A Knowledge Discovery System for Detecting and Visualizing Knowledge Evolution of a Research Field

Wei Sun, Xuefu Zhang
AII of CAAS
Beijing, China

Huai Wang
SASMAC, NASG
Beijing, China

Abstract—**The paper proposes a knowledge discovery system for detecting and visualizing knowledge evolution patterns of a research field. It is mainly focused on co-word technology and core-based algorithm of tracking knowledge evolution. Firstly, the paper defines six kinds of knowledge evolution patterns systematically. Moreover, the paper illustrates the complex architecture of the system which contains four levels, i.e., basic data layer, pre-process layer, visualization layer and analysis layer. The paper elaborates key technologies involved in the system construction, knowledge structure building, knowledge evolution pattern detection and visualization. Then, as an example, the knowledge evolution patterns of hybrid rice field across 17 years are analyzed using 22 core journals of related fields, which verify the feasibility of the system preliminarily.**

Keywords-knowledge evolution; knowledge structure; evolution patterns; algorithm; hybrid rice

I. INTRODUCTION

With the widespread use of computer network, researchers have urgent needs for specialized and in-depth knowledge services, such as analysis services about hot spots, fronts and a series of evolutions in a research field gradually. Therefore, it has been one of the hot spots in the knowledge discovery field currently to identify and reveal phenomena and laws of knowledge evolution dynamically. Science mapping is the basis of knowledge evolution analysis. It is able to depict scientific structure and its dynamic characteristics. Co-citation and co-word network are main technical methods of knowledge mapping construction, but the majority of current studies [1-6] are mainly focused on the community evolution and little research has been carried out in order to analyze a pattern or a side of knowledge evolution. Even fewer application systems start with the various stage of knowledge evolution and are used for providing user with analysis services of dynamic knowledge evolution.

As the co-citation analysis result has a certain lag at the longitudinal studies, and it is more effective in mapping research front and intellectual base [7]. The main aim of this paper is to propose a knowledge discovery system for detecting and visualizing knowledge evolution patterns of a research field. It is mainly focused on co-word technology and core-based algorithm of tracking knowledge evolution. Firstly, six kinds of knowledge evolution patterns are defined systematically. Secondly, the complex architecture of the system is illustrated which contains four levels, i.e., basic data layer, preprocess layer, visualization layer and analysis layer. For a better understanding of the method for constructing the system, key technologies involved in the system construction, knowledge structure building, knowledge evolution pattern detection and visualization are elaborated. Then, as an example, the knowledge evolution patterns of hybrid rice field across 17 years is analyzed using 22 core journals of related field, which verify the feasibility of the system preliminarily.

This paper is organized as follows. Section 2 analyzes and defines knowledge evolution patterns. Section 3 introduces KEA (knowledge evolution analysis system) Architecture. Section 4 illustrates key technologies involved in the system construction. Section 5, uses the system for analyzing the knowledge structure and knowledge evolution patterns in hybrid rice, and verifies the feasibility of the system. Finally, some conclusions are drawn in Section 5.

II. KNOWLEDGE EVOLUTION PATTERNS OF A RESEARCH FIELD

There is a mechanism of evolution during scientific development process. The knowledge evolution in the paper focuses on all kinds of variations (knowledge birth, death, splitting, merging, rebirth, transfer) which are demonstrated by themes in a research field during the knowledge evolution process. Six kinds of knowledge evolution patterns are as depicted in Fig.1.

