Metrics Based Research Assessment and Evaluation

Proceedings of the National Workshop on Using Different Metrics for Assessing Research Productivity

Indian Statistical Institute, New Delhi, 16-17 February, 2012

Editors
S.M. Dhawan
N.K. Khatri
A. Ratnakar

ISBN: 978-93-5067-501-4

Mapping the Structure and Development of Science Using Co-citation Analysis

Ganesh Surwase, B. S. Kademani and K. Bhanumurthy
Scientific Information Resource Division,
Bhabha Atomic Research Centre, Trombay, Mumbai-400085 (India).

Abstract

Co-citation analysis is a unique method used for studying the cognitive structure of science and assessing the research productivity. It is a research tool for examining the intellectual development and structure of the scientific discipline. This paper illustrates principles, techniques and applications of co-citation analysis. It also introduces the newly emerging co-citation analysis softwares, especially SciVal Spotlight and CiteSpace. Co-citation analysis is based on grouping together the papers that are frequently cited in pairs. Combined with single-link clustering and multidimensional scaling techniques, co-citation analysis can literally map the structure of specialized research areas as well as science as a whole.

Keywords: Co-citation analysis, co-citation matrix, co-citation network, bibliographic coupling, knowledge mapping, SciVal Spotlight, CiteSpace.

Introduction

In this age of information explosion, many attempts have since been made for the bibliographic control. In the beginning the bibliographies were published so as to save the time and money required for literature search and subsequently to acquire the resources. In 1955 concept of citation indexing was introduced by Eugene Garfield and many citation databases (SCI, INSPEC, CA, BA) came into existence. Development of world wide web in 1990s, enabled publishers to host the full-text databases (ScienceDirect, INIS-NCL) of research/journal articles on the Internet.

The citation databases established the forward and backward citation linkages among the research papers. With the help of bibliographies and citation databases the quantitative growth and development of science can be measured. The citation techniques viz. bibliographical coupling and co-citation analysis paved a way for mapping the structure of scientific disciplines.

Growth and Development of Scientific Information

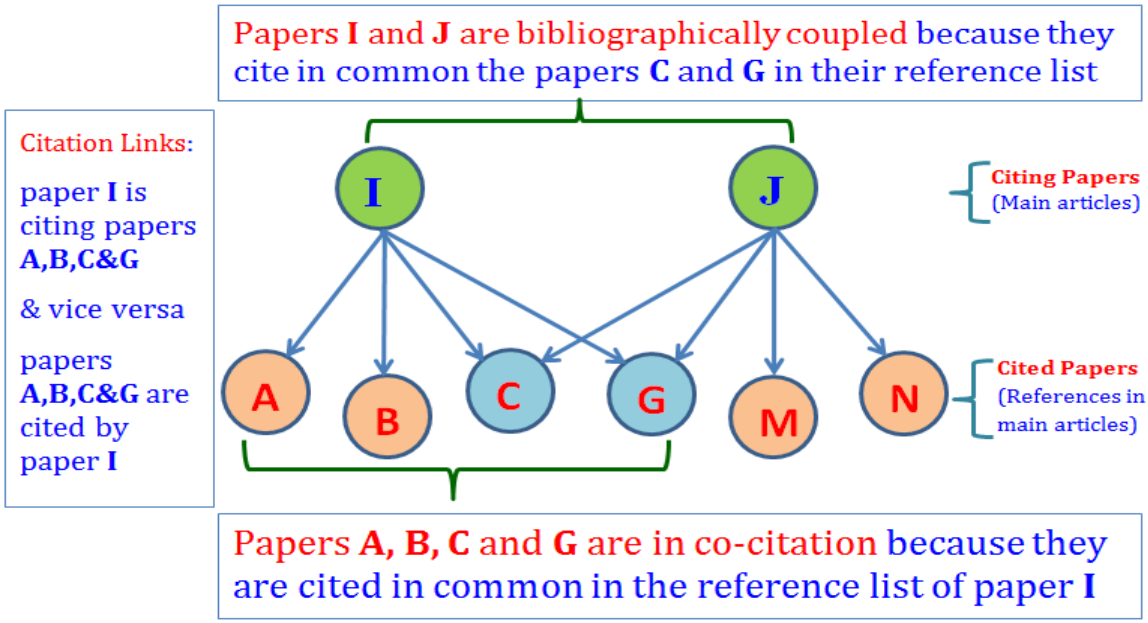
Egghe (2005) described informetrics as a broad term comprising all-metrics studies related to information science, including bibliometrics, scientometrics and webometrics. It deals with the quantitative aspects of information, including production, dissemination and use of all forms of information. With the help of bibliographies and citation databases, we can measure the growth and

development of science in general, while bibliographical coupling and co-citation analysis enables us to map the evolutionary structure of scientific disciplines.

