ATLANTIS PRESS

Advances in Computer Science Research, volume 93

2nd International Conference on Mathematics, Modeling and Simulation Technologies and Applications (MMSTA 2019)

J Algorithm for Scientific Knowledge Discovery: Taking Economic Growth Theory as an Example

Qianhui Zhao[1], Qi Qian[2] and Zhaohua Jiang[3]

[1]Dongbei University of Finance and Economics, Dalian, China

[2]Johns Hopkins University, Baltimore, USA

[3]Institute of Science and Technology Management, Dalian University of Technology, Dalian, 116023, China

***Abstract*—Text mining, intelligent algorithms and knowledge maps are the research frontiers and hotspots of current scientific knowledge discovery. However, there is currently no theoretical basis for such research—for example, the theoretical framework for the structure of scientific knowledge systems and the lack of case studies for scientific knowledge discovery. The paper proposes that most scientific knowledge systems are composed of concept set [A], concept set [B], and concept set [C]. Different scientific knowledge systems have different cognitive patterns [P], so the scientific knowledge system consists essentially of four concept sets. And the four concept sets are represented in the four quadrants of the "concept coordinate system", and then study the evolution process from concept to model block to model to model system, revealing new concepts and new structures in the process of scientific knowledge discovery. And through the construction case of the economic growth model of The Synergy Theory, it is shown that the knowledge map provides a powerful analysis tool for the "four-set analysis method" proposed by the J system methodology, thus greatly expanding, deepening and innovating the Swanson's knowledge discovery method which is based on non-relevant literature–ABC model.**

Keywords—scientific knowledge discovery; J algorithm; ABC model

Swanson first discovered the existence of recessive associations in the medical literature[1]. He inferred the recessive associations between [A] and [C] through the complementarity in content of two non-relevant literatures containing concept set [A] and concept set [C]. (Dong Fenghua, Lan Xiaoyu, 2004). For example, [A] indicates that the intake of a substance may cause a physiological change [B], and the physiological change of [B] triggers a disease of an organ [C], so that useful information [A] can be obtained in [C], and this is not found in a single literature, and through the intermediate link [B], this recessive associations can be extracted. Thus, Swanson refers to the knowledge discovery method based on non-relevant literature as the ABC model. The J system theory studied in this paper is the theoretical basis of the ABC model, and the J system method and J algorithm based on the J system theory are further expansion, deepening and innovation of the ABC model.

I. INTRODUCTION

How to find and construct concept sets of scientific knowledge system? For the two sets of non-relevant literature A and C, the associations between the two can be established through an intermediate word or intermediate literature so as to find B. The process of discovering the associations between A and C is called the non-relevant literature knowledge discovery method. The algorithm consists of a two-step knowledge discovery process combining open and closed, and form namely, hypothesis formation and hypothesis testing[2] .

Open mode: the open knowledge discovery process can be expressed as looking for C (A→B→C) from A. The algorithm flow is as follows:

Search the document containing the A word in the Medline database and save the title to file a. Cut out all the words that co-occur with the A word to form a co-occurrence list B;

Count the word frequency of B in the Medline database and file a, and calculate the words with large relative frequency appearing throughout Medline [3];

Remove words that are obviously inappropriate. Search the Medline title with the remaining words, and find the target word and filter to get the word list u. [4];

Use u to retrieve the Medline database title and save the title to file c. Screening C words artificially. Finally, a list of C-B word pairs is formed;

Search the Medline database title with C and repeat the second step to get a list of candidate C words. Form an AB-BC co-occurrence connection.

Closed mode: the process can be expressed as A→B←C. It can be summarized as searching the literature containing the A and C word titles from the Medline database and extracting the words in the title by computer to form a B word list, and then extracting the common word B from the B word list to find the corresponding literature title [5].

