International Journal of Database Theory and Application
Vol.9, No.7 (2016), pp.169-178
http://dx.doi.org/10.14257/ijdta.2016.9.7.15

A Case Based Reasoning Algorithm for Enterprises' Integration of Informatization and Industrialization

Lifeng Li, Shifeng Liu and Danqing Li

Beijing Jiaotong University
School of Economics and Management
Beijing, China
13113153@bjtu.edu.cn

Abstract

The integration of informatization and industrialization is depth integration of informatization and industrialization in many fields, which is the new developing way for informatization and industrialization. It is a complicated process of integration for enterprises, which is necessary to learn from the experience of other enterprises to improve efficiency. But facing numerous cases, it is hard for enterprises to pick up the right cases, because of shorting of effective recommended algorithm in academic field and practical application. Base on the studying the existing recommend algorithm, this paper try to design new algorithm with analyzing the structure and characters of the cases. First of all, the case structure is analyzed and are classified and the information of cases are classified and graded according to the degree of integrity; secondly, AHP is used to give weights to different attributes; then, the recommended algorithm is illustrated; finally, the cases similarity are calculated and the high similarity cases are outputted to the users.

***Keywords**: The integration of informatization and industralization; recommened algorithm; case based reasoning; similarity calculation*

1. Introduction

In China, the integration of informatization and industrialization is the Chinese characteristic way to develop industry and information technology, which is mutual promotion and win-win process. Similar to early information construction, the process of the integration is systematic and complicated. It easily lead to the failure, if any deviation happens. So that it will be useful to draw on the successful experience of the integration by enterprises for the construction. But there are great differences between enterprise in different industries, whether in the management process or in the production process. It is hard to pick up the right cases from numerous cases in database. Additionally, a single case or the cases from different industries can't meet the enterprises' needs. Hence, it is urgent to design the case reasoning algorithm of enterprise' integration of informatization and industrialization. According to their self-characteristics and needs, enterprises offer the key words and the algorithm will calculate similarity of cases in the database. The highest similarity of cases will be recommended to the enterprise. Because of unstructured data and numerous cases, the accuracy and speed of the case similarity calculation is the key to the algorithm. So the case retrieval algorithm has high research value.

This paper is organized as follows. In the following section, literature review is given to show the current research in this filed and pointed out what theories are used in this paper. Subsequently, the characteristics and the structure of the cases are analyzed and the cases are classified according to the degree of integrity. The process of case reasoning and the design of case reasoning algorithm are given in the future section. Finally, a data set is used to test the algorithm and the result show that it is effective and efficiency.

ISSN: 2005-4270 IJDTA

2. Literature Review

Case-based Reasoning method is one of the important solution in the field of artificial intelligence and a way to solve the problem by imitating human thought. It can take the solution for the old problem as the way to solve new problem. Because of no need to store prior knowledge, it can eliminate the difficult problem that the knowledge is difficult to obtain in the general knowledge system. CBR originated from cognitive science research by Roger Schank in AI Laboratory of Yale University in the 1980s. In the 1982s, Schank proposed the dynamic memory theory with memory organization packet as the core, which considered the earliest thought of CBR in the artificial intelligence field [1]. The first case based reasoning system is CYRUS, developed by Janet Kolodner (1983), which is a basic question answering system with the all the travel and meeting information for former U.S. Secretary Cyrus Vance. The case store model for the system became the basis for later case-based reasoning system, including MEDIATOR (Simpson, 1985), PERSUADER (Sycara, 1988), CHEF (Hammond, 1989), JULIA (Hinrichs, 1992) and CASEY (Koton, 1988) [2].

In aspect of case representation, Doyle [3], at the University of Berlin, proposed a case-based reasoning language based on XML. Wang Yue [4] proposed a representation method based on Simulation and structure of case through analyzing the description of the case structure and the problems they faced in case reasoning. Kai Bo Zhou [5] combined XML and object oriented technology to puts forward a kind of object - oriented case representation method.

