Analysis on the Technology Front of Workflow Based on Knowledge Mapping

JI Yong, ZHANG Mier
School of Management, Dalian University of Technology, Dalian 116023, China
jiy@neusoft.com, zhmill@dlut.edu.cn

Abstract—**Workflow is an important fundamental technology in computer field. We can make more reasonable usage of the technology through an accurate understanding of that. By using the information visualization tool (CiteSpace), social network analysis tool (Bibexcel) and statistics tool (SPSS) together, we can do some quantitative research of workflow related literatures which come from the Web of Science database with the theory of knowledge mapping and technology foresight. As a result, the author confirmed the top 5 hotspots and the top 3 research communities of workflow area by using word frequency analysis, and identified the top 3 evolution phases and the evolution trends of workflow technology by using word burst detection analysis.**

Keywords- knowledge mapping; technology foresight; workflow; co-citation analysis

I. INTRODUCTION

Workflow is a rapid development and widely used technology. Its main feature is to enable process automation and to make people coordinate with a variety of applications. In 1993, the foundation of WFMC (Workflow Management Coalition) [1] indicated that the workflow technology had become a mature software technology and entered into a rapid development track. As the workflow system can provide enterprise level automation controlling, monitoring, analysis and other functions, the technology has been considered as one of the most important fundamental technologies in computer field. So, in today's fast-growing economy environment, it is more meaningful to raise the information level by the way of tracking and mastering the technical development trends in workflow area.

Technology foresight is a general technology which can identify the strategic research areas and choose the maximize contribution to the economic and social benefits through the means of systematic study on scientific, technical, economic and social development in a long period [2]. Technology foresight helps to deeply discern the future technical developments and their potential impact on the economy [3]. With the rapid development of information technology, an enormous mass of scientific literature data brought us a very big challenge of improving the efficiency of technology foresight. As a useful supplement to the traditional methods of technology foresight, the knowledge mapping which based on bibliometrics analysis has been gradually developed in recent years. Especially in the complex technical areas, the method of the massive data processing capacity is not only efficient to enter the relevant technical fields, but also more precise to understand the knowledge structure, research hotspot and development trend of related fields by means of visual analysis tools.

II. RESEARCH APPROACH

Scientometrics reveals that the citation information among the scientific papers always represent the closely relationship of related discipline content. Through the citation cluster analysis, particularly from the mesh relationship between the citations, we can prove the structure and affinity of related disciplines. Also, we can do some analysis on the background, development overview, breakthrough achievements and future direction of a specific discipline, which reveals the dynamic structure and certain development laws of scientific development [4]. Knowledge mapping is an international emerging visual research method which based on the citation analysis theory of scientometrics and the development of information science. Its purpose is to apply the visualized methods to reveal the development and evolution trends of some disciplines, technology diffusion and dissemination laws, and the relations among authors.

In this paper, based on the document Co-citation Analysis and other mapping theory of scientific knowledge, we make some comprehensive use of Multi-dimensional Scaling, Cluster Analysis, Factor Analysis and Multivariate Statistical Analysis to identify the hotspot and research communities of workflow domain. And we do some visual analysis of scientific bibliography data (in particular, citation data and keyword data) to show the development trends and fronts of some disciplines by use of the information visualization tool (Citespace).

A. Multi-dimensional scaling and cluster analysis

1) Co-word Analysis: We can use Bibexcel (social network analysis tool) to do co-word analysis. In this paper, we choose top 50 high cited words according to the key words frequency of searched literatures. After co-word analysis, the Bibexcel generated co-word matrix. And then the results are converted into the correlation matrix by the SPSS software. The theory of co-word analysis defined a relationship type as co-occurrence when two key words appeared in one literature. The more frequency of co-occurrence keywords, the closer of their research subject [5]. Also, the relationship reflected as a closer distance between them in the knowledge map.

This work is supported by National Natural Science Foundation of China (70872015) and Specialized Research Fund for the Doctoral Program of Higher Education (20090041110008).

2) Multi-dimensional Scaling Analysis: We can use SPSS software to do multi-dimensional scaling analysis. The method uses the low-dimensional space (usually two-dimensional space) to express the relationship among the objects, and using the plane distance to reflect the degree of similarity among objects. In a co-word map, the key words (each point) position shows the similarity among keywords. A high degree of similarity of the key words crowd together to form a research hotspot. The more keywords in the middle indicate that the more key word associated with it, and the more central in the subject. On the contrary, the more lonely, the more in the periphery. Therefore, by multi-dimensional scaling analysis, it is easy to determine a subject hotspot or a discipline location in some research field [6].