Mapping the Structure of Science

Evolutionary structure of science can be mapped by establishing the relatedness among the research papers/topics based on links connecting individual documents in two ways: direct (backward and forward) citation and indirect (bibliographic or co-citation coupling) citation (Surwase, Sagar, Kademani & Bhanumurthy 2011). Figure-1 illustrates the concept of citation, co-citation and bibliographic coupling.

Figure-1: Concept of citation, co-citation and bibliographic coupling



Bibliographic coupling (shared references) and co-citation (shared citations) are association measures based on citation analysis. Bibliographic coupling is a fixed and retrospective in nature while co-citation analysis is dynamic and forward looking. Bibliographic coupling and co-citation based mapping techniques are more popular and accurate than direct citation (Boyack & Klavans 2010).

Bibliographic Coupling

Kessler (1963) introduced the concept of bibliographic coupling, demonstrated the usefulness of the phenomenon and argued for its application as an indicator of subject relatedness. Documents are said to be bibliographically coupled if they share one or more bibliographic references. Figure-2 illustrates the concept of bibliographic coupling.

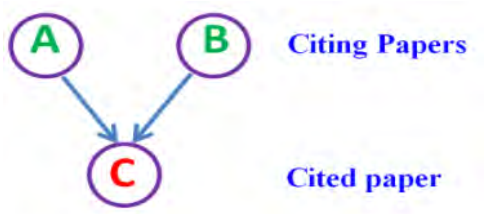


Figure-2: Papers A and B both cite paper C, hence paper A and B are bibliographically coupled, even they do not directly cite each other.

Formula for bibliographic coupling matrix:

$$Bab = \sum_{k=1}^{n} Lak \; Lbk$$

Where, B = Bibliographic coupling matrix
Bab = #common refs. in paper 'a' and 'b'
K = Refs. in paper 'a' and 'b' respectively
L = Matrix of refs. in paper 'a' and 'b'

Bibliographic coupling is used to extrapolate subject similarities among the two documents. The technique of bibliographic coupling is implemented to establish document relatedness in various databases like 'Web of Science', 'CiteSeer.IST' and 'The Collection of Computer Science Bibliographies'.

Co-citation Coupling

Small (1973) and Marshakova (1973) independently proposed co-citation-coupling as a variation of bibliographic coupling, to measure relatedness of documents by their co-citation frequency. The two documents are said to be co-cited if they appear simultaneously in the reference list of a third document. Figure-3 illustrates the concept of co-citation coupling.

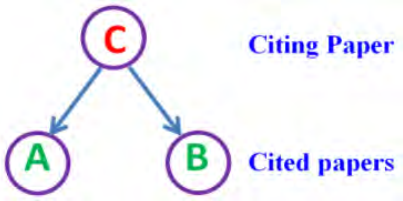


Figure-3: Papers A and B are cited by paper C, hence papers A and B are in co-citation, even they do not directly cite each other.

Formula for co-citation matrix:

$$Cab = \sum_{k=1}^{n} Lka \; Lkb$$

Where, C = Co-citation matrix
Cab = #papers that cite 'a' and 'b' in pair
K = papers citing 'a' and 'b' respectively
L = matrix of citations to paper 'a' and'b'

The co-citation frequency is defined as the frequency with which two documents are cited together in subsequent literature. When the same pairs of papersare co-cited by many authors, clusters of research begin to form. Co-cited papers in clusters tend to share some common theme (Small & Garfield 1985, 1993). The co-citation frequency count is directly proportional to the strength of co-citation coupling.