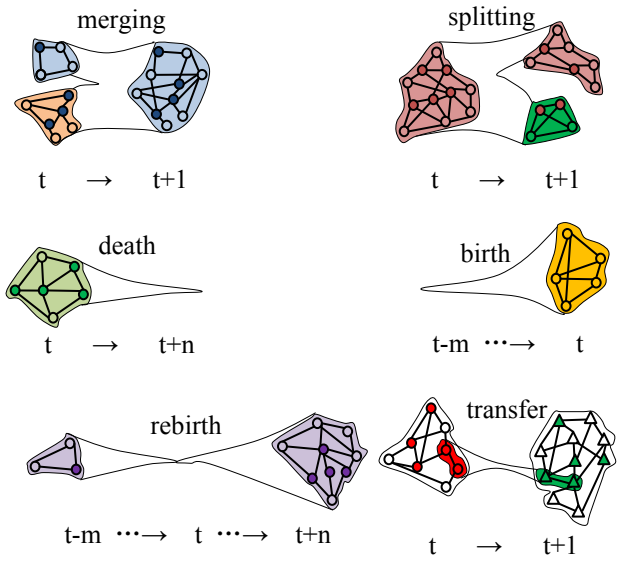

Figure 1. Diagram of Six Kinds of Knowledge Evolution Patterns

A. Knowledge Birth

During the knowledge evolution process of a field in a timeline, if an unrelated theme is a new one relative to themes in the previous any time slice, we call it newborn knowledge. The kind of theme may be a new research spot in the future, and we call the phenomenon a knowledge birth pattern. Analyzing the pattern is great helpful for researchers to select valuable scientific issues or hot spots.

B. Knowledge Transfer

Knowledge transfer in broad sense refers to the flow of knowledge from one node to another one. The node refers to the sender and receiver of knowledge.

During the knowledge evolution process of a field in a timeline，knowledge transfer in narrow sense refers that one theme in a time slice only transfer to another theme in the next time slice. Its stream is one-way and caused by means of knowledge borrow. Therefore, analyzing the pattern is helpful to analysis of knowledge borrow and reference between fields or disciplines, which can provide reference for finding valuable knowledge to researchers. In this paper, the knowledge transfer refers in particular to be in narrow sense.

C. Knowledge Splitting

During the knowledge evolution process of a field in a timeline, one or fewer themes in a time slice transfer to multiple or more themes in the next time slice. Its knowledge streams are multi-direction which is caused by trans-boundary research generally. We call the phenomenon knowledge splitting pattern. Analyzing it contributes to analysis of association between fields or disciplines and provides reference to researchers for finding new and trans-boundary scientific problems further.

D. Knowledge Merging

During the knowledge evolution process of a field in a timeline, multiple or more themes in a time slice converge on one or fewer themes in the next time slice. Its knowledge streams are multi-direction which is caused by joint research generally. We call the phenomenon knowledge merging pattern. Analyzing it contributes to analysis of cross and fusion association between fields or disciplines and is helpful for researchers to solve scientific problem in their own field with more frontier, effective and diversified knowledge.

E. Knowledge Death

During the knowledge evolution process of a field in a timeline, as a theme own important research value no longer, or researchers have been interested in it no longer, or related research tasks have been changed, the theme disappears or is transferred totally to others in the latter time slices, that is to say, the theme does not exist. We call the phenomenon knowledge death pattern. Analyzing it is helpful for researchers to grasp hot spots and avoid false proposition in scientific research.

F. Knowledge Rebirth

In some particular conditions, some newborn knowledge may be treated as the rebirth of a formerly theme although there is a snapshot gap between death of the old one and birth of the new one. We call the phenomenon knowledge rebirth pattern. One possible reason is from lack of some experiment datasets in a specified snapshot, but it is more likely to be the consequence of temporally lower activating rate of the people in question. Researching knowledge rebirth is helpful for researchers to clear scientific problem.

III. KEA Architecture

As is shown in Fig. 2, KEA Architecture is divided into four levels, namely, basic data layer, pre-process layer, visualization layer and statistical analysis layer.