Case retrieval algorithm is the core part of case reasoning. There are three kinds of general methods of case retrieval: the nearest neighbor method, the inductive index and the knowledge guidance. With the expansion of the application of case reasoning, many scholars in specific areas, put forward some new methods, such as case clustering analysis, neural network based case retrieval model, rough set theory combined with the fast case retrieval model and so on. In order to improve the efficiency of retrieval, Ma [6] used clustering analysis to classify the cases before case retrieval. Meng Yanni, [7] studied the model based on neural network. Haibo Zhou [8] compared the different data types of case attributes and proposed the similarity calculation method according to several different types of interval data.

In aspect of setting case attribute weights, Zhansi Jiang [9] uses the similarity deviation method to give attributes weight by calculating the minimum values of the sum of similarity deviation square. Xiaoyan Wei [10] uses the analytic hierarchy process (AHP) to study the weight setting problem of the multi-dimensional optimization algorithm. Gareth [11] study the application of genetic algorithm in the problem of setting the case attribute weight. Gu and Li propose a new case retrieval method called FRAWO, in which emphasis is put on the problems of similarity calculation of fuzzy and interval attributes of cases using trapezia based fuzzy set and the dynamic weight of a case is adjusted by adopting PULL&PUSH strategy [12].

The basic theory of case based reasoning algorithm has few improvement and the innovation focus on the new application in different fields. In the practical application aspect, the research of the application of case based reasoning in decision support system is given by Albert [13]. Zhang [14] has done research on the application of case-based reasoning on the disaster relief auxiliary decision support system. Some domestic scholars research the theoretical framework of emergency decision making based on case based reasoning support system. Yanchao Feng, who integrates the CBR into the budget management, manages the new project budget according to the information of matching case [15]. Case based reasoning is also integrated with complex medical diagnosis (Subhagata, Suvendu 2013) [16] and question answering (Weis 2015) [17].

3. Case Characteristics Analysis

3.1. Case Content

There are four main parts in a case, including enterprise profile, the construction situation, the construction program, achievements and experience. Do not use abbreviations in the title or heads unless they are unavoidable.

Table 1. Case Content

Name	Content
Enterprise profile	Introduction of the enterprise Enterprise scale
The construction situation	The timetable of the construction Investment scale Infrastructure
The construction program	R&D investment The source of information equipment Manning level Information system construction
Achievement and experience	Construction achievement Application Competitiveness Economic and social performance Construction experience

3.2. Case Characteristic

The structure of the case includes three parts: the first part is the basic information of the case, which is shown to the readers after the searching, but it is set as the identification parameters rather as the inference parameters; the second part is the basic parameters, which is divided to the basic information of the enterprise and the integration. These information is used as the inference parameters; the third part is the program of the integration, including the integration program and the information of supporting enterprises. This part is what users want to learn for integration construction. Also they are parts of important parameters for searching.

Table 2. Case Characteristic

Name	Content
The basic information of the cases e	Case name Case ID Case abstract Key words Case grading
Case parameters	Industry involved Enterprise scale Geographical location Beginning and ending time Investment
The integration program	Business for information The source of hardware The source of software

Name	Content
	The kinds of software The number of information related staffs
Supporting enterprises	Consulting enterprise name Implementing enterprise name Hardware equipment supplier Software equipment supplier

3.3. Case Classification

The case level can be classified to ten levels according to the case completeness. In order to calculate the similarity conveniently, the levels are assigned to 1, 0.95, 0.9, 0.85, 0.75, 0.7, 0.65, 0.6 and 0.45.

4. The Reasoning Algorithm

4.1. Case Retrieval Process

In this study, the searching strategy uses a combination of inductive method and the nearest neighbor algorithm, which is the nearest neighbor method with index. In searching, the process of case searching is divided to two parts. When there are a large number of cases, the inductive retrieval method is used according to the directory index until the number of cases is less than threshold vale. This step is equivalent to a classification of the decision-making problems and make out that which category of the cases belong to. In the searching results, the similarity is calculated by using the nearest neighbor method. The cases are sorted by the similarity and the highest ones are recommended to the users.

According to the user needs, the case searching is divided to two parts: basic search and advanced search. Based on the four attributes of Industry category, enterprise scale, the ending time of integration and the case classification, the similarity is calculated in basic search. Advanced search needs eight more attributes, includes aggregate investment, information business, program origin, hardware and software equipment origin, the kinds of information systems, the number of staffs, the suppliers and related service enterprises.