3) Cluster Analysis: Statistically, we defined the method which classifies things according to certain requirements and the rules as cluster analysis. Clustering is to convert some heterogeneous group into some higher cluster or sub-group with a more similarly structure. In clustering process, it does not need to do any pre-defined categories for classification. The data is clustered by their similarity, and the meanings of clusters are relying on the qualitative description which can be learned later. In this study, we use "hierarchical cluster analysis" as the research method. For example, for the key words in the cluster, the first is to consider each separate key word as a cluster, and then merges the nearest two clusters. By recalculating the distance among all clusters in the same way, each iteration reduces a cluster, until all of the key words are clustered.

B. Factor Analysis

The basic purpose of factor analysis is to use a few factors to describe the link among the indicators or elements. The methods try to merge several variables which closely related into one factor, and trying to use less number of factors to restore most of the information being analyzed. Among these methods, the principal component analysis always transforms a given set of variables first, and converts them into a group of unrelated variables while maintaining the variable without changing the total variance. Also, the process is always making the first principal component to have the maximum variance value, and so on. So, we can easily determine the communities or the distribution information of a given subject area by combining the principal component analysis and Scientometrics methods together.

C. Trend Analysis

This study uses CiteSpace 2.1 toolsets which developed by Dr. Chen Chaomei as a means of information visualization and technology trend analysis in the field. In the process of using the tool, the first, we do some pre-analysis on the plain text data downloaded from the Web of Science database. The second, we define a reasonable interval period (Time Slice) according to the time span of the range selection. The third, we import the literature and citation data of related fields into the CiteSpace project, and set the relevant parameters as the following: Node Types-Cited Reference; Term Sources-Title, Abstract, Descriptor, Identifiers; Term Selection-Burst Phases; Pruning-Minimum Spanning Tree, Pruning Sliced Tree, Pruning the Merged Network. At last, we generate the map of evolution trends by the "Time-zone View" functionality provided by CiteSpace software.

D. Data sources

The data used in this paper is derived from Web of Science database provided by the ISI (Information Sciences Institute). On the types of literature, considering the fact that a lot of papers occurred in the meeting with the rapid development in computer field, we took some proceeding papers in addition to the article types. In the keywords selection, first we got the core literature sets searched by the keyword of "Workflow", and then added some high-frequency words in the core workflow areas, such as WFMS, BPEL, and WFMC as the supplemented searching keywords which extracted from the core literature abstract sets by the method of word frequency analysis. In the field coverage, we chose a number of computer cross-disciplinary fields areas besides the computer-related fields. Base on the retrieval principle above, we retrieved 6528 documents together with the cumulative citation frequency of 19,256 times in the whole time period. From the searching results, we found that the number of workflow and related citation is steady growing in recent years, which reflects the fact that the workflow is still a burgeoning technology.

III. Analysis and Results

A. Workflow Research front

Based on the top 30 high frequency keywords refined from 6528 literatures of workflow area, we got the co-word knowledge map (see Figure 1). We can find four obvious core knowledge groups by using multi-dimensional scaling and cluster analysis methods, and the largest group implies two subgroups which have not very clear boundary.

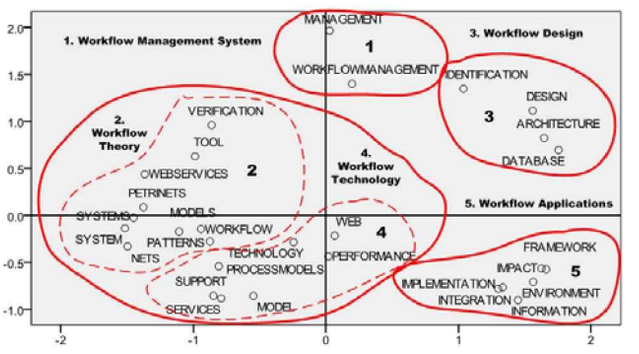

Figure 1. The co-word knowledge map of workflow area

In order to grasp the workflow research hotspot more accurately, we extracted five principal components through principal component analysis (see Table I). The cumulative variance contribution rate reaches to 94.668%, so it means a higher representation of workflow area. Based on the result of principal component analysis, the biggest knowledge group can be further divided into two separate subgroups. Accordingly, the workflow area can be divided into five research fronts.

Based on the analysis of high frequency keywords in each subgroup, the research fronts can be summarized as: 1. workflow management system; 2. work flow theory; 3. workflow design; 4. workflow technology; 5. workflow application.

TABLE I. THE PRINCIPAL COMPONENT ANALYSIS OF WORKFLOW RESEARCH FRONT

Principal Factor	*Extraction Sums of Squared Loadings*		
	Total	*% of Variance*	*Cumulative %*
1. Workflow Management System	13.757	47.439	47.439
2. Workflow Theory	6.541	22.555	69.994
3. Workflow Design	3.830	13.207	83.202
4. Workflow Technology	1.910	6.586	89.788
5. Workflow Applications	1.415	4.880	94.668

In the knowledge map of workflow research front, the variance contribution rate of workflow management system component reaches to 47.439%, indicating that the workflow management system is the core of well-deserved in workflow area. In addition, the distributed location between workflow theory and workflow technology is relatively close in the knowledge map, showing that the two components have a relation of influencing each other and promoting each other. In particular, the recent major innovations of workflow technology almost had a profound impact on workflow theory. If we analyze the five areas in the knowledge map with a clockwise direction, we will find: the workflow theory and workflow technology constitute the research base of workflow, and the workflow management system which based on the customized workflow design promotes the popularity of workflow applications.