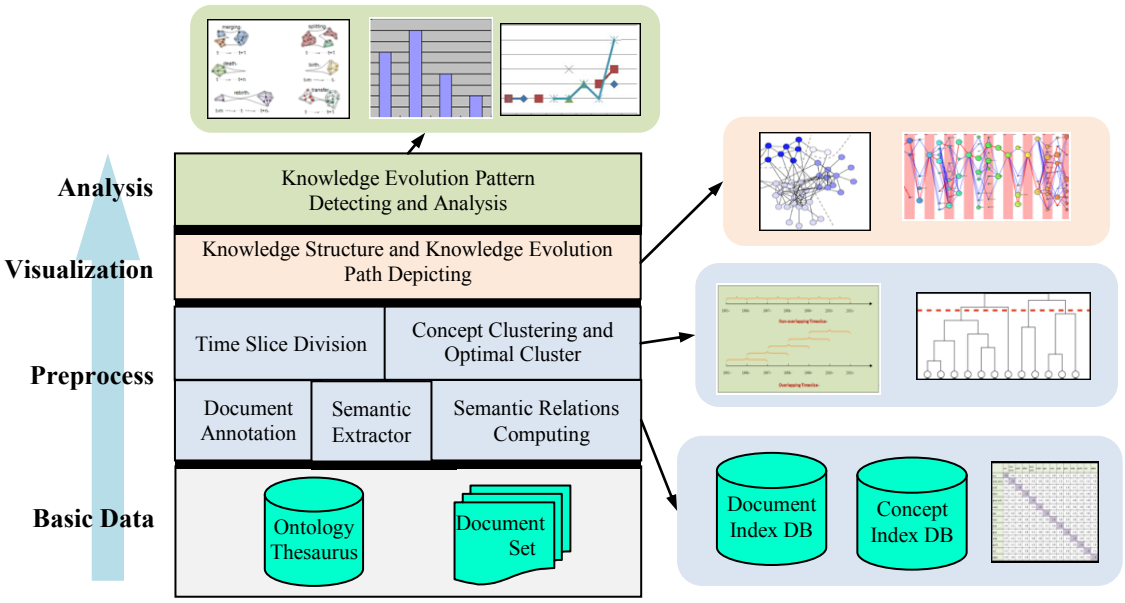

Figure 2. KEA Architecture

1) Basic data layer: This layer is data sources layer, including domain ontologies, thesauri, and document. Cleaned and standardized data can be used as the data basis of KEA system.

2) Pre-process layer: This layer is responsible for data processing before visualization. It consists of six parts, the text labeling generates the document index database, the semantic extraction generates the concept index database, and the calculation of semantic relations generates the concept association matrixes. On the basis of the concept association matrixes, combined with the time slice partition strategy and clustering algorithm, the layer will cluster concepts in every time slice respectively, and generates optimal clusters with optimal cluster strategy.

3) Visualization layer: This layer is responsible for the visual depiction of the knowledge structure and the knowledge evolution. It focuses on depicting elements of thematic cluster and thematic associations between clusters and within cluster in knowledge structure. It also depicts thematic streams in time zone view including the continuous evolution paths of same themes, the paths of knowledge splitting, merging, death and birth, the paths of knowledge transfer, as well as thematic name, the degree of thematic importance and other relevant information.

4) Statistical analysis layer: This layer is responsible for detecting knowledge evolution patterns including knowledge merging, splitting, birth, death, transfer and rebirth, and providing the statistics of themes in different time slice involved in every knowledge evolutionary pattern.

The pre-process layer, visualization layer, the statistical analysis layer can be interacted with. The user retrieves data through the pre-process layer and understands the elements of knowledge structure and streams of knowledge evolution through visualization layer. The visualization layer calls the data generated by pre-process layer through real-time interaction between the user and system interface and then completes the presentation of knowledge evolution path such as thematic information of knowledge structure, concept information and association information between themes etc.. Through statistical analysis layer, users can know about the specific themes involved in all kinds of knowledge evolution patterns and the essence of the knowledge evolution. In order to improve the response time of the system, all data generated before the visualization can be processed in advance, that is to say, the response time of the system can be improved by mean of visualizing the non-real time processed data.

