Visualization of Individual's Knowledge by Analyzing the Citation Networks

Tze-Haw Huang and Mao Lin Huang
Faculty of Information Technology
University of Technology, Sydney, Australia
{thhuang@it.uts.edu.au, maolin@it.uts.edu.au}

Abstract

Visual analysis of knowledge domain is an emerging field of study as science is highly dynamic and constantly evolving. Behind the scene, a knowledge domain is formed and contributed by enormous researchers' publications that describe the common subject of the domain. There is large number of significant activities have been carried out to visualize and identify the knowledge domains of research projects, groups and communities. However, the research on visualizing the knowledge structure at individual level is relative inactive. It is difficult to track down the individual's contribution to the subject and the degree of the knowledge they possess.

In this paper, we are attempting to visualize the individual's knowledge structure by analyzing the citation and co-authorship relational structures. We try to analyze and map author's documents to the knowledge domains. By mapping the documents to knowledge domain, we obtain the skeleton of knowledge structure of an individual. Then, we apply the visualization technique to present the result.

Keywords--- **co-authorship, information visualization, information analysis**

1. Introduction

Gaining the knowledge is the cognitive process of integrating experience, perception and reasoning into a representation which reflects the author's confidentiality of understanding the subject domain. In academic research, the approach of proofing the author's understanding of the knowledge is by transforming the knowledge into document that is also known as scientific literature and publishes it to a particular knowledge domain for knowledge sharing across community boundaries. Discovery and mapping the knowledge structure has also gained enormous awareness in academic research. Mapping knowledge is one of emerging topic in KVD that uses sophisticated data mining analysis and visualization techniques to fairly identify the research areas, experts and etc.

The advance of Internet in late ninety has made virtually any information ubiquitous with aim to promote knowledge sharing and idea exchanging as a result the scientific publications has grown exponentially in the mean time. There were various scientific literature search engines such as DBLP and SCSI have maintained enamors dataset of categorized publications grouping into knowledge domains, these high volume of data has intrinsic raised the interest of researchers to explore the knowledge experts within the domains and visualizing the knowledge structure and boundary of experts across domain. Visualization techniques are critical to explore the individual knowledge acquisition and identification of active knowledge experts in which have always interesting the government bodies, research organizations and laboratories as they are persistently searching for experts for purpose of research funding and award cutting-edge project.

Intrinsically, attempting in the visualization of the knowledge and experience of an individual is complex as it involves the collection and analysis of scientific literatures grouping by an author in order to build the skeleton of individual's knowledge boundary. The ability to map the knowledge involves database mining and document analyzing. More importantly, the effective presentation sense to transform the abstract data into intuitive way.

We are neither focusing on identifying the most cited, influential authors nor does scientific collaboration in general in face these topics have already received widely discussion. Instead, we were searching for approaches to present an individual researcher's knowledge structure by analyzing the given documents. It is possible that an individual's knowledge is spawned across domains that is also known as knowledge branching.

This paper is organized as follow in Section 2, we introduced the approaches that we used to pre-processing the document and mapped a document into knowledge domain. In Section 3, we presented the knowledge structure analysis visualization and finally a conclusion in Section 4.

2. Knowledge Mapping

Figure 1 illustrates a framework of the knowledge mapping, which consists of five steps:

