

CiteSpace II: Visualization and Knowledge Discovery in Bibliographic Databases

Marie B. Synnestvedt MEd,^{1,2} Chaomei Chen PhD,² John H. Holmes PhD¹

¹Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, Philadelphia PA

²College of Information Science and Technology, Drexel University, Philadelphia PA

Abstract. This article presents a description and case study of CiteSpace II, a Java application which supports visual exploration with knowledge discovery in bibliographic databases. Highly cited and pivotal documents, areas of specialization within a knowledge domain, and emergence of research topics are visually mapped through a progressive knowledge domain visualization approach to detecting and visualizing trends and patterns in scientific literature. The test case in this study is progressive knowledge domain visualization of the field of medical informatics. Datasets based on publications from twelve journals in the medical informatics field covering the time period from 1964-2004 were extracted from PubMed and Web of Science (WOS) and developed as testbeds for evaluation of the CiteSpace system. Two resulting document-term co-citation and MeSH term co-occurrence visualizations are qualitatively evaluated for identification of pivotal documents, areas of specialization, and research trends. Practical applications in bio-medical research settings are discussed.

INTRODUCTION

The scientific literature has been estimated to grow at a rate of 6% per year [1,2]. Record counts collected from the PubMed database shows a fifty-percent increase in the number of records indexed by year of publication over the past fifteen years (Figure 1). With this growth rate in scientific literature come ever increasing challenges for investigators and clinicians to become acquainted with the core literature of their field, conduct literature reviews, keep abreast of a field, and search for relevant documents. This growth of the literature is reflected in the concomitant growth in the size and complexity of bibliographic databases.

We feel that there are strong parallels between bibliographic databases and clinical data warehouses, and that citation data is suitable for a Knowledge Discovery in Databases (KDD) approach that uses specialized data mining tools. The KDD approach to data analysis is usually a retrospective analysis of

data and does not involve consideration of experimental design and related concepts [3]. KDD has been defined as the automated or convenient extraction of patterns representing knowledge explicitly stored in large databases, data warehouses, or other large repositories. The process of evaluating data, analyzing patterns, and extracting knowledge is analogous to the sorting, cleaning, and grading process involved in mining minerals [4]. The knowledge discovery process is applied to explain existing data, make predictions or classifications, or summarize contents of large databases to support decision making [5].

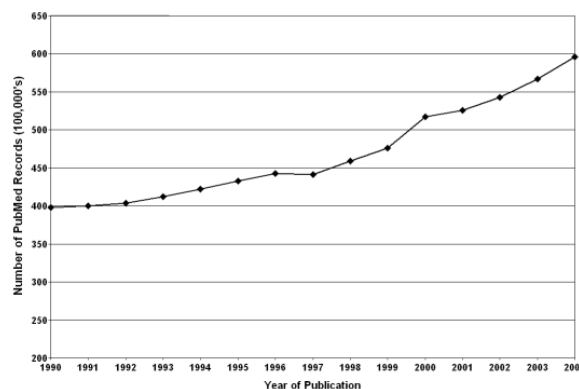


Figure 1. Number of PubMed Records by Year of Publication 1990 – 2004

THE CiteSpace II APPLICATION

This article presents a description and case study of CiteSpace II, a Java application which combines information visualization methods, bibliometrics, and data mining algorithms in an interactive visualization tool for extraction of patterns in citation data. A pilot study [6] of medical informatics applied document co-citation analysis (DCA) combined with Pathfinder Network Scaling (PFNET), visualization, and animation to develop a 3-dimensional (3-D) knowledge landscape to a limited dataset based on AMIA publications. Animated 3-D models vividly depicted the growth of the field, but they were cognitively demanding. CiteSpace II incorporates

substantial changes since our previous report. Due to space limitations, a brief summary of the theoretical and methodological basis on which CiteSpace II was developed is presented here. Detailed reports can be found in Chen, 2004 and Chen, 2005 [7, 8].

The primary goal of CiteSpace II is to facilitate the analysis of emerging trends in a *knowledge domain*. Knowledge domains are modeled and visualized as a time-variant duality between two fundamental concepts in information science – *research fronts* and *intellectual bases*. The concept of a research front was originally introduced by Price [1]. In a given field, a research front refers to the body of articles that scientists actively cite. Persson [9] made a distinction between a research front and an intellectual base (p. 31): “In bibliometric terms, the citing articles form a research front, and the cited articles constitute an intellectual base.”