Co-citation Analysis

Citation of a document reflects the merit (quality, significance, impact) of that document (Smith 1981). Citation analysis is based on the premise that authors cite papers they consider to be important to substantiate and development of their research. As a result, heavily cited articles are likely to have exerted

a greater influence on the subject than those less frequently cited (Small 1974, 1977). Figure-4 illustrates the formation co-citation clusters.

Figure-4: Diagrammatic representation of co-citation analysis, co-citation matrix, co-citation frequency counts, co-citation links and cluster formation.

Co-citation analysis

Paper	Reference list
I	A
	B
	C
	G

Paper	Reference list
J	C
	G
	M
	N

Co-citation matrix

	A	B	C	G
A	AA	AB	AC	AG
B	BA	BB	BC	BG
C	CA	CB	CC	CG
G	GA	GB	GC	GG

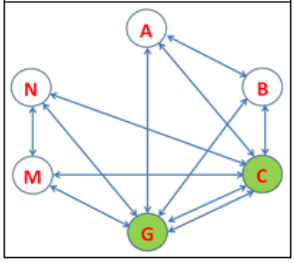
Co-citation matrix

	C	G	M	N
C	CC	CG	CM	CN
G	GC	GG	GM	GN
M	MC	MG	MM	MN
N	NC	NG	NM	NN

Pairs of papers	Co-citation count
AB	1
AC	1
AG	1
BC	1
BG	1
CG	2
CM	1
CN	1
GM	1
GN	1
MN	1
Total	12

Paper	Co-citation links
A	3
B	3
C	6
G	6
M	3
N	3
Total	24

Co-citation linkages



Current papers from publication window (institute) are assigned to the co-citation clusters based on their references.

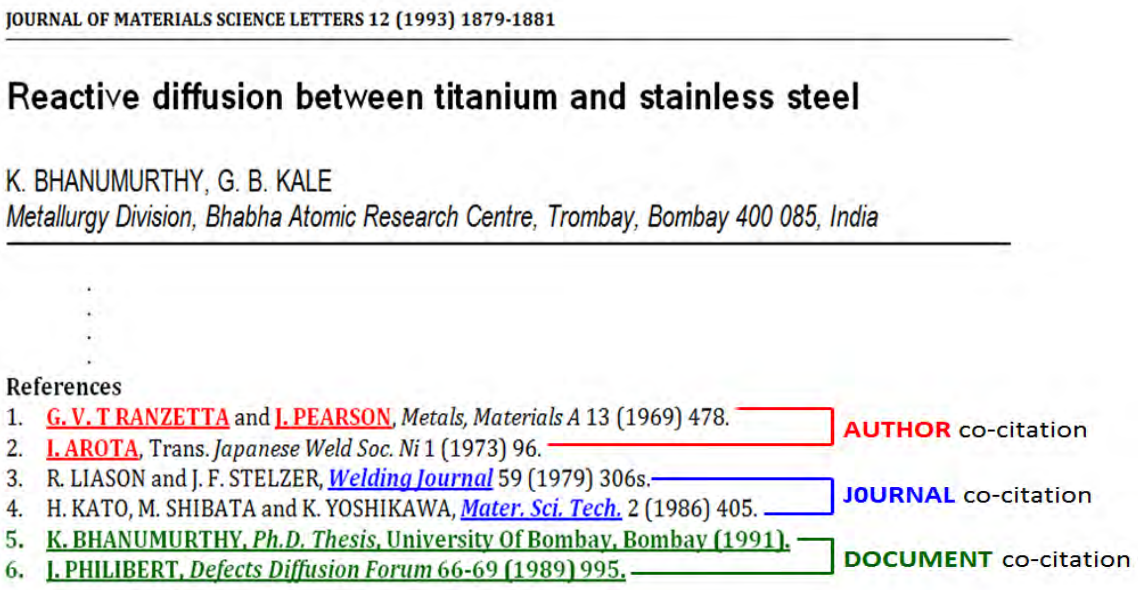
Co-citation analysis is based on grouping together the papers that are frequently cited in pairs. Intellectual linkages are established among the co-cited papers (Leydesdorff & Vaughan 2006). Combined with single-link clustering and multidimensional scaling techniques, co-citation analysis can literally map the structure of specialized research areas as well as science as a whole (Small & Garfield 1985). Co-citation analysis is a unique method for studying the cognitive structure of science.