The process of case-based reasoning, as show in Figure 1:

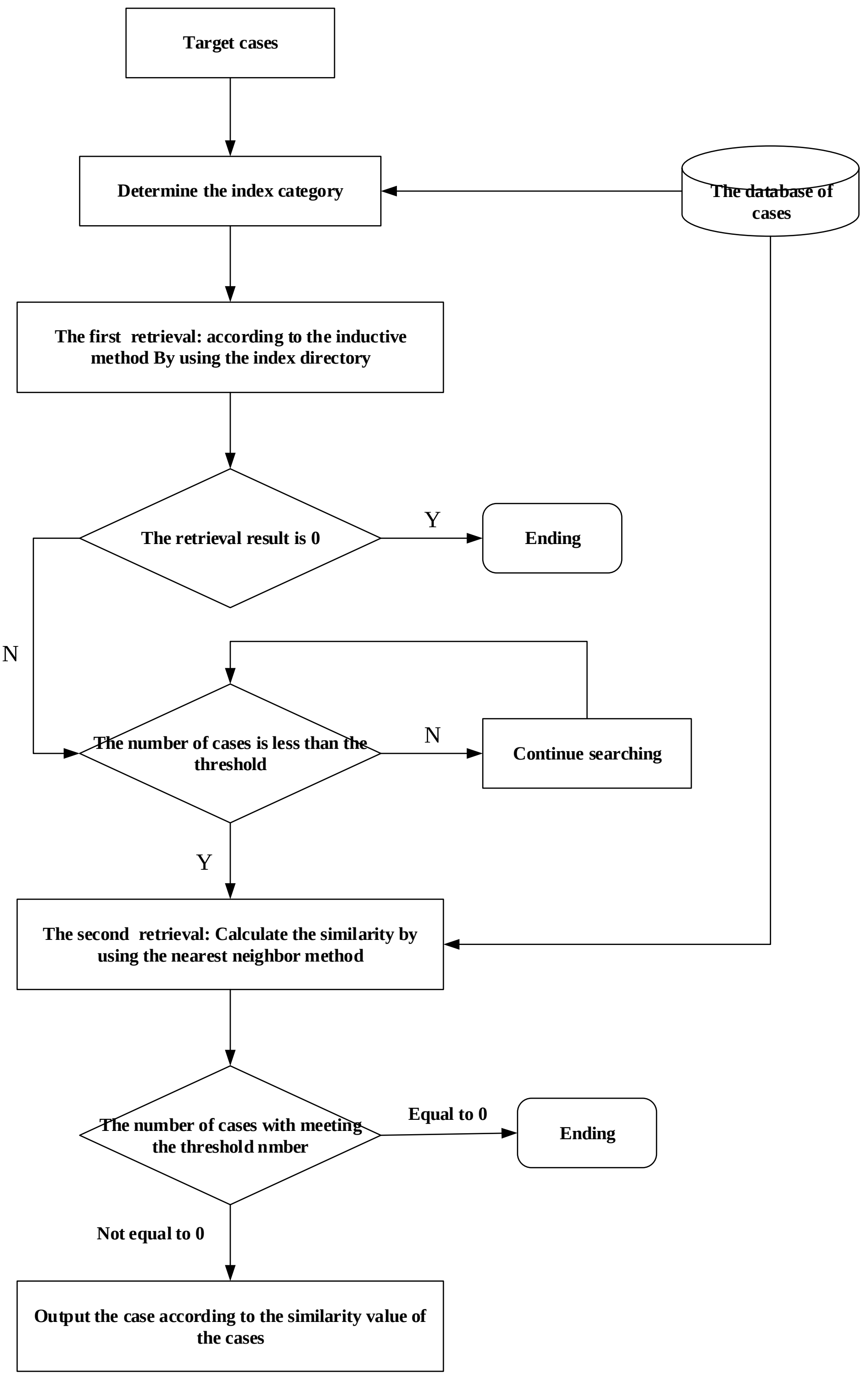
Target cases
Determine the index category
The database of cases
The first retrieval: according to the inductive method By using the index directory
The retrieval result is 0
Y
Ending
N
The number of cases is less than the threshold
N
Continue searching
Y
The second retrieval: Calculate the similarity by using the nearest neighbor method
The number of cases with meeting the threshold nmber
Equal to 0
Ending
Not equal to 0
Output the case according to the similarity value of the cases

Figure 1. Process of Case Reasoning

4.2. Weight Setting

Because that each case attributes have different effect on the description of the case, so that different weights are arranged to the attributes. The key attribute has a great influence on the retrieval result, and the secondary attribute has less influence on the retrieval result.