The following is a discussion of various algorithms and methods based on computer knowledge discovery of non-relevant literature:

A. Gordon and Lindsay Based on Word Frequency Statistics

This method inherits and extends Swanson's research, and its analysis object becomes the complete field of the Medline record. The steps are as follows (Lindsay RK.Gordon MD,1999) :

B. Weeber's Two-step Knowledge Discovery Process and DAD System

Weeber used computer processing techniques and proposed a lexical analysis system based on compound concept[6]. On a two-step basis, Weeber developed the DAD system, which deals with the subject of the specification UMLS (Computerized Information Retrieval Language Integration System) in Medline records, and successfully reproduced the discovery process of Swanson's "Renault's disease" and "fish oil".

C. Co-occurrence Analysis of Johannes Stegmann

Co-occurrence analysis is a powerful tool for knowledge discovery[7],which is mainly applied to keywords in complementary literature, and clusters keywords according to centripetality and density. Centripetality is used to measure the degree of interaction between disciplines. Density is used to measure the degree of aggregation of words, It is the internal strength of a class.

D. Pratt and LitLinker of Yetisgen-Yildiz

They use ARM association rule mining technology, which is an unsupervised learning. ARM differs from co-occurrence in that it can co-occur in three or four words. The algorithm of Pratt and Yetisgen-Yildiz uses the Medline knowledge base and UMLS [8].

E. Padmini Srinivason's Text Mining Algorithm

Padmini Srinivason's algorithm is used to discover hypotheses that rely entirely on the MeSH vocabulary and is also based on the Medline database in the medical field. He created a Profile (topic-related subject heading) for each topic, and the user only has to give the initial interest topic and the subject type of the profile to be generated[9].

F. Wei Huang et al.'s New Connection Prediction Algorithm

This algorithm can find the recessive associations between A and C. The novelty is that the target word C and the intermediate word B belong to the same genus relationship in the hierarchical relationship, And there is no association between the starting word A and the target word C. Wei Huang and others' test sets are also Medline databases, and the algorithm only processes MeSH words in Medline records[10].

G. The Associated Concept Space of Van der Eijk et al.

The scientific literature is often fragmented, which means that certain scientific questions can only be answered by combining information from various articles. Therefore, a learning algorithm that uses co-occurrence data as input can be used to map concepts into multidimensional space. The resulting conceptual space allows exploring the neighborhood of a concept and discovering the relationships between potential, new concepts.

Van der Eijk discovered that "insulin" and "ferritin" are tightly positioned together and that these concepts did not appear together in a previous series of abstracts. So infer that they are relevant, and search PubMed for 212 articles containing two words. ACS can reveal the connotative information [11].

The goal of this approach is to achieve visualization between concepts and concepts, but it is quite difficult.

H. Stochastic Model of Wren et al

The research field of Wren et al. is also medicine, and the database is still Medline. They extract concepts from the title and abstract of the literature record and it is done from the free text of multiple databases to the mapping of concepts. The acronym problem was solved by the acronym universal detection method invented by Wren [12].

I. BITOLA by Hristovski et al.

BITOLA is a literature-based interactive biomedical knowledge discovery system designed by Hristovski et al. They used the MeSh thesaurus and the title and abstract in the Medline record. Hristovski and others use association rules to mimic Swanson's A→B→C open knowledge development model[13].

J. Xiao hua Hu et al. Bio-sbKDS

Bio-sbKDS is also based on the Medline database. The specific implementation is to download the literature data from PubMed. The processing fields include MeSH words, title, and abstract [14] .

The above 10 algorithms have greatly developed the ABC model discovered by Swanson, and there are also some practical applications, and some complete software that can be used; but most of them involved concept set [A], concept set [B], concept set [C] is a simpler association rule mining, and is less integrated with current scientometric methods or patent metrology methods, knowledge mapping methods, text mining methods, and so on. Therefore, based on the J system theory and the J system method, this paper combines the subject-based research of scientific metrology method, knowledge mapping method and text mining method to construct the J algorithm of knowledge discovery. Almost in sync with Swanson, Jiang Zhaohua's preliminary paper on the scientific knowledge system consisting of four concept sets, "On the Three Collections", was recommended by Professor Sun Mutian of Harbin Normal University to the magazine "Sink Science" published in 1985[15], Jiang Zhaohua et al. published a paper in Journal Seeking Truth, expounding the relatively complete and clear theoretical framework of "the scientific knowledge system consists of four concept sets", and proceeded from the materialist dialectics; the system science research in the Chinese Academy of Sciences With the help of Professor Shu Guangfu, Jiang Zhaohua (2000) published "J-System Reconstructability: A formalized Study" in "Int. J. General Systems" in 2000 [16] , and systematically proposed "J System Theory" and "J Methodology". In 2018[17], Liu Jianhua and Jiang Zhaohua (2018) jointly published by the China Science Press and the German Springer Publishing Group, "The Synergy Theory on Economic Growth: Comparative Study between China and Developed Countries", further expanded and deepened this study.)