IV. KEY TECHNOLOGY FOR CONSTRUCTING KEA

A. Knowledge Structure Building

1) Rules for semantic extraction

Treating standardized Concepts (Noun phrases) as nodes of thematic cluster can more fully reflect the contents of the theme. Therefore, the paper tags the phrase in document set semantically with the POS tagging technology [8], and then extracts noun phrases from the titles, abstracts and keywords of the literature. On this basis, the noun phrase concepts will be merged and standardized by means of making concept matching between professional thesauri, ontology and extracted noun phrases. Standardized noun phrases can be used as indexing terms of literature.

2) Clustering and optimal cluster selection

Amongst many clustering methods and algorithms, in this paper we apply a method proposed by Newsman which is able to deal with large networks with relatively small calculation time and requires no parameters from users. The algorithm is based on the idea of modularity [9]. In this paper, not only modularity Q but also silhouette [10] is used to select optimal cluster automatically. The silhouette value of a cluster, ranging from -1 to 1, indicates the uncertainty that one needs to take into account when interpreting the nature of the cluster. The value of 1 represents a perfect separation from other clusters. The modularity score ranges from 0 to 1. A low modularity suggests a network that cannot be reduced to clusters with clear boundaries, whereas a high modularity may imply a well-structured network. For every clustering result with nonnegative silhouette, we will calculate the index O one time for selecting optimal cluster respectively. The index O_i is defined as follows:

$$M = \frac{\sum_1^n (Si huatte_i + Q_i)/2}{n} (0 < i < n) \qquad \tau$$

$$O_i = |Sihuatte_i - M| + |Q_i - M| \qquad (2)$$

"i" is the ith calculation of clustering result. The maximum of O_i is the value of index O for selecting optimal cluster.

3) Automatic cluster labeling

In this paper, in order to extract a term to label a certain cluster, we extend the *tf*idf* [11] term ranking algorithm to clusters, and the *tf*idf* weight of term *i* for indexing cluster *s* is given by

$$W_{i,s} = tf_{i,s} \times idf_i \times \log\left(\frac{N}{df_s}\right).$$

As title reflects the important content of the article, term *i* of cluster labels is selected from noun phrases in the titles of articles of each cluster *s*. Where $tf_{i,s}$ is the number of occurrences of term *i* in the titles.

B. Knowledge Evolution Patterns Detection

Compare to non-core nodes, core nodes are representative and reliable and will be more accurate and effective to track knowledge evolution. In this paper, we use core-based algorithm [12] for reference to detecting knowledge evolution path, that is to say, we take advantage of not all nodes but core nodes to track thematic evolution. Specific depicting method of knowledge evolution path and detecting method of knowledge evolution patterns are defined in table 1. We apply the vote strategy-based algorithm to select core nodes [12].

TABLE I. RULES FOR DEPICTING AND DETECTING OF KNOWLEDGE EVOLUTION PHENOMENA

Knowledge evolution pattern	Evolution path	Detecting index
Knowledge birth	(1) $Core(C_i^{(t)}) \cap Node(C_j^{(t+1)}) \neq \emptyset$ (2) $Core(C_j^{(t+1)}) \cap Node(C_k^{(t-m)}) \neq \emptyset$ (3) $C_k^{(t-m)} => C_i^{(t)}$	A thematic cluster has no predecessor.
Knowledge death		A thematic cluster has no successor.
Knowledge splitting		A thematic cluster has more than one successor
Knowledge merging		A thematic cluster has no predecessor.
Knowledge transfer	(1) $Node(C_i^{(t)}) \cap Node(C_j^{(t+1)}) \neq \emptyset$ and $C_{size}(Node(C_j^{(t+1)}))$ (2) $(Core(C_j^{(t+1)}) \cap Node(C_k^{(t-m)}) = \emptyset$ $C_k^{(t-m)} => C_i^{(t)})$ or ($C_i^{(t)}$has no ancestor)	A thematic cluster has and only has one successor, and the successor one has no ancestor.
Knowledge rebirth	(1) $Core(C_{death}^{(t)}) \cap Node(C_{birth}^{(t+1)}) \neq \emptyset$	A birth thematic cluster appears in the evolution track of an old one.