B. *Workflow Research Community*

Based on the top 30 high frequency authors retracted from the literatures of workflow area, we got the author co-cited knowledge map by using multi-dimensional scaling and cluster analysis methods (see Figure 2). From the map, we can find three obvious author communities, and it precise matches the result obtained by the principal component analysis method (see Table II). Through the analysis of high frequency keywords of the highly cited literatures in each community, the workflow research communities can be summarized as: 1. workflow theory; 2. workflow applications; 3. workflow standards.

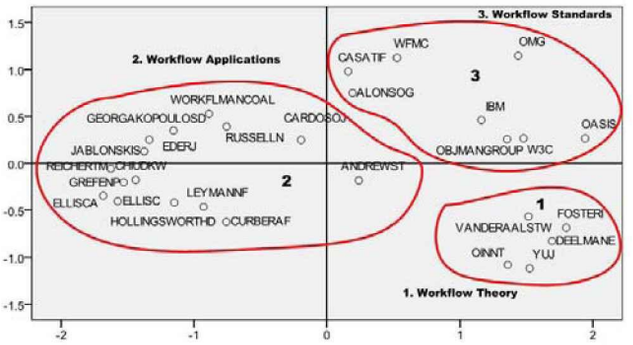

Figure 2. The author co-cited knowledge map of workflow area

TABLE II. THE PRINCIPAL COMPONENT ANALYSIS OF WORKFLOW RESEARCH COMMUNITY

Principal Factor	*Extraction Sums of Squared Loadings*		
	Total	*% of Variance*	*Cumulative %*
1. Workflow Theory	13.073	46.689	46.689
2. Workflow Applications	4.672	16.686	63.374
3. Workflow Standard	2.982	10.651	74.026

1) Workflow Theory: This community owns the largest variance contribution rate (46.689%), but contains only five core members, indicating that the relevant research of workflow theory is very stable. It is worth mentioning that a representative of the community named as Van der Aalst has a amazing record whose accumulated literature citations reaches to 3573 times. So, he is worthy of the leading authority in workflow research area. His research results on the Petri net [7] and workflow patterns [8] have become the classic theoretical model of workflow.

2) Workflow Applications: This community has more members with more uniform distribution. It indicates that the research of workflow has not yet formed clear factions, and this due to the fact that no focus in various workflow scenarios. The research on workflow model and application framework of the representative in this field named as Georgakopoulos (cited 507 times) consisted the foundation of common workflow application model [9].

3) Workflow Standards: Although this community has the lowest variance contribution rate, the literatures are relatively new and have the trends of accelerating aggregation. In the aspect of workflow standards setting, as the representative of enterprise researchers in workflow area, IBM has gradually broken the situation of fighting each other composed by WFMC, OMG, W3C and other standards organizations in recent years. Accordingly, a BPEL (Business Process Execution Language) [10] centric situation is gradually formed. As shown in the knowledge map of workflow starndard research community, the fact that IBM has a relative central position also confirmed its significant impact on the setting of workflow standards.

C. *Evolution Trends of Workflow Technology*

After importing the workflow literatures and citation data into the CiteSpace project, we can detect the burst phases from numerous keywords by using the "Time-zone View" functionality. Also, we can get the time distribution of word frequency and the evolution trends of workflow technology by word frequency changing tendency (see Figure 3). Based on the gathering of knowledge shown in the map, the evolution of workflow technology can be divided into three phases: the foundation phase (1995-1999), the connotative consolidation phase (2000-2004) and the denotative development phase (2005 - present).