1. Document mining
2. LSA
3. KD mapping
4. Knowledge quality analysis
5. Knowledge structure analysis

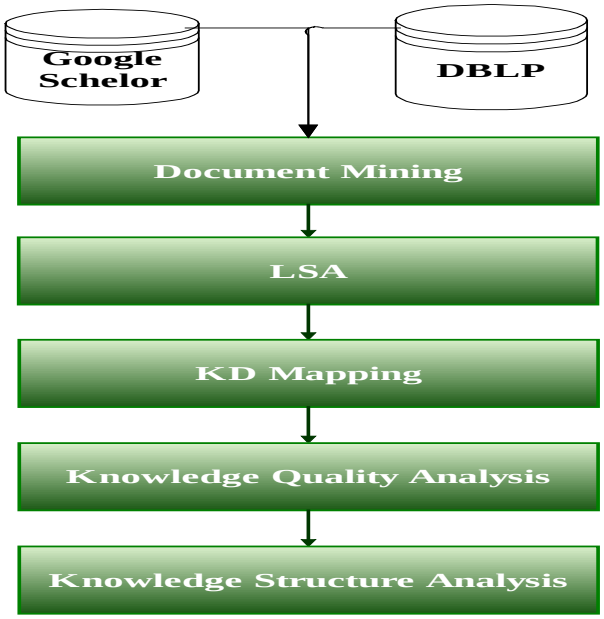

Figure 1: RFCD mapping steps

2.1 Data Source

The dataset used for analyzing and visualization were ported from the well-known online DBLP (http://www.informatik.uni-trier.de/~ley/db/), which is one of the largest Computer Sciences bibliographic data source available on the Internet. It provides information on major computing journals and conference proceedings between year 1936 and 2006. The distinctive advantage of DBLP over another scientific bibliographic database such as CiteSeer is the easily identification of authors. DBLP provides full author names in a publication and CiteSeer uses only the initial of authors which will often cause confusions when multiple result-set has been returned when querying on an author name and also performing insert operation to the database that violates the constraints. Thus, the decision was made to use DBLP as our preference of dataset.

A summary of statistics of DBLP studied is given in Table.1

Number of authors	442,886
Number of papers	678,296
Average authors per paper	2.40
Average papers per author	3.67
Conference accepts most papers	Communications of the ACM, 6892

Table.1: The statistics studied was based on DBLP and as a benchmark. Please note the data collected by DBLP may not be complete.

2.2 knowledge Domain Classification via RFCD

Scientific knowledge is disseminated in many domains. A knowledge domain is a content of particular knowledge or an area of expertise that scientists learn about which is easily accessible with information offered in structured and consistent way. Thus, a collection of texts, files and images does not form a knowledge domain unless the various bits of information found are integrated and related.

Our knowledge domain classification is based on Australia Research Fields, Courses and Disciplines Classification (RFCD) [4] as shown in Figure.1 which is the collective name for a set of related classifications and the research activities were classified according to the field of research undertaken and it is generally reflects the overall structure of disciplinary fields. The RFCD is organized in a hierarchical structure with 24 divisions, 139 disciplines and 898 subjects, in Figure 2 has shown the subset of RFCD structures. The rich set of research area classification provided by RFCD is sufficient to cover the majority of identifiable knowledge domains.

2.3 Text Pre-Processing

Building blocks of mapping scientific document to a knowledge domain require the data mining and development of pre-defined vocabularies. By given a corpus of published document d_i with a set of document keywords $\{w_{i,j}, j=1,..., j=N\}$ where $w_{i,j}$ represents a keyword that summarized the content of document to a subject domain. The keywords may come in many forms such as verbs, norms or adjectives, in order to reduce all forms of the word to a base or stemmed form, we applied Porter stemming algorithm [2] on $w_{i,j}$ by removing the suffix to reduce the related words to the same stem (i.e. networking → network).

The purpose of preliminarily applying word stemming on each keyword is to improve the matching efficiency by reducing the number of unique words in keyword population μ and facilitate the automatic document to subject domain mapping process.

RFCD		
Discipline	Subject	Description
280100		**Information System**
	280101	Information System Organization
	280102	Information System Management
	208103	Information Storage, Retrieval & Management
	280104	Computer-Human Interfaction
	28105	Interface and Presentation

Figure 2: Fraction of RFCD

2.4 Latent Semantic Analysis

A document is a collection of words that composed together to describe a particular knowledge. Latent semantic analysis (LSA) has been used to relate the knowledge domain to a document; it is a technique in natural language processing by extracting and representing the contextual meaning of words by statistical computation to a corpus of text. LSA utilizes SVD [6] that is a matrix decomposition technique for uncovering latent data structures while removing noise. SVD is defined as below and the results were shown in Figure 3 and 4.

$$A = USV^T \qquad (1)$$

Where A is the decomposed term-document matrix, U and V are orthogonal and S is a diagonal matrix.