Note: $C_i^{(t)}$is thematic cluster of index i in snapshot t; $Core(C_i^{(t)})$ is Core node set of $C_i^{(t)}$; $Node(C_i^{(t)})$ is Node set of $C_i^{(t)}$; $C_i^{(t)} \rightarrow C_j^{(t+1)}$ represents $C_i^{(t)}$ is a predecessor of $C_j^{(t+1)}$ or $C_j^{(t+1)}$ is a successor of $C_i^{(t)}$; $C_i^{(t-m)} => C_j^{(t)}$ represents $C_i^{(t-m)}$ is an ancestor of $C_j^{(t)}$; If there is an evolving chain $C_i^{(t-m)} \rightarrow C_x^{(t-m+1)} \cdots \rightarrow C_j^{(t)}(m \geq 1)$, $C_i^{(t-m)} => C_j^{(t)}$

C. Knowledge Evolution Visualization

Knowledge structure depiction is a foundation of knowledge evolution depiction. In the knowledge structure interface of KEA, as shown in Fig.3, the concept nodes in same theme are represented in a same color randomly. The nodes in a same cluster are very close to each other, while the nodes in different cluster are isolated from each other. Every thematic cluster is silhouetted by a kind of color and is labeled by a thematic name. For a large data set, users can choose whether or not to use Pathfinder to reduce the thematic network.

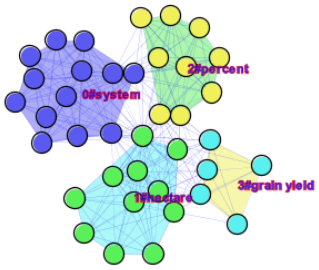

Figure 3. A network case of knowledge structure in rice field

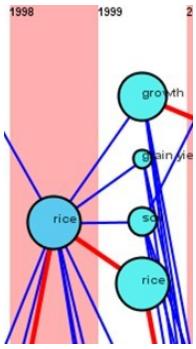

Figure 4. A case of knowledge splitting in two continuous time stamps

The knowledge evolution path contains components including nodes and associations between nodes. The knowledge evolution interface of KEA is a time zone view in which thematic nodes are distributed in their time stamp. Different from knowledge structure graphs, the knowledge evolution ones use a node to represent a thematic cluster and edges thematic stream. In the interface of the system, the blue and red lines represent the theme streams of knowledge evolution including knowledge birth, death, merging and splitting. Red lines represent same theme evolution streams among them. The green lines represent knowledge transfer streams. Fig. 4 is a case of knowledge splitting in two continuous time stamps.

V. SYSTEM PERFORMANCE

We collected 39659 papers on rice that had been published in 22 journals from 1995 to 2012 as data source. Three major thesauri, AGROVOC, CAB and NAL, are used to standard concept.

TABLE II. STATISTICS CASE FOR THEMATIC CLUSTERS OF KNOWLEDGE STRUCTURE

Number	Cluster label(number of concepts in the cluster)	Concept(fre)
0#	system(13)	other normal variety (1); system tgm (1); seed production (1); bentazon (1); marker (1); f-2 population (1); genetic mapping (1); bel (1); temperature fluctuation (1); development (2); use (2); response (2)
1#	hectare(11)	nitrogen nutrition (1); hybrid rice (6); grain (3); hectare (2); nitrogen (2); level (2); application (2); graded level (1); growth (1); effect (1); yield (1);
2#	percent(9)	percent (2); rice (2); fao (2); china (2); effort (1); period (1); government (1); country (1); indonesia (1);
3#	grain yield(5)	heterosis (1); hybrid (1); grain yield (2);panicle (2); number (2);

We make a fuzzy search in the "title" of the data set with "hybrid rice" and "one year" as a time slice, and then we have got 35 papers and 1864 concepts. The knowledge structures in every year snapshot and merged one from 1995 to 2012 are generated. Fig.3 is a knowledge topology of hybrid rice field in 2002 from which we can see it contains 4 theme clusters clearly, "system", "hectare", "percent" and "grain yield". The more detail information of the clusters is as described in Table 2. The knowledge topology is consistent with the 2002 time slice in the time zone view shown in Fig.5.