Types of Co-citation Analysis

The unit of co-citation analysis determines the type of co-citation analysis. Author co-citation analysis, document co-citation analysis and journal co-citation analysis are the major types of co-citation analysis. Co-citation analysis has been used to map the topical relatedness of clusters of authors, journals or articles. Author co-citation analysis (White & Griffith 1981) aims to identify the groups of authors who were cited together in relevant literature and focuses on a network of cited authors connected by co-citation links. Author co-citation analysis has been used in analyzing the intellectual structure of science (Culnan 1987). Journal co-citation is of interest to the collection manager concerned with developing core journal lists, selecting journals and evaluating collections that serve particular research-oriented constituencies (McCain 1991). Document co-citation analysis studies a network of co-cited references. The assumption underlying document co-citation analysis is that too often co-cited documents are related

to one another, and address the same broad research questions, but without necessarily agreeing with each other. These cited documents serve as symbols for scientific ideas, methods, and experiments (Chen, Ibekwe-SanJuan & Hou 2010; Steven 2002).

Figure-5: Different types of co-citation links



Depending on the unit of co-citation analysis and thresholds, both the macro-structure and micro-structure of science can be mapped. Figure-5 provides the examples of various types of co-citation links. Using these various types of co-citation links, one can perform co-citation analysis.

Co-citation and Correlation Matrix

Co-citation clusters can be generated by using a modified cosine index based on co-citation counts for similarity, and to run the resulting matrix of cosine values through a visualization program which would assign each reference paper an (x, y) position on a 2-D plane. Formula for measuring the relatedness (Cosine coefficient) between two papers 'I' and 'J' (Chen, Ibekwe-SanJuan & Hou 2010; Leydesdorff 2005) is:

$$W_{ij} = \frac{|x \cap y|}{\sqrt{|x| * |y|}}$$

where,

W_{ij} = cosine coefficient for paper **i** and **j**
$|x|$ = set of citations to paper **i**
$|y|$ = set of citations to paper **j**
$|x \cap y|$ = set of citations to papers **i** and **j**

Multidimensional scaling is a class of techniques that uses proximities among any kind of objects as input. The chief output is a spatial representation, consisting of a geometric configuration of points (objects) on a map. The Minkowski distance metric provides a general way to specify distance in a multidimensional space (Steyvers 2002):

$$d_{ij} = \left[\sum_{k=1}^{n} \left|x_{ik} - x_{jk}\right|^{r}\right]^{1/r}$$

where,

d_{ij} = proximity measure between papers **i** and **j**
n = number of dimensions
x_{ik} = value of dimension **k** for paper **i**
x_{jk} = value of dimension **k** for paper **j**
r = 2

A Euclidian metric is appropriate when the stimuli are composed of integral. In practice, the Euclidian distance metric is often used because of mathematical convenience in multidimensional scaling procedures.

Co-citation Analysis Steps

Co-citation analysis has a fairly consistent sequence of steps, regardless of whether the objects of study are articles, authors or journals:

- Selection of the core set of items for the study (ACA/DCA/JCA).
- Retrieval of co-citation frequency information for the core set.
- Compilation of the raw co-citation frequency matrix.
- Correlation analysis to convert the raw frequencies into correlation coefficients.
- Multivariate analysis of the correlation matrix, using principle components analysis, cluster analysis or multidimensional scaling techniques.

Combined with single-link clustering and multidimensional scaling techniques, co-citation analysis can literally map the structure of particular topic of research or science in general (Small & Garfield 1985). The results of co-citation analysis can be presented in various formats such as tables, charts, graphs etc. as per characteristics of the data and objectives of the study.