The retrieval result objectively reflects the similarity degree between the target problem and the existing cases in the database, and the accuracy of the retrieval is improved.

The expert evaluation is taken for the weight setting, then the expert opinion is analyzed by the AHP (analytic hierarchy process method).

Table 3. Weight Setting

Definition	Introduction	Rating
Equality important	Two factors are equally important	1
Slightly important	This factor is a little more important than the other one	3
Obviously important	This factor is more important than the other factors	5
Very important	This factor is significantly more important than the other factors	7
The most important	This factor is the most important than the other one	9
2, 4, 6, 8 are the intermediate values between the adjacent judgment		

The steps of using AHP to assign weight, as follows:

Establishing judgment matrix: It assumes that there are n attributes and the judgment matrix A is composited by mutual contrast between every two attributes. It indicates the importance of the first I attribute relative to the j attribute

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \tag{1}$$

Calculating the product of each line of the matrix element and it is denoted as M_i

$$M_i = \prod_{j=1}^{n} a_{ij} \quad \left(i, j = 1, 2, \ ...\text{n}\right) \tag{2}$$

Calculating's M_i n root mean square

$$\overline{W_i} = \sqrt[n]{M_i} \tag{3}$$

Normalizing and getting each attributes' weight, denoted as W_i

$$\sum_{i=1}^{n} W_i = 1 \tag{4}$$

Author names and affiliations are to be centered beneath the title and printed in Times New Roman 12-point, non-boldface type. Multiple authors may be shown in a two or three-column format, with their affiliations below their respective names. Affiliations are centered below each author name, italicized, not bold. Include e-mail addresses if possible. Follow the author information by two blank lines before main text.

4.3. Case Attribute Classification

Before discussing the similarity of each attribute, we must firstly classify the attributes. Different types of attributes have different definitions of similarity. There are three types of attributes:

(1)Numerical continuous attributes: Attribute values are represented by continuous numerical value, for example: the finishing time of the fusion case, the number of total investment and information related personnel are all belong to this type, which are known as numeric attributes.

One method of normalization is to standardize the numerical value and to make the characteristic value of the mean value equal to the variance. The mean value of the attribute X is m and the standard deviation is S.

$$m = \sum_{i=1}^{n} x_i / n \tag{5}$$

$$s = \sqrt{\sum_{i=1}^{n} \left(x_i - X\right)^2 / n} \tag{6}$$

The standard procedure for attribute X can be described as:

$$X^* = \frac{X - m}{s} \tag{7}$$

(2)Enumerative attributes: The characteristic value is not only the grade order relation, but also has the attribute of the quantity relation. For example, the geographical location of the fusion case, the source of hardware and software.

(3)Ordinal attributes: The desirable feature value has a sequence of relation, but there is no quantitative relation. It is called hierarchical attribute, such as a case rating, which belong to this type. To facilitate the calculation of similarity, this paper has assigned the 10 grades of the case rating, and transformed into numerical attributes, so the similarity calculation is the calculation method of numerical attributes.

4.4. Similarity Calculation between the Attributes

(1) Numerical continuous attributes

Since that all numerical attribute values are positive in the case, so after the standardization of the value, it is possible to apply minimum value method to calculate the similarity between numerical attributes, namely:

$$sim(x_i, y_i) = \begin{cases} 0 & x_i or y_i = null \\ \dfrac{\min(x_i, y_i)}{\max(x_i, y_i)} & x_i \neq y_i \\ 1 & x_i = y_i \end{cases} \tag{8}$$

(2) Similarity calculation of enumerative attributes:

For the similarity calculation of the enumeration type attribute, the method of direct matching is used to determine the similarity between the two attribute values according to the matching of the string. In addition, due to the user in the formulation of the fusion schemes, one may not make just one decision, such as the kinds of information system in a program is impossible only OA or ERP. Enterprises may choose kinds of systems like OA, ERP, OM in a program at the same time.

$$sim(x_i, y_i) = \begin{cases} 0 & if\ \forall x_i \neq y_i OR y_i = null \\ \dfrac{count(x_i = y_i)}{\max(count(x_i), count(y_i))} & if\ \exists x_i = y_i OR x_i = all \end{cases} \tag{9}$$

x_i = all means that the user fill in "all", y_i =null means that the characteristic value of the attributes are null in the source cases.