II. CONCEPT SET AND MODULE OF SCIENTIFIC KNOWLEDGE SYSTEM

A. Concept Set in Scientific Knowledge System

Through the analysis of a large number of scientific knowledge systems, we can see that most scientific knowledge systems are constructed (composed, constituted, or compounded) by three "concept sets", and the three "concept sets" are generally written as the concept set [A], the concept set [B], the concept set [C].First, we analyze several typical scientific knowledge systems:

Newton system is the simplest system in the physical system, it is a particle, it has quality $m, m \geq 0$; positioning r in a certain frame of reference, r constantly changing in time t, there is speed $(v = dr / dt)$, and perhaps acceleration $a = d^2 r / dt^2$.

Newtonian system = {quality variable, dynamic variable, motion variable}

The reason why the particle is accelerating, according to Newton, is because the force is acting on it. Thus, (r, v, a) is the motion parameter of the particle system, while f is the powertrain parameter and m is the internal parameter. We see that the simple particle is composed of three "parts" into $\{m, f, (r, v, a)\}$.

Thus, according to Newton's cognitive model p_N, Newtonian mechanics knowledge system, concept set [A], including quality m concept; concept set [B], including force f concept; concept set [C], including position concept r, velocity concept $(v = dr / dt)$ and acceleration Concept ($a = d^2 r / dt^2$).

The so-called integral knowledge system. According to the cognitive models of mathematicians, the three concept sets are:

①$[A]:\rightarrow f(x)$;②$[B]:\rightarrow \int_{\Sigma} dx$; ③$[C]:\rightarrow F$.

The real class is a system R, which is constructed from a set of positive real numbers R_+, zero set R_0, and a set of negative real numbers R_-, $r = (R_+, R_0, R_-)$

The so-called control system, according to Wiener's P_W, is such a system in which the following three concept sets are given:

①[A]: input set; ②[B]: feedback set; ③[C]: Output set.

So why are the three concept sets of Newtonian systems (m, f, a), and the three conceptual sets of control systems (input sets; feedback sets; output sets)? Why are the three concept sets of different types of systems different? Obviously there is something basic working in it and we call this kind of thing." cognitive model", writing [P].

This shows us that the scientific knowledge system J is constructed from three conceptual sets according to the cognitive model [P]. Different scientific knowledge systems have different sets of concepts because of their different cognitive patterns [P].

Thus, the so-called knowledge system J is constructed from three concept sets $[A],[B],[C]$ according to the cognitive model [P], namely:

$J = [A] \circ [B] \circ [C]$ or $J = ([A],[B],[C])$ or $J = [P]([A],[B],[C])$ .here，$[A],[B],[C]$ are the formal symbols of the three concept sets, and " o " is the interconnection and interaction between the three concept sets. Cognitive mode [P] is also a set of concepts, so the knowledge system consists essentially of four sets of concepts [P],$[A],[B],[C]$.

B. Conceptual Coordinate System

The J system approach first represents the four sets of concepts in the four quadrants of the "concept coordinate system".

For example, the “quantity” and “price” in the market structure problem are the two coordinate axes of the conceptual coordinate system; the “entity elements” and “relational elements” in the economic growth problem are the two coordinate axes of the conceptual coordinate system. See Figure 1 for details.

C. Module in the Scientific Model

In Cobb-Dagras production function model $Y = aK^{\alpha}L^{\beta}$, a, K^{α}, L^{β} are three modules; In the general equation $\frac{d\theta_i}{dt} = T_i + P_i \quad (i = 1, 2, \cdots, n)$ of Beta Langfi's open system, a $\frac{d\theta_i}{dt}$, T_i and P_i are three modules.