Knowledge Domains

Document Keywords	A = K1	K2	K3
Visualization	1	0	1
Map	1	1	0
DBLP	2	0	0
Citation	1	1	1

Figure 3: Co-Word Matrix

A_K

	k1	k2	k3	Total
Visualization	0.997	0.5001	0.5001	1.9972
Map	0.997	0.5001	0.5001	1.9972
DBLP	2.0012	0.0003	0.0003	2.0018
Citation	1.0006	0.9997	0.9997	3
Total	4.9958	2.0002	2.0002	

Figure 4: SVD Calculation

We adopted the SVD to obtain the document to knowledge domain mapping by treating document as knowledge domain and the keywords as terms. The term-document matrix A is constructed from a given corpus of a scientific document which gives the relationship between terms and documents. LSI transforms this into a relationship between the terms and concepts. Each concept is a vector and the elements within it were assigned the weight. The most significant words found would be used to identify to which knowledge domain that this document is most suitable mapped to.

Applying the technique on the document published by an author, the function will return the knowledge domain the input document is most likely mapped to. We could obtain the knowledge domains possessed by an individual or more precisely the knowledge and research experience possessed and classified by knowledge domains as described in (1).

$$rfcd = LSA(d_{k1}.....d_{kn}) \qquad (2)$$

where a set of corpus $d_{k1}.....d_{kn}$ from a document is analyzed using SVD with the most likely mapped RFCD code is suggested and in fact, a RFCD represents a knowledge or research domain. Thus, the overall knowledge structure of an author is therefore given by:

$$RFCSs \in \sum_{i=0}^{N} f_{lsa}(d_{i,k1}....d_{i,kn}) \qquad (3)$$

2.5 Knowledge Quality Analysis

By applying the derivatives in Section 3.4, we have obtained the skeleton of one's knowledge structure by mapping the document to RFCD. It is difficult to determine the knowledge ranking of an author simply by counting the number of documents in each RFCD grouping. The quality of the knowledge should be determined by the citations received instead of quantity.

Scholars often cited the published work of others in their own research work to gain supportive evidence of their work. Such citations can be used to estimate the impact of scientific publications of an author. The well-known indicator for the quality of a publication is its impact factor [5] which provides a quantitative unit for each journal proportional to the average number of citations per paper published in the previous two years by Thomson ISI in the Journal Citation Report (JCR). Unfortunately, conference proceeding publications were not covered by JCR. Nevertheless, Google Scholar makes it possible to easily access citation data of over millions of publications and authors.

The assumption behind the use of citation count to measure the knowledge quality is that citation generally reflects the utilization, contribution and quality of a scientific publication in international aspect. Furthermore, the publication venue is also an important

factor that should be taken into account when considering the author's knowledge quality analysis.

Thus, citation count is one of measures that we used to consider the quality of a publication by an author. If we have concluded by counting citations to determine the quality then that would introduce the bias and fairness problems. That is, the quality of referencing publications might vary if we are comparing the quality of cited publications. WWW is one of example that we could picture to have same phenomenon. For example, Google uses page link count as part of algorithms to determine the page ranking when responding to users search query. In a general sense, a page is getting popular as the number of links increase by other web pages. However, web pages with relative few links may also be prominence if links come from other prominence web pages. Therefore, an author might be an expert in the knowledge domain even though he or she has published few publications with little citation received and come from other prominence journals.

3. Visualization

In previous section, we have grouped the documents into knowledge domains of an author. In this section, we would present the result using various visualization approaches.

3.1 Citation Structure

We used 2D ring to visualize the change of citations received by a document over time. Number of citations received by a document of an author reveals the importance of the document in the published venue of domain. The author's knowledge has been acknowledged as established and thought by others when creating new knowledge.