In particular, similarity calculation of industry classification attributes needs classified: If the user submits the industry classification for the first class industry, then use the formula (9) to calculate the similarity; otherwise, if the user submitted the industry classification for the two industry, then use the formula (10) to calculate the similarity. Specific formulas are as follows:

$$sim(x_i, y_i) = \begin{cases} 0 \\ 0.6 \\ 1 \end{cases} \tag{10}$$

Sim(x_i , y_i) = 0, if primary industry matching; Sim(x_i , y_i) = 0.6, if primary industry matching, the second industry does not match; Sim(x_i , y_i) = 1, if the second industry.

(3)Similarity calculation of Ordinal attributes

For hierarchical attribute in the case, in order to facilitate the calculation similarity between the scale of the enterprise, as the size of large, medium, small and micro assign respectively 4, 3, 2, 1, and the similarity calculation method is based on the distance:

$$sim(x_i, y_i) = \begin{cases} 1 - \dfrac{|x_i - y_i|}{\max(R_x) - \min(R_x)} & x_i \neq y_i \\ 1 & x_i = y_i \end{cases} \tag{11}$$

x_i , y_i are the same type of numerical value from two cases. R_x is the value collection of attribute x in all cases. Max (R_x) is the biggest value min (R_x) is the smallest value.

4.5 Case Similarity Calculation

Target case X is consist of 4 attributes. X= （x_1，x_2，x_3，x_4），represents the one attributes of the target case. Similarity, Y= （y_1，y_2，y_3，y_4），represents the one attributes of the target case. According to the nearest neighbor algorithm, the similarity function between the target case and the source case is:

$$sim(X,Y) = \sum_{i=1}^{4} W_i \times sim(x_i, y_i) \tag{12}$$

In the formula, the weight value of every attributes are equal to W_i.

Target case X is consist of 15 attributes, X= （x_1，x_2，……，x_{15}），x_i is one attribute of the target case, Similarity, Y= （y_1，y_2，……，y_{15}），y_i represents the one attributes of the target case. According to the nearest neighbor algorithm, the similarity function between the target case and the source case is:

$$sim(X,Y) = \sum_{i=1}^{15} W_i \times sim(x_i, y_i) \tag{13}$$

In the formula, the weight value of every attributes are equal to W_i.

5. Empirical Study

The threshold is set to number 50, 100 and 150 and the number of cases is set to number1000, 2000 and 3000. Through five times searching, five cases with the highest similarity are recommended and the searching time and average searching time are recorded.

5.1. Basic Retrieval

Table 4. Results of Basic Retrieval

CN	T	Case Retrieval Time (Second)					AVG
		C1	C2	C3	C4	C5	
1000	50	0.010	0.020	0.008	0.037	0.042	0.02
	100	0.067	0.083	0.054	0.037	0.042	0.05
	150	0.066	0.082	0.055	0.037	0.052	0.05
2000	50	0.052	0.099	0.022	0.044	0.018	0.04
	100	0.154	0.087	0.117	0.146	0.132	0.12
	150	0.144	0.098	0.205	0.150	0.160	0.15
3000	50	0.092	0.044	0.065	0.067	0.031	0.06
	100	0.114	0.248	0.066	0.105	0.140	0.13
	150	0.102	0.175	0.066	0.249	0.212	0.16

5.2. Advanced Retrieval

Table 5. Results of Advanced Retrieval

CN	T	Case Retrieval Time (Second)					AVG
		C1	C2	C3	C4	C5	
1000	50	0.027	0.041	0.014	0.060	0.066	0.04
	100	0.105	0.133	0.101	0.060	0.072	0.09
	150	0.116	0.139	0.098	0.061	0.078	0.09
2000	50	0.067	0.152	0.044	0.081	0.029	0.07
	100	0.072	0.169	0.162	0.256	0.183	0.16
	150	0.209	0.212	0.337	0.181	0.202	0.22
3000	50	0.100	0.103	0.105	0.120	0.049	0.09
	100	0.200	0.329	0.115	0.228	0.006	0.18
	150	0.200	0.319	0.113	0.231	0.339	0.24