The model is constructed by modules A, B, C, P.

A knowledge system is constructed from three concept sets $[A],[B],[C]$ based on cognitive patterns [P]. The model is constructed from modules A, B, C, P .The modules A, B, C, P here are generally determined by$[A],[B],[C]$ [P], and there is no simple correspondence between them. However, in general, A is more determined by $[A]$ - generated on the basis of $[A]$; likewise, B is more determined by $[B]$ - generated on the basis of $[B]$; likewise, C is more The $[C]$ -determined - generated on the basis of $[C]$; the same, P is more $[P]$ decided - generated on the basis of $[P]$.

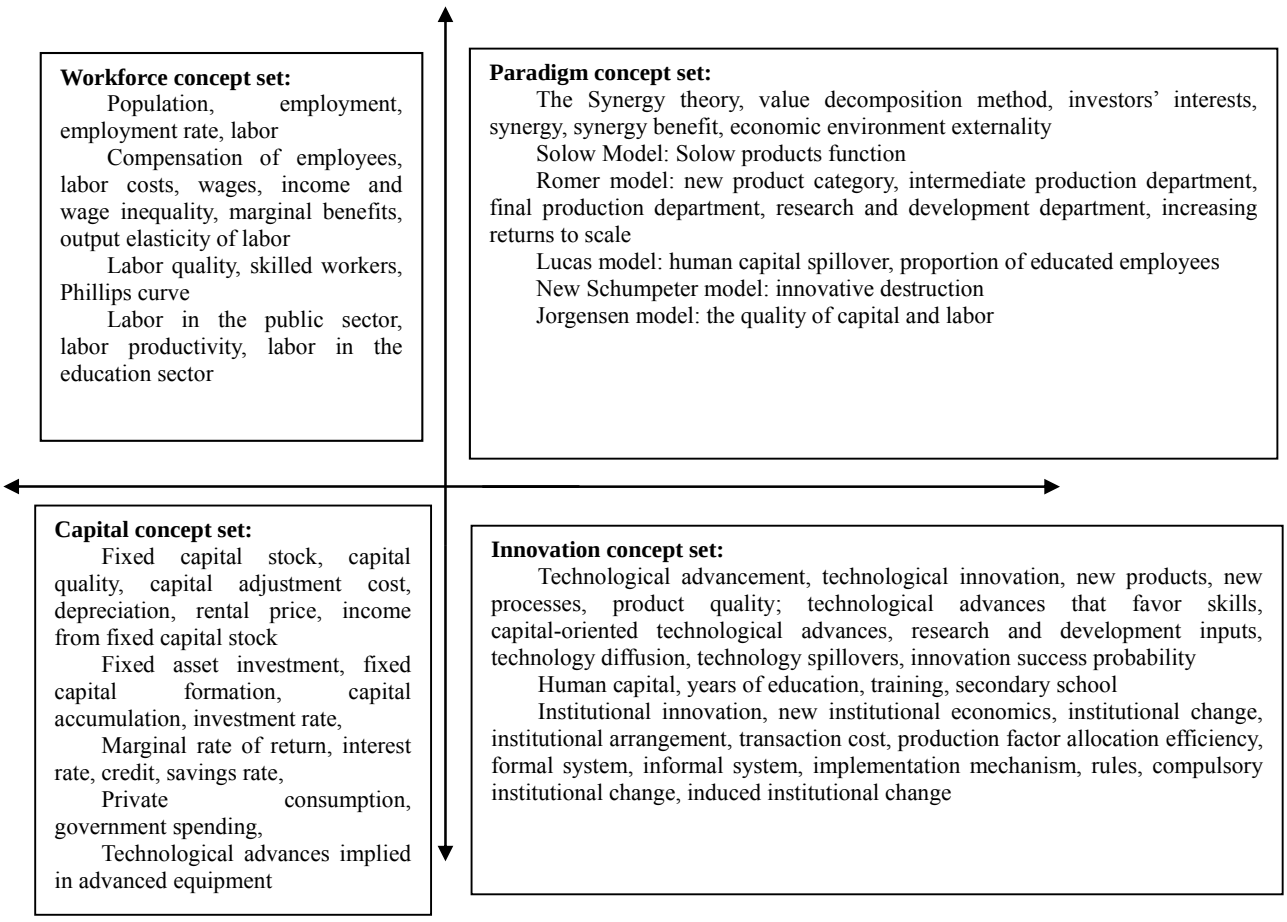

FIGURE I. CONCEPTUAL COORDINATE SYSTEM OF ECONOMIC GROWTH THEORY

In general, the structural relationship between module A, B, C, P and the whole system is:

$$J = P(A, B, C) \tag{1}$$

The three more specific forms of equation (1) are

$$J = A \circ_1 B \circ_2 C \tag{2}$$

$$J = A \pm B \pm C \tag{3}$$

$$J = A \bullet B \bullet C \tag{4}$$

For (3), a well-known economic example is GDP = investment + consumption + net exports; and in economics, another well-known example is Cobb-Dagras production function model $Y = aK^{\alpha}L^{\beta}$, Among them, a, K^{α}, L^{β} three modules, and Y (gross domestic product) is the overall nature of the economic system [18]. Jones and Manuelli (1990) have constructed production function models of the form

$$Y = AK + \Omega(K, L) \tag{5}$$

A simple form is

$$Y = AK + BK^{\alpha}L^{\beta} \tag{6}$$

From the point of view discussed here, it makes the incomplete and the complete model should be consisted of three modules.

Module A determines module C through module B, and module C reacts to module A through module B. Its formal model is

$$BOA \Rightarrow C \qquad A \Leftarrow {}^{C}\!/_{B} \tag{7}$$

Among them, "O" and "/" are the modes of action of B and A, C and B; "$\Rightarrow$" "$\Leftarrow$" is read as "decision" and "reaction".

In the Newtonian mechanics knowledge system,

according to Newton's cognitive model p_N , The relationship between the quality concept, the force concept f , and the acceleration concept ($a = d^2 r / dt^2$) is $f = ma$.

III. J Algorithm for Knowledge Discovery

A. J Algorithm Steps

The J-system approach considers the discovery process of scientific knowledge to be: accumulation of knowledge, selection of new research directions, and further exploration. This is divided into three steps.