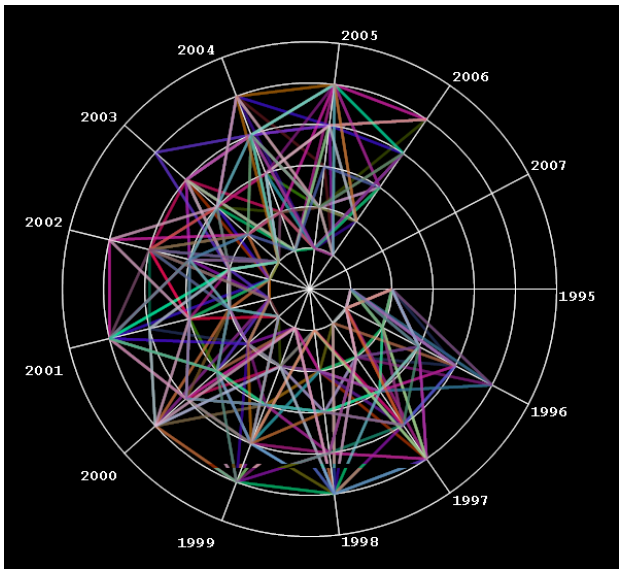

Figure 5: Citations received for each document

Figure.5 shows the citations received for all the documents mapped to knowledge domain of an author over time. Comparison between citations received could be used to identify the area of knowledge that an author has mostly acknowledged in author's knowledge structure. For example, suppose an author has expertise in two knowledge domains as a result of grouping documents published by applying the knowledge mapping technique described in Section.3. By comparing the citations received over time across knowledge domains that could help to understand an author's main knowledge stream.

In Figure 6, we present a view of all the citations received by knowledge domain.

Similarly, in Figure 7 presents a historical view of acknowledge by citation of individual document. These citation analyses of document enhance the view towards the author's past work recognition.

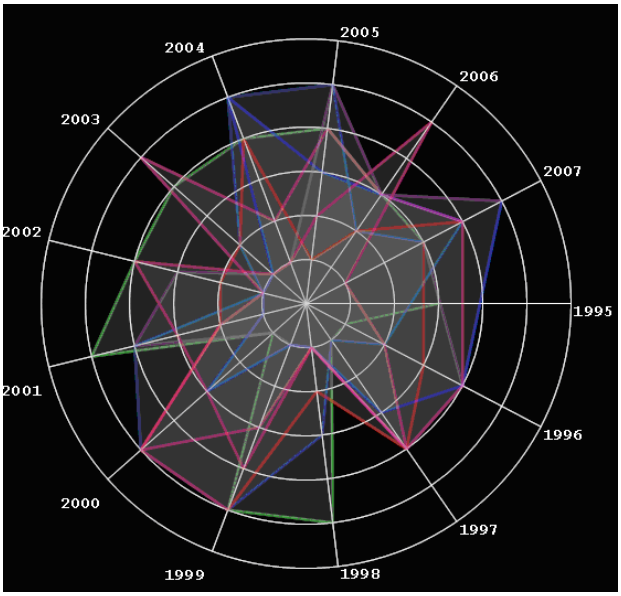

Figure 6: Citation received for a knowledge domain

3.2 Knowledge Structure

The skeleton of an individual has emerged; we used a 3D orbit view to present the knowledge structure of an individual. The center is a knowledge core with textures attached as shown in Figure.8. We had used a most appropriate texture to represent each RFCD. The knowledge core in the center of the 3D graph reveals the knowledge combinations of an author by corresponding textures.

Figure 8 presents a knowledge structure of an individual derived from the analysis in Section.2. Each node in the graph represents a publication which is a portion of knowledge of an author that he or she possesses and the coloring was used to differentiate the RFCD in order to identify the node for which knowledge that it belongs to. The color for each RFCD was not pre-defined instead it has been assigned randomly to each newly identified RFCD. The size of node represents the

citation count received over time and their distance towards the knowledge core reveals the quality.