6. Conclusion

In this paper, based on the user's needs, case based reasoning is divided into basic and advanced searching. The searching process is divided into two retrieval process for case based reasoning: The first retrieval search all cases in the case library by filtering the input conditions and the second search calculate the similarity of the target case between the cases based on the results of the first search and output the cases with more high value than the threshold. Meanwhile, the searching process, procedure and threshold are illustrated. Firstly, in order to complete the similarity calculation of the second retrieval, AHP method is used to assign the weight of the attributes; Secondly, for the two forms of numerical attributes, standard normal distribution formula is improved and the method of similarity calculation is discussed; at last, empirical study is used to test the algorithm.

Acknowledgements

This work was supported by a grant from the Program of Beijing Social Science Foundation Project (No. 15JGB101).

References

[1] C. S. Roger, Dynamic memory, "A theory of reminding and learning in computers and people", Cambridge University Press, **(1983)**.

[2] L. Alireza, "A decision support system for solving quality problems using case-based reasoning", Total Quality Management and Business Excellence, vol. 14, no. 6, **(2003)**.

[3] C. Lorcan, D. Doyle, and P. Cunningham, "Representing Similarity for CBR in XML", Advances in Case-Based Reasoning, 7th European Conference, ECCBR 2004, Madrid, Spain, Proceedings, **(2004)**.

[4] Y. Wang, J. Fan and S. G. Tian, "Case representation method of case based reasoning expert system", Journal of Shanghai University of engineering and technology, vol. 19, no. 1, **(2005)**.

[5] K. Zhou, "An object-oriented case representation method based on XML", Journal of Wuhan University of Technology: information and Management Engineering Edition, vol. 27, no. 3, **(2005)**.

[6] M. Shixia, J. Li, and D. Liu, "The Case Retrieval Strategy Based on Hierarchical Clustering", Web Mining and Web-based Application, WMWA '09. Second Pacific-Asia Conference on IEEE, **(2009)**.

[7] Y. Meng and Z. Fang, "A case retrieval method based on ART-2 neural network", Journal of Information, **(2006)**.

[8] K. Zhou, S. Feng and F. Li, "Study on the similarity computation model based on the characteristic of case attributes", Journal of Wuhan University of Technology: information and Management Engineering Edition, vol. 25, no. 1, **(2003)**.

[9] Z. Jiang, L. Chen and N. Luo, "The nearest neighbor case retrieval and similarity analysis", Computer integrated manufacturing systems, vol. 13, no. 6, **(2007)**.

[10] X. Wei, "Research on multi-dimensional optimization retrieval algorithm based on case reasoning", Journal of Taiyuan University of Technology, Shanxi, **(2008)**.

[11] R. B. Gareth and S. Petrovic, "Selecting and weighting features using a genetic algorithm in a case-based reasoning approach to personnel rostering", European Journal of Operational Research, vol. 175, no. 2, **(2006)**.

[12] G. Dongxiao, "Case retrieval and weight optimization method research and application", Journal of systems engineering, vol. 24, no. 6, **(2009)**.

[13] A. A. Albert and S. Dutta, "Case-based decision support", Comm. ACM, vol. 41, no. 5, **(1998)**.

[14] M. L. A. Zhang, D. L. Zhou, and J. F. N. Jr., "A Knowledge Management Framework for the Support of Decision Making in Humanitarian Assistance/Disaster Relief", Knowledge & Information Systems, vol. 4, no. 3, **(2002)**.

[15] F. Yanchao, "Research on the R&D budget management based on case based reasoning", Journal of science progress and Countermeasures, vol. 27, no. 23, **(2010)**.

[16] C. Subhagata, "A Case-Based Reasoning system for complex medical diagnosis", Expert Systems, vol. 30, no. 1, **(2013)**.

[17] W. K. Heinz, "A case based reasoning approach for answer re-ranking in question answering", arXiv preprint arXiv, **(2015)**.