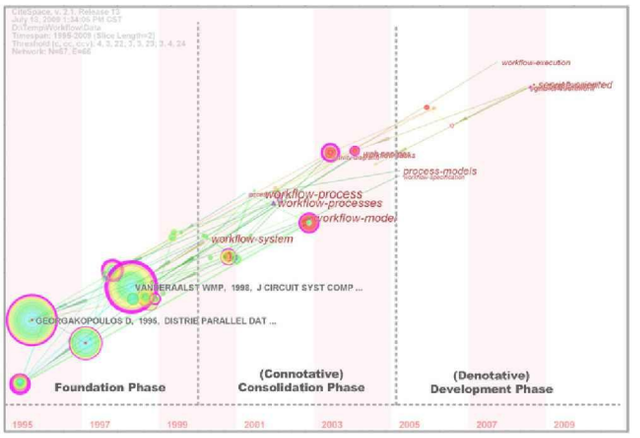

Figure 3. The knowledge map of workflow evolution trends

1) Foundation Phase: The focus of this phase is the definition of basic workflow theoretical model. We can clearly find out that the two core clusters consist of Georgakopoulos and Van der Aalst are the two largest clusters in the whole knowledge map. On one hand, it's due to the fact of earlier published, more subsequent referenced; on the other hand, it reflects that the basic structure and running architecture have not major changes with the passage of time. This conclusion also can be confirmed by the cited history of cluster owned by Van der Aalst in micro-perspective (see Figure 4). In the figure we can see, the cited cases owned by Van der Aalst do not decrease significantly over time. Even in 2009 when the paper published 10 years later, it remains a more active state of being cited. Therefore, either from the macro or micro perspective, the phenomenon of highly cited literatures in basic phase indicates that both the workflow structure and the running architecture are have been developed to a relatively mature stage. Accordingly, its technology basis will have no major changes within a short period, which established a good foundation for the following technical development.

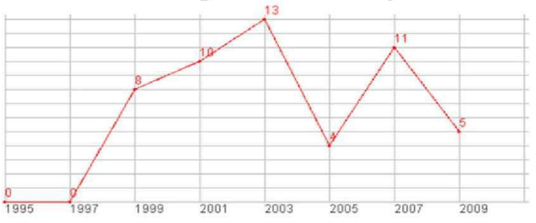

Figure 4. The cited history of cluster owned by Van der Aalst

2) Connotative Consolidation Phase: This phase has a significant feature: two high frequency words of Workflow Model and Workflow Process occurred alternately, and reflect as medium-sized cluster. This indicates that the work of this stage is based on the previous stage, and its main contribution is to consolidate the basic model and architecture of workflow.

3) Denotative Development Phase: In this phase, there are not any outstanding clusters which may be due to the late to the published literatures. The research hotspots are more concentrated on the Web Service, Service Oriented and Grid Environment. This shows significant differences with the former two phases which consist of Workflow Process and Workflow Model. On the background of the popularity of SOA (Service Oriented Architecture) and Grid Computing, as a major computer software middleware technology, workflow must do some adaptive changes for suiting the tide. Also, it explains why the recent workflow researches focus on Service and Grid. We believe that the trend will remain gradually evolving towards a higher development stage. But the related impact will only change the external workflow interaction mode, and the basic model or running architecture will not suffer the revolutionary impact.

IV. CONCLUSION

By using the methods of citation analysis, cluster analysis and information visualization tool on the literatures of workflow area, we can accurately find out the research hotspots, research communities and evolution trends of workflow. In conclusion, the workflow research hotspots include workflow management system, workflow theory, workflow design, workflow technology and workflow applications; the workflow research communities include workflow theory, workflow applications and workflow standards; the workflow evolution trends are towards the direction of Service and Grid.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (70872015) and Specialized Research Fund for the Doctoral Program of Higher Education (20090041110008).

REFERENCES

[1] WFMC. http://www.wfmc.org/.

[2] Martin, B.R., Johnston, R. Technology foresight for wiring up the national innovation system - Experiences in Britain, Australia, and New Zealand. Technological Forecasting and Social Change, vol. 60, pp. 37-54, 1999.

[3] Salo, A., Cuhls, K. Preface - Technology foresight - Past and future. Journal of Forecasting, vol. 22, pp. 79-82, 2003.

[4] ZHAO Kai, CHEN Yong-jun. On Technical Paradigm of Circular Economy--The"XR" Principle. China Industrial Economy, vol. 6, pp. 44-50, 2006.

[5] Polanco, X. Co-word analysis revisited: Modelling co-word clusters in terms of graph theory, Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics, pp.662-663, 2005.

[6] Liu Lin-qing. Mapping knowledge domains of research with document co-citation. Studies in Science of Science, vol. 02, pp. 155-159, 2005.

[7] Van der Aalst, W.M.P. The application of Petri nets to workflow management. Journal of Circuits Systems and Computers, vol. 8, pp. 21-66, 1998.

[8] Van der Aalst, W.M.P., Barros, A.P., ter Hofstede, A.H.M., Kiepuszewski, B. Advanced workflow patterns. In: Etzion, O., Scheuermann, P. (eds). Cooperative Information Systems, Proceedings. Berlin: Springer-Verlag Berlin, pp. 18-29, 2000.

[9] Georgakopoulos, D., Hornick, M., Sheth, A. An overview of workflow management: From process modeling to workflow automation infrastructure . Distributed and Parallel Databases, vol. 3, pp. 119-153, 1995.

[10] BPEL. http://www.oasis-open.org/committees/wsbpel/.