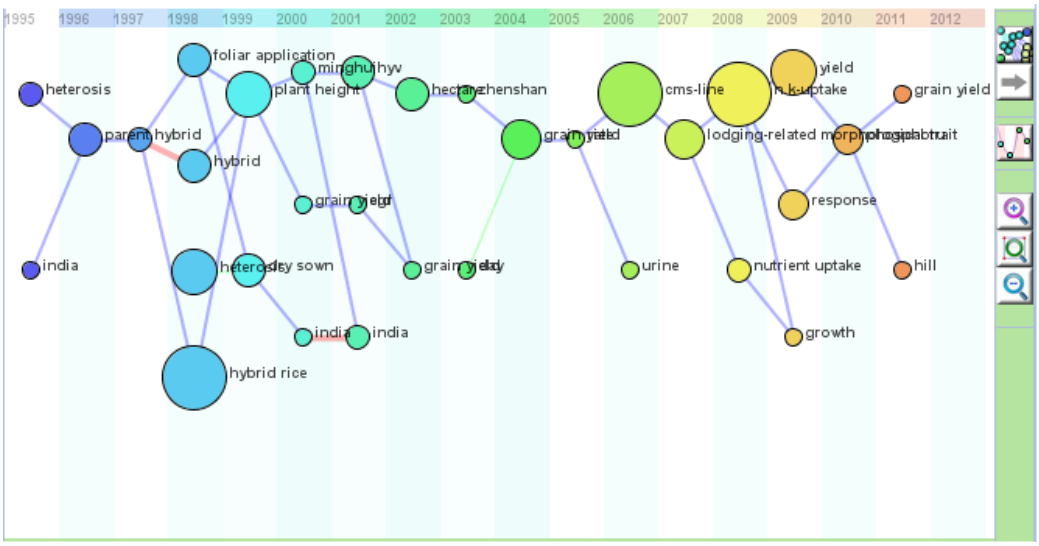

Figure 5. A time-zone view of knowledge evolution in hybrid rice field

TABLE III. STATISTICS FOR KNOWLEDGE EVOLUTION PATTERN OF HYBRID RICE FIELD

Knowledge evolution pattern	The former time slice		The latter time slice	
	Year	*Cluster*	*Year*	*cluster*
Knowledge merging	1995	heterosis, india	1996	parent
	1998	foliar application, hybrid, hybrid rice	1999	plant height
	2000	minghui, india	2001	india
	2001	hyv, cgr	2002	grain yield
	2008	n k-uptake, nutrient uptake	2009	growth
	2009	yield, response	2010	phosphoru
Knowledge splitting	1997	hybrid	1998	foliar application; hybrid; hybrid rice
	1998	foliar application	1999	plant height; dry sown
	1999	plant height	2000	minghui; grain yield
	2000	minghui	2001	hyv; india
	2001	hyv	2002	hectare; grain yield
	2005	rate	2006	cms-line; urine
	2007	lodging-related morphological trait	2008	n k-uptake; nutrient uptake
	2008	n k-uptake	2009	yield; response; growth
	2010	phosphoru	2011	grain yield; hill
Knowledge transfer	1998	heterosis	1999	dry sown
	2003	day	2004	grain yield
Knowledge birth	1997	—	1998	heterosis
	2002	—	2003	day
Knowledge death	1998	heterosis	1999	—
	2001	india	2002	—
	2002	grain yield	2003	—
	2003	day	2004	—
	2006	urine	2007	—
	2009	growth	2010	—
	2011	grain yield	2012	—
	2011	hill	2012	—