Case Study-1: Scival Spotlight

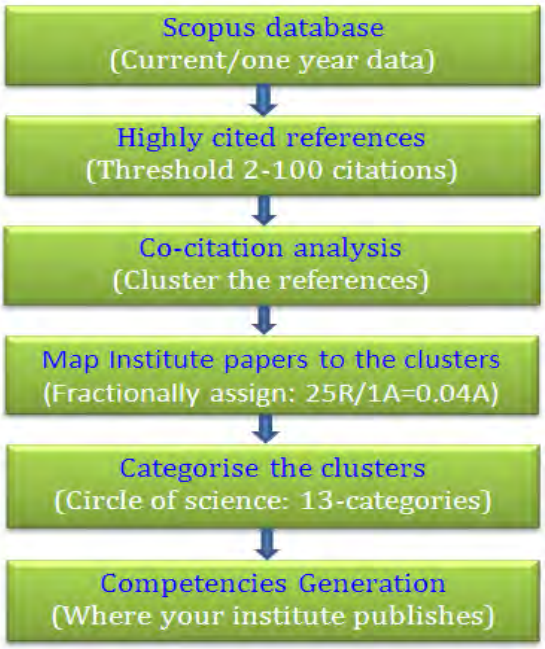
SciVal Spotlight is Internet-based proprietary software brought out by Elsevier Publishers. It reveals the fine structure of science using co-citation analysis and visualisation algorithms. This method adopts a reference based classification, and gives an alternative methodology for identifying science leaderships in three ways (Elsevier 2009, 2010):

- Publication Leadership: The institution with the greatest relative article share (RAS) exhibits publication leadership. RAS means how many articles within each competency are from your institute relative to the institution ranked #1 in that competency.

- Reference Leadership: The institution with the greatest relative reference share (RRS) exhibits reference leadership. RRS is the number of highly-cited reference articles authored by the institution divided by the closest competitor.
- Thought leadership: The institution with the greatest State of the Art (SotA) value exhibits thought leadership. SotA value (in years) indicates the recentness of articles cited in the institution's papers compared to the average recency within the competencies.

The diagrammatic representation of steps involved in SciVal Spotlight is given in Figure-6. To identify the competencies for an institute, one year Scopus database is selected to create a mini database of highly cited references in that year. Co-citation network of these highly cited references is created. Based on the references, papers from your institute are mapped to the clusters of highly cited references network. Competencies are generated where your institute is leader or tends to be the leader.

Figure-6: Steps involved in generating the maps in SciVal Spotlight



The assignment of a current paper to more than one reference cluster creates two critical statistics, first an indication of the size of the cluster and another cluster: cluster relatedness measure. The results are visualised as circle of science and various graphs and statistical reports are generated.

Case Study-2: CiteSpace

CiteSpace is a free co-citation analysis software created by Chaomei Chen, that users can download to visualize and analyze trends in scientific literature. The primary source of input data for CiteSpace is the Web of Science, and CiteSpace also provides simple interfaces for obtaining data from PubMed, arXiv, Astrophysics Data System (ADS), and National Science Foundation (NSF) Award Abstracts. Search records from Derwent World Patents Index (DWPI) can also be visualized in CiteSpace

(Chen 2004, 2006). Table 1 and 2 illustrate the record (text file) input to the CiteSpace and Figures 7 and 8 give an output i.e. maps generated by CiteSpace software.

Table-1: Input data for co-citation analysis using CiteSpace

Cited Paper	Citing Paper	References in Citing Paper
Bhabha HJ, Daniel RR, 1948, Meson scattering with nuclear excitation, ***Nature,*** 161 (4101): 883-884.	Sinha MS, 1949, Photograph of a shower produced by a pi-meson and a pi-mu-decay, ***Physical Review,*** 75 (11): 1757-1759.	Bhabha HJ, 1948, ***Nature***, 161:883
		Occhialini GPS, 1948, ***Nature***, 162:168
		Lattes, 1947, ***Nature***, 160:486
		Rochester GD, 1947, ***Nature***, 160:855
		Heitler W, 1942, ***P Camb Philos Soc***, 38:296
		Williams EJ, 1939, ***Proc R Soc Lon-A***, 169:531

Table-2: Author Co-citation Matrix

AUTHOR	Bhabha HJ	Heitler W	Lattes	Occhialini GPS	Rochester GD	WilliamsEJ
Bhabha HJ						
Heitler W	HeitlerW--BhabhaHJ					
Lattes	Lattes--BhabhaHJ	Lattes--HeitlerW				
Occhialini GPS	OcchialiniGPS--BhabhaHJ	OcchialiniGPS--HeitlerW	OcchialiniGPS--Lattes			
Rochester GD	RochesterGD--BhabhaHJ	RochesterGD--HeitlerW	RochesterGD--Lattes	RochesterGD--OcchialiniGPS		
Williams EJ	WilliamsEJ--BhabhaHJ	WilliamsEJ--HeitlerW	WilliamsEJ--Lattes	WilliamsEJ--OcchialiniGPS	WilliamsEJ--RochesterGD	