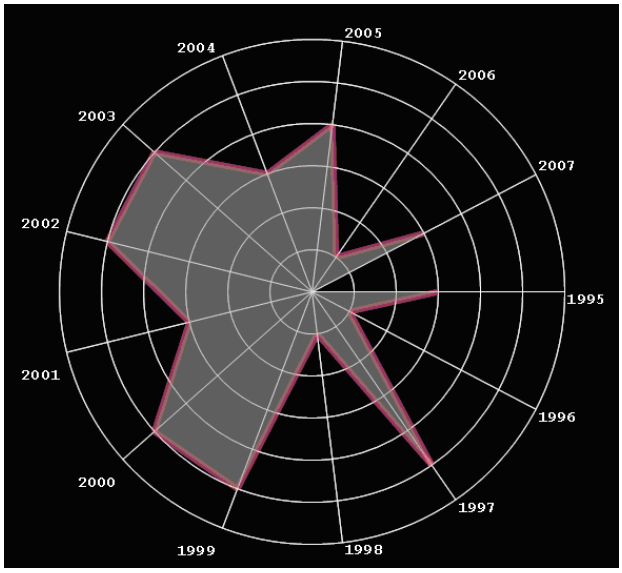

Figure 7: Citation received of a document

The node closes to the centre or knowledge core suggests the publication has received the widely acceptance by citations. It also suggests that the author has the strong understanding of the knowledge in that domain. However, the size of the nodes close to the knowledge core might not be relative larger than other nodes scattered around the graph as we have implemented the quality fairness by taken the quality of cited papers by other authors into consideration. For example, if a publication has been cited by other authors of whom publication has been published to influential venues or journals then a stronger weight would be assigned as a result of stronger weight a node's coordinate would be placed near the center as shown in Figure 8 which presents an author possesses knowledge structure of four areas of research knowledge domains by analyzing the publications and mapped into corresponding RFCDs.

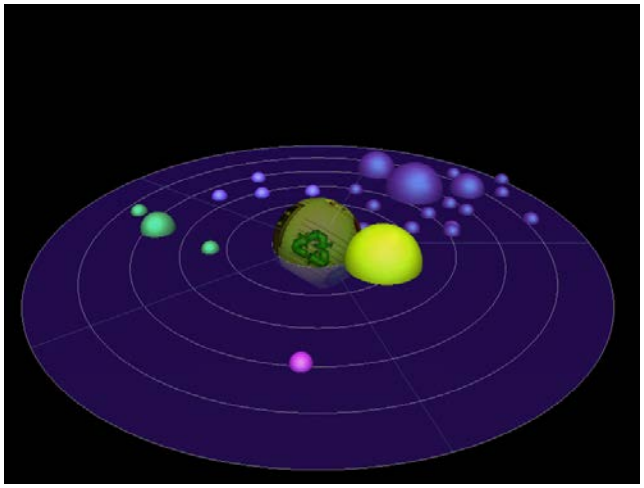

Figure 8: 3D Knowledge Structure

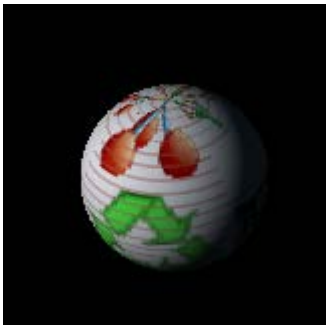

Figure 9: Knowledge Core

4. Conclusion

We aimed to apply the visualization techniques, information retrieval and mapping approaches developed to the context of visualizing knowledge structure of research expertise. The study of identifying knowledge experts and important papers at knowledge domain level had been widely discussed and those serve as foundation of our work. However, instead of exploring general level knowledge structures, we focusing on researching the knowledge structure at individual level by analyzing and grouping publications into well defined knowledge domains.