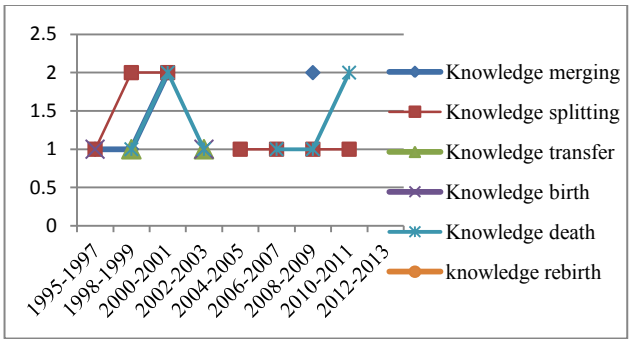

Figure 6. Statistics of each knowledge evolution patterns from 1995 to 2012

Fig. 5 is a time zone view which describes the thematic distribution and knowledge evolution of hybrid rice from 1995 to 2012. As shown in Fig.5, there are two same themes evolution paths, "hybrid" and "india". They are both lasted two years. Generally the bigger a theme is, the longer it lasts. Table 3 is detail information of clusters involved in every knowledge evolution pattern. See Fig. 6, from 2001 to 2011, knowledge splitting, knowledge death and knowledge merging patterns happened more frequently, which indicates trans-boundary and joint research activities become more frequent in the hybrid rice field in recent years. Of course, with these activities also appear some themes that have no research value or experimental basis. In contrast, knowledge transfer, knowledge birth and rebirth patterns happen little, which shows current research themes have become stable in the field relatively.

VI. CONCLUSTION

The main contributions of the paper are the definition of six kinds of knowledge evolution patterns and design of knowledge discovery system for detecting and visualizing knowledge evolution patterns of a research field based on co-word technology and core-based algorithm of tracking knowledge evolution.

The preliminary results analysis of knowledge structure and knowledge evolution in hybrid rice field indicates that the knowledge discovery system the paper proposes is feasible to some extent.

With the limitation of time, space and experimental conditions, there are some disadvantages in the system the paper proposes. (i) The effect of clustering in the system needs to be improved further. (ii)The paper validates the feasibility of the system only with papers in hybrid rice field preliminarily. So, the system still needs to be verified further and adjusted with more extension data. Besides, we will pay attention to the thematic life cycles during knowledge evolution.

ACKNOWLEDGEMENTS

This work is supported by National Key Technology Research and Development Program of the Ministry of Science and Technology of China during the "12th Five-Year Plan"(No. 2011BAH10B06).

REFERENCE

[1] Chen, C., Ibekwe-SanJuan, F., & Hou, J.. The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. Journal of the American Society for Information Science and Technology, 2010, 61, 1386–1409.

[2] Kandylas, V., Upham, S. P., & Ungar, L. H. Analyzing knowledge communities using foreground and background clusters. ACM Transactions on Knowledge Discovery from Data, Vol. V, No. N, December 2008, PP 1–34.

[3] Leydesdorff, L., & Rafols, I.. A global map of science based on the ISI subject categories. Journal of the American Society for Information Science and Technology,2009, 60, 348–362.

[4] Small, H., & Upham, S. P.. Citation structure of an emerging research area on the verge of application. Scientometrics, 2009, 79, 365–375.

[5] Small, H.. Tracking and predicting growth areas in science. Scientometrics, 2006, 68, 595–610.

[6] Upham, S. P., & Small, H.. Emerging research fronts in science and technology: Patterns of new knowledge development. Scientometrics,2010, 83, 15–38.

[7] Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 2006, 57(3): 359–377.

[8] Chen C. Turning points: The nature of creativity[M]. Springer, 2011.

[9] Newman, M.E.J.. Fast algorithm for detecting community structure in networks. Physical Review E, 2004, 69, 066133:1–15.

[10] Rousseeuw, P.J.. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 1987, 20, 53-65.

[11] Salton, G., Yang, C.S., & Wong, A.. A Vector Space Model for Information Retrieval. Communications of the ACM, 1975, 18(11), 613–620.

[12] Wang Yi, Wu Bin, Yang Shengqi. CommTracker: A core-based algorithm of tracking commuity evolution. Journal of Frontiers of Computer Science and Technology, 2009, 3(3): 282–292.