The first step: based on the concept set [A] and [B], we can retrieve the literature, construct the text set U, perform text mining, and discover a series of research topics (cognitive models), P1, P2...Pi...Pj...

The second step: select one or several topics Pi, Pj, use the concept set to conduct literature retrieval to build the text library V, perform text mining, find a series of concepts, and select the concept set [C].

The third step: select a new theme Pk, based on [A], [B], [C], mining the literature, forming the text W, scientific research between [A], [B], [C] relationship.

B. Application in the Economic Growth Model

The following is an example of the establishment of a co-association theoretical model of economic growth[19].

The first step: discover a series of research topics

First, select the online version of the Web of Science database built by the National Institute of Scientific Intelligence (ISI). The Web of Science citation index method is used to reveal the intrinsic relationship of keywords in the economic growth literature, reflecting the development of economic growth theory. The search formula is set to: subject = ("economic* growth" and model*), a total of 744 data is obtained. Each data record mainly includes the title, author, abstract and citation of the literature).

The economic growth theory consists of three aspects: production function model, Solow model, and endogenous growth model. The so-called production function, according to the well-known American economist Samuelson PA , is a technical relationship It is used to illustrate the maximum output that a particular input (production factor) can produce, and different economists often construct different forms of production functions. The Solow model assumes a savings-investment conversion rate of 1; the scale return of an investment is a constant. Exogenous variables are: population growth rate, technological progress rate; endogenous variables are: output growth rate, capital growth rate. The Solow model proves that the path of economic growth is stable. In the long-term practice, only scientific and technological progress is the source of growth. The endogenous growth model includes Schumpeter's growth theory. One of the main tasks is to explain the reasons for the differences in economic growth rates and to explain the possibility of sustained economic growth.