Since each object is identical to itself, one half of the co-citation matrix, usually the lower triangle below the diagonal is used. Metrics such as burstness, centrality, modularity, and silhouette metrics described above are implemented and incorporated in various visualizations in CiteSpace. For example, purple rims of nodes indicate the importance of nodes in terms of betweenness centrality (≥0.1). Three (Cosine, Dice, and Jaccard) types of link similarity measures are supported in CiteSpace (Chen, Ibekwe-SanJuan & Hou 2010).

Figure-7: Document co-citation analysis map generated by using CiteSpace

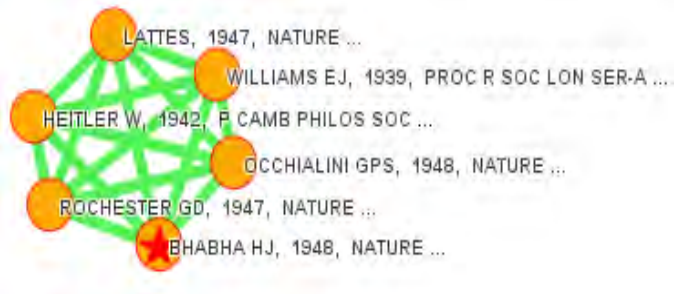


Figure-8: Author co-citation analysis map generated by using CiteSpace

Interactive functions in CiteSpace correspond to three levels of units of analysis. At the network level, functions operate on networks, including global visualizations of networks: a node-and-link cluster view and a timeline view. At the cluster level, functions operate on individual clusters such as showing all the citers to a cluster or hiding a cluster. At the basic entity level, functions are restricted to individual entities, for example, showing the citation history of a cited reference.

Steps for visualizing the information with CiteSpace (Chen, Ibekwe-SanJuan & Hou 2010):

- **Collect Data**- The primary source for data is Web of Science, and default input data format is ISI Export Format.
- **Create a Project**- Consists of two directories i.e. input data files and files generated by CiteSpace for analysis and visualization.
- **Adjust Parameters**- Change time slicing, node types, term sources, term selection, links, pruning, and visualization options.
- **Generate Visualizations**- Available visualizations include Cluster View, Time-Zone View, Show Networks by Time Slices, and Show Merged Networks.
- **Explore Visualizations**- depends upon the nature of data and objective of the study.
- **Generate Clusters**- CiteSpace uses a spectral clustering algorithm to decompose a network, and the resultant clusters are mutually exclusive (one item to one cluster).
- **Generate Cluster Labels**- sources for the label terms are title, abstract or index.

CiteSpace applies progressive network analysis, and focuses on nodes that play critical roles in the evolution of a network over the period of time. Such critical nodes are candidates of intellectual turning points.

Conclusion

The study demonstrates how co-citation analysis can be exploited to identify and map the historical structure of science, core sets of articles, authors, or journals in a given field of study. Commercial and open source software tools currently available in the market have further simplified the task of co-citation analysis. Not only do such software help us to identify science leadership, but also help us visualize the current picture of science on two/three dimensional scale, discover intellectual structure of science, and reveal important clues to the developments expected in science. Invariably co-citation analysis is done at two levels: document level and author level. Document co-citation analysis (DCA) reveal research patterns in science better than author co-citation analysis (ACA) does. Secondly, co-citation based applications (SciVal Spotlight) use reference based classification and deal subject classification of journals at topic level. This approach to journal classification is better than the classification approach that citation databases (WoS and Scopus) apply to journals, since citation-databases-based journal classification is too general or vague.

Co-citation analysis requires forming clusters. However, assigning labels to clusters is the most challenging task, more so because sometimes labels to clusters donot represent real reasons underlying their formation.

Co-citation analysis has demerits too, such as loss of relevant papers, inclusion of non-relevant papers, over-representation of theoretical articles, time lag between emergence of new specialties and capturing of them in a co-citation map, and subjectivity inherent in the setting of threshold levels. Some of these shortcomings can be overcome by addressing semantics of labels and using appropriate computing technology.

