: refers to analyze and understand the input questions. Its understanding means mainly includes the keywords analysis method and logical analysis method (through the study of the syntactic analysis of the whole sentence to understand the whole sentence). Question classification: the classification problem now mainly includes: according to the question word (most cases use this way); according to the searching answer type. Because there are different types of problems, the searching keywords is different, the classification method is different also. Keywords information matching and sentence similarity computing: according to the mentioned earlier sentence similarity calculation model to compute the distance between the calculating question sentence and the answering question sentence in the warehouse.Getting the answer: refers to through the matching process (that is the process according to the question type to select the possible answers and assess them, and then by filter return the more accurate candidate answers), to calculate weight of the candidate answer sentence to select the best answer. System evaluation: In a sense, intelligent question-answering system is a knowledge management system. It saves the existing knowledge resources to reasonable structure, and provides the most natural and effective querying means.The implementation process of intelligent question-answering system is shown in figure 2.

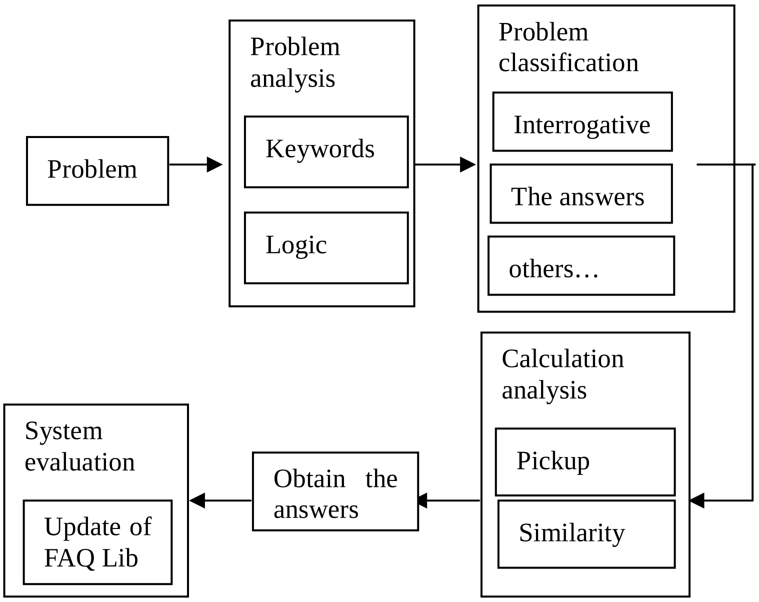

FIG. 2: The intelligent question-answering system process in general.

CONCLUSION

The study on the current intelligent question-answering system still has many problems which need to be solved based on the careful research and practice. And in this paper, the intelligent question-answering system based on the new sentence similarity computing model, comprehensively sums up all kinds of methods that current intelligent answering-questions system query the sentence similarity, analyzes their characteristics, and designs a kind of workflow to make full use of all kinds of intelligent question-answering resources. And aiming to the characteristics of intelligence and the openness of the system, this paper analyzes the key technologies to realize the intelligent question-answering system. From the perspective of the use effect of this intelligent question-answering system, it can fully meet the needs of the remote education system for intelligent answering questions. In some sense, intelligent question-answering system can also play the role of a good teaching system.

REFERENCES

[1] Chaudhri V., Question Answering System:Papers from the 1999 Fall Symposium, *Technical Report FS-98-04(November),AAAI*, **1999**.

[2] Chinese Knowledge Information Processing Group, Categorical Analysis of Chinese, *ACLCLP Technical Report # 93-05*, Academia Sinica, **2009**.

[3] K. Chidananda Gowda; E. Diday, *IEEE Transactions on Systems,Man and Cybernetic*,**2012**,22(2), 195-199.

[4] YimingYang,An evaluation of statistical approach to text classification, Computer Science Department,Carnegie Mellon University, **1997**.

[5] G.Cao;J.Y.Nie;J.Gao;S.Robertson, Selecting Good Expansion Terms for Pseudo-Relevance Feedback,*In Proceedings of the 3lst annual international ACM SIGIR conference on Research and development in information retrieval(SIGIR08)*,**2012**, 243-250.
[6] Yaakov Engel;ShieMannor;Ron Meir, *IEEE Transactions on Signal Processing*, **2004**, 1(2), 145-152.
[7] Bin Li; Ting Liu. *Computer application research*,**2003** (12), 67-71.
[8] Faguo Zhou;Bingru Yang, *Computer engineering and application*, **2008**,12(1),56-61.
[9] RueyShiang Shaw; ChinFeng Tsao;Peiwen Wu, *Expert Systems With Applications*,**2012**, 39 (14), 11593-11606.
[10] Thomas P Y, *IEEE Trans. on Visualization & Computer Graphics*, **2006**, 12(4), 640-648.
[11] A.K.McCallum; K.Nigam,A comparison of event models for naive Bayes text classification,*In Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, **2009**, 137-142.
[12] HansPeter Kriegel;Karsten M, *Data Mining and Knowledge Discovery*, **2013**,15(l),87-97.
[13] KehYih Su, *The 23(rd) International Conference on Computational Linguistics Proceedings of the Main Conference* **2010**,2(3),195-201.
[14] GuangjunGuo; Fei Yu;Zhigang Chen; Dong Xie, *Journal of Computers*,**2011**, 6(2),377-386.
[15] Dongqing Wu;Fengjian Yang; Chaolong Zhang, *Journal of Software*, **2013**, 8(1),192-199.
[16] Peck C., *Journal of pharmacokinetics and biopharmaceutics*, **2010**, 12(5), 134-142.