The main concepts of economic growth theory include: labor, human capital, physical capital (public capital, foreign direct investment), technology (innovation, creative destruction), institutions, economic environment externalities.

The research methods based on economic growth theory mainly include: dynamic stochastic general equilibrium model (DSGE model), parameter estimation method, panel data method, econometric method, vector autoregressive model, cointegration test, optimization control method, development path Simulation, grey prediction.

The main problems related to economic growth theory research are: economic growth issues, energy consumption issues, carbon emissions, energy, environment, trade and other major issues.

Step 2: Discover a range of concepts

Select an online version of the Web of Science database built by the National Institute of Scientific Intelligence (ISI) was selected. The Web of Science citation index method is used to reveal the role of innovation in promoting economic growth and to reflect the relationship between economic growth theory. Search for the SCI-EXPANDED database in Web of Science, set the storage time = all years, and the search formula is set to: (Human capital or Innovation or Patents or Institution or Technological change) and (economic * growth) Or TI=(endogenous growth), a total of 3261 data (a full record for each data file). The bicomb software is used to obtain the co-occurrence matrix, and the ucitet software is used to draw the knowledge map.

The size of the point represents the centrality. The largest centrality is economic growth, human capital, endogenous growth, and growth. The biggest breakthrough is human capital and economic growth, this shows that the role of human capital in economic growth has received more attention; R&D (research and development) can promote the development of innovation, so innovation can be measured by R&D expenditure; human capital has a strong relationship with the number of years of education, so human capital can be measured by the number of years of education per capita; There are many factors that promote economic growth, including human capital, education, research and development, innovation, taxation, and knowledge. Human capital is the main factor.Therefore, human capital can be expressed in terms of years of education, and technology is represented by R&D expenditure.

The third step: study the calculation method of each concept

Select an online version of the Web of Science database built by the National Institute of Scientific Intelligence (ISI) was selected. The Web of Science citation index method is used to reveal the role of innovation in promoting economic growth and to reflect the relationship between economic growth theory. Search for the SCI-EXPANDED database in Web of Science, set the storage time = all years, the search is

set to: TI = ((institution or externalities) and (economic * growth)) or TI = (new institutional economics or Resource allocation efficiency), a total of 663 data (a full record of each data file). The bicomb software is used to obtain the co-occurrence matrix, and the ucitet software is used to draw the knowledge map.

It can be seen from the knowledge map analysis of Fig. 8 that the data envelopment analysis method can be used as a measure method for measuring the role of institutional innovation in economic growth.

Data Envelopment Analysis (DEA) is a new field of cross research in operations research, management science and mathematical economics.It is a quantitative analysis method for the relative effectiveness evaluation of comparable units of the same type.It based on multiple input indicators and multiple output indicators and uesd a linear programming approach.The DEA method and its model have been widely used in different industries and departments since it was proposed by the famous American operations research experts A.Charnes and WWCooper in 1978, and it is unique in dealing with multi-indicator inputs and multi-indicator outputs. Advantage.

Through the above three steps, it is determined that the direct factors determining economic growth include labor, human capital, physical capital (public capital, foreign direct investment), science and technology (innovation, creative destruction), institutions, and economic environment externalities;It is also possible to use the length of education to represent human capital and to use R&D funds to represent technology; On this basis, the economic growth model of the The Synergy Theory is established.Through the above three steps, it is determined that the direct factors determining economic growth include labor, human capital, physical capital (public capital, foreign direct investment), science and technology (innovation, creative destruction), system, economic environment externalities; and education can be used The term indicates human capital, and R&D expenditure is used to represent science and technology; and the data envelopment analysis method can be used to measure the role of institutional innovation in economic growth, and on this basis, the economic growth model of the co-association theory is established.

IV. CONCLUSION

Most scientific knowledge systems are composed of concept set [A], concept set [B], and concept set [C]. Different scientific knowledge systems have different cognitive patterns [P], so the scientific knowledge system consists essentially of four concept sets. The J system approach first represents the four sets of concepts in the four quadrants of the "concept coordinate system" and then explores the relationship among these concepts.