Acknowledgement

Authors are highly indebted to Dr. Chaomei Chen, Associate Professor, College of Information Science and Technology, Drexel University, USA and developer of CiteSpace, for his co-operation and guidance in the installation and use of CiteSpace software.

References

Boyack, K.W. and Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2389–2404.

Chen, C. (2004). Searching for intellectual turning points: progressive knowledge domain visualization, *Proc. Nat. Acad. Sci*., vol. 101(Suppl.), pp. 5303-5310.

Chen, C. (2006) CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature, *Journal of American Soc. Info. Sci. and Technology*, vol. 57, no. 3, pp. 359-377.

Chen, C, Ibekwe-SanJuan, F, Hou, J. (2010). The structure and dynamics of co-citation clusters: a multiple-perspective co-citation analysis, *Journal of American Soc. Info. Sci. and Technology*, vol. 61, no. 7, pp. 1386-1409.

Culnan, M.J. (1987). Mapping the intellectual structure of MIS, 1980-1985: a co-citation analysis, *MIS Quarterly*, vol. 11, no. 3, pp. 341-353.

Egghe, L. (2005). Expansion of the field of informetrics: origins and consequences, *Information Processing and Management*, vol. 41, no. 6, pp. 1311-1316.

Elsevier BV (2009). Co-citation analysis: the methodology of SciVal Spotlight, *(http://www.americalatina.elsevier.com/sul/pt-br/scival/pdf/Co-citation_Analysis_SciVal_Spotlight.pdf).*

Elsevier BV (2010). Identifying organizational inefficiencies in research institutions, (*http://www.info.scival.com/UserFiles/SciVal_Perspectives_Identifying_Organizational_Perspective.pdf*)

Kessler, M.M.1963, 'Bibliographic coupling between scientific papers', *American Documentation*, vol. 24, pp. 123-131.

Leydesdorff, L. (2005). Similarity measures, author co-citation analysis, and information theory, *Journal of the American Society for Information Science & Technology*, vol. 56, no. 7, pp. 769-772.

Leydesdorff, L. and Vaughan, L. (2006). 'Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment, *Journal of the American Society for Information Science and Technology*, vol. 57, no. 12, pp. 1616–1628.

Marshakova, I.V. (1973). A system of document connection based on references, *Scientific and Technical Information Serial of VINITI, vo.* 6, no. 2, pp. 3-8.

McCain, K.W. (1991). Mapping economics through the journal literature: an experiment in journal co-citation analysis, *Journal of the American Society for Information Science*, vol. 42, no. 4, pp. 290-296.

Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents, *Journal of the American Society for Information Science* (JASIS), vol. 24, no. 4, pp. 265-269.

Small, H. (1974). Co-citation in the scientific literature: a new measure of the relationship between two documents, in: *Essays of an Information Scientist,* vol. 2, pp. 28-31.

Small, H. (1977). A co-citation model of a scientific specialty: a longitudinal study of collagen research, *Social Studies of Science,* vol. 7, pp. 139-166.

Small, H. and Garfield, E. (1985). The geography of science: disciplinary and national mappings, *Journal of Information Science*, vol. 11, pp. 147-59.

Small, H. and Garfield, E. (1993). Co-citation analysis of science: Henry Small on mapping the collective mind of science, *Current Comments*, May 10, 1993.

Smith, L.C. (1981). Citation analysis, *Library Trends,* pp. 83-106.

Steven, N, Chu, CH, Raghavan, V. (2002). Visualization of document co-citation counts, in: *Proceedings of the 6th International Conference on Information Visualization*, London, England, July 2002.

Steyvers, M. (2002). Multidimensional scaling, in: *Encyclopedia of Cognitive Science.* Macmillan Reference Ltd.

Surwase, G., Sagar, A., Kademani, B.S., and Bhanumurthy, K. (2011). Co-citation analysis: an overview, *National Conference on Beyond Librarianship* (NCBL-2011). September 16-17, 2011, Mumbai (India), pp. 196-206.

White, H.D. and Griffith, B.C. (1981). Author co-citation: a literature measure of intellectual structure, *Journal of the American Society for Information Science* (JASIS), vol. 32, no. 3, pp. 163-171.

---X---