The research result also suggests the potential of applying information indexing and retrieval in the field of knowledge visualization. We have employed such approach to map the document into well defined RFCD that each code in fact represents a classified knowledge domain. It automates the process of mapping the document semantics to pre-defined domain definitions. However, the pre-requisite is that the domain definition must be pre-defined for a document to know which domain that it should map to.

In the research, we have contributed to the visualization of individual knowledge domains and mapping of document into knowledge domain using RFID code by applying various semantic analysis techniques. In the meantime, we are also investigating the feasibility of extending the research into visualizing the entire knowledge structure of a domain.

References

[1] C. Chen and R. J. Paul, “Visualizing a Knowledge Domain's Intellectual Structure”, IEEE Comput. 34(3), pp65-71, March, 2001.

[2] M. Porter. An algorithm for suffix stripping. Program, 14(3):130-137, 1980.

[3] Nieminen P, Carpenter J, Rucker G, Schumacher M, “The relationship between quality of research and citation frequency”, BMC Medial Research Methodology 2006, 6:42

[4] RFCD - http://www.arc.gov.au/apply_grants/rfcd_codes.htm

[5] T. Opthof, “Sense and nonsense about the impact factor”, Cardiovasc Res 1997, 33(1), pp1-7

[6] Berry, M. W, “Large-scale sparse singular value computations” Int. J. Supercomputer, vol. 6, no. 1, pp. 13–49, 1992

[7] W. Ke, K. Borner, and L. Viswanath, “Major Information Visualization Authors, Papers and Topics in the ACM Library”, IEEE Symposium on Information Visualization (INFOVIS'04).

[8] C. Chen and L. Carr, “A Semantic-Centric Approach to Information Visualization”, International Conference on Information Visualization, pp 18-23, 1999

[9] T.K. Landauer, P.W. Foltz and D. Laham, “Introduction to Latent Semantic Analysis”, Discourse Processes, 25, pp259-284, 1998

[10] S. Wasserman and K. Faust, “Social Network Analysis “,Cambridge University Press, Cambridge, 1994

[11] M. Newman, “Scientific Collaboration Networks: I. Network Construction and Fundamental Results”, Physical Review E, 64(1):016131, 2001.

[12] Yoshikane. Fuyuki, Nozawa. Takayuki and Tsuji. Keita, ”Comparative Analysis of Co-authorship Networks Considering Authors' Roles in Collaboration: Differences between the Theoretical and Application Areas”, ISSI 2005, July, 2005, vol.2, p.509-516.

[13] C. Cotta, J.J. Merelo, “The Complex Network of Evolutionary Computation Authors: an Initial Study”, Physics/0507196, 2005

[14] E. G. Berkowitz and M. R. Elkhadiri, “Creation of a Style Independent Intelligent Autonomous Citation Indexer to Support Academic Research”, Proceedings 15th Midwest Artificial Intelligence and Cognitive Science Conference, pp. 68-73, 2004

[15] A. Goldenberg and A. Moore, “Bayes Net Graphs to Understand Coauthorship Networks KDD”, Workshop on Link Discovery: Issues, Approaches and Applications, 2005

[16] B. Aleman-Meza, M. Nagarajan, C. Ramakrishnan, A. Sheth, B. Arpinar, L. Ding, P. Kolari, Anupam Joshi, and T. Finin, “Semantic Analytics on Social Networks: Experiences in Addressing the Problem of Conflict of Interest Detection”, WWW 2006, Edinburgh, Scotland

[17] P.A. Chirita, A. Damian, W. Nejdl and W. Siberski, “Search Strategies for Scientific Collaboration Networks”, CIKM 2005, Bremen, Germany

[18] B. Lee, M. Czerwinski, G. Robertson, B. B. Bederson, “Understanding Eight Years of InfoVis Conferences Using PaperLens”, INFOVIS'04, Vol 0, pp216.3

[19] Z. Huang, H.C Chen, F. Guo, J.J Xu, S. Wu, W.H Chen, "Visualizing the Expertise Space," *hicss*, p. 10038b, Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 1, 2004.