A module is composed of several variables that represent concepts. It is the symbolic entity with a clear meaning. The model consists of simple modules, but the modules can be very complex. From the concept to the model block, to the model, to the evolution of the model system, the scientific modeling logic can be revealed.

Through the construction case of the economic growth model of The Synergy Theory, it can be explained that the knowledge map provides a powerful analysis tool for the "four-set analysis method" proposed by the J-system methodology. The J algorithm greatly expands and deepens Swanson's knowledge discovery method based on non-related literature the ABC model.from the accumulation of scientific knowledge, the selection of new research directions, and the further exploration.

ACKNOWLEDGEMENT

This paper is the result of a project of the Henan Department of Transportation of China，a project of the State Intellectual Property Office of China, a key project of the Chinese Academy of Engineering,a project of the National Natural Science Foundation of China. Thanks for Waseda University professor Watada Junzo guidance.

REFERENCES

[1] Swanson D R. Undiscovered Public Knowledge[J].Library Quarterly, 1986,56(2):103-118 .

[2] Dong Fenghua, Lan Xiaowei. Literature-based knowledge discovery tool—Arrowsmith[J]. Journal of Information,2004,5,51-54.

[3] Hao Liyun. Guo Qizhen. Research progress in knowledge discovery of non-related literature [J]. Journal of Information,2006,3,342-348.

[4] Swanson D R. Smalheiser N R. An interactive system for finding complementary literature: a stimulus to scientific discovery[J].Artificial Intelligence.1997.91(2);183-203.

[5] Harter SP. Scientific inquiry: a model for online searching[J].Journal of the American society for information science.1984.35(2):110-117.

[6] Lindsay R K,Gordon M D. Literature-based discovery by lexical statistics[J].Journal of the American Society for information Science. 1999.50:574-587.

[7] Weeber. Klein, De Jong-van den Berg. Using Concepts in Literature-Based Discovery; Simulating Swanson’ s Raynaud-Fish Oil and Migraine-Magnesium.Discoveries[J].Journal of the American Society for Information Science and Technology,2001,52(7);548-557.

[8] Cui Lei,Zheng Huachuan. Research progress on knowledge extraction and mining from MEDLINE database[J]. Journal of Information,2003,4,425-433.

[9] Pratt W, Yetisgen-Yildiz M, LitLinker. Capturing Connections across the Biomedical Literature[M]. Proceedings of the International Conference on Knowledge Capture (K-Cap’03). Florida, October 2003.

[10] Srinivasan P. Text Mining: Generating hypotheses from Medline[J], Journal of the American Society for Information Science and Technology. 2004,55(5);396-413.

[11] Huang W, etc. Mining scientific literature to predict new relationships[J]. Intelligent Data Analysis. 2005, 9 (2), 219-234.

[12] Van der Eijk C. Van Mulligen E. Kors. JA. et al. Constructng an associative concept space for literature-based discovery[J].Journal of the Society for Information Science and technology.2004,55(5);436-444.

[13] Wren JD. Garner HR. Heuristics for identification of acronym-definition patterns within text; towards an automated construction of comprehensive acronymdefinition dictionaries[J]. Medthods of Information in Medicine.2002.41;426-434.

[14] Hu X, Zhang X, Yoo l. et al. A semantic approach for mining hidden links from complementary and non-interactive biomedical literature[DB/OL].2006.

[15] Jiang Zhaohua,LIU Xiaoting.From general system theory to J-system[J]. Seeking Truth, 1989,9(5):32-38.

[16] Jiang Zhaohua. J-System Reconstructability: A formalized Study[J].Int. J.General Systems, 2000,29(3):363-371.

[17] Liu Jianhua,Jiang Zhaohua.The Synergy Theory on Economic Growth: Comparative Study Between China and Developed Countries[M].Science Press & Springer,2018.

[18] Zhao Xiping.Research and practice of automatic control and hydraulic simulation in the Yellow River Diversion Project [M]. Beijing: China Water Resources and Hydropower Press, 2006.

[19] Jiang Zhaohua, Li Yafei, et al.Research on The Synergy Theory, Dalian University of Technology paper, 2019.