New features of CiteSpace II are related to three central concepts: 1) Kleinberg’s *burst detection algorithm* is adapted to identify emergent research front concepts [10], 2) Freeman’s *betweenness centrality metric* is used to highlight potential pivotal points [11], and 3) heterogeneous networks. A knowledge domain is conceptualized as a mapping function between a research front and its intellectual base. This mapping function provides the basis of a conceptual framework to address three practical issues: 1) identifying the nature of a research front, 2) labeling a specialty, and 3) detecting emerging trends and abrupt changes in a timely manner. CiteSpace collects *n-grams*, or single words or phrases of up to four words, from titles, abstracts, descriptors, and identifiers of citing articles in a dataset. Research front terms are determined by the sharp growth rate of their frequencies. Two complementary views for analyzing and visualizing 2-D co-citation networks are designed and implemented: cluster views and time-zone views. The new methods in CiteSpace II have improved the clarity and interpretability of visualizations so as to reduce the user’s cognitive burden as they search for trends and pivotal points in a knowledge structure.

The CiteSpace II application has two major interface components. The first component is used for designating the data and analysis parameters, and is shown in Figure 2. The primary source of data for CiteSpace analysis is the Web of Science from which data must be downloaded prior to using CiteSpace. CiteSpace II also allows users to download citation data directly from PubMed. Research front terms are extracted by first running the Burst Detection option. Users specify the range of years to be analyzed a

time, the length of time slices within the time interval; and three sets of threshold levels for citation counts, co-citation counts, and co-citation coefficients (*c*, *cc*, *ccv*). The specified thresholds are applied to the earliest, middle, and last time slice. Linear interpolated thresholds are assigned to the rest of slices. Network pruning, merging, and layout options are also set by users. The second interface component allows users to interact with and manipulate the visualization of a knowledge domain in several ways. Visual attributes of the display as well as a variety of parameters used by the underlying layout algorithms can be adjusted. Figure 3 illustrates a zoomed view of an author co-citation cluster that has been marked with marquee selection, and the resulting display of associated MeSH headings and retrieval of related article abstracts from PubMed.

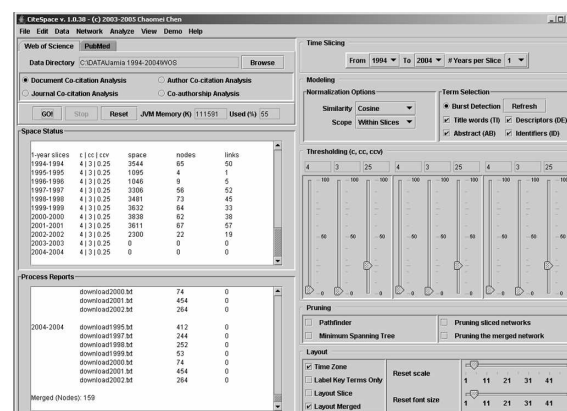


Figure 2. CiteSpace II Interface for Configuring Analysis

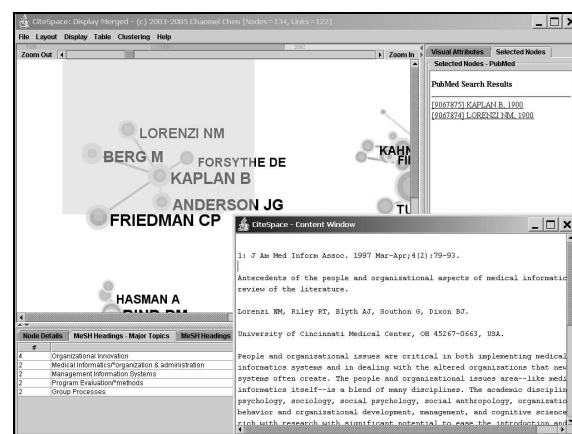


Figure 3. CiteSpace Visualization Interface.

METHODS

Two new datasets for analysis of Medical Informatics were developed as a testbed for CiteSpace II. The Institute for Scientific Information's (ISI) Journal Citation Reports list of medical informatics journals for 2003 was cross-referenced against a list of medical informatics journals from AMIA [12]. The twelve journals that both resources had identified as important or relevant to medical informatics were selected for study. These twelve journals were also checked against the NCBI journals database for publication history, and the journals which were predecessors of some of the current journals were identified. Citation data was exported from Web of

Science, and a query was run against the PubMed database from within CiteSpace. Because ISI has indexed meeting abstracts under journal names instead of conference proceeding names, meeting abstracts were excluded from the WOS data. This resulted in a WOS dataset of 11,952 citation records covering forty years from 1964-2004 and the closely equivalent time period and journals dataset of 13,369 records from PubMed (Table 1). The datasets cover a larger period of time than Morris and McCain's 1998 journal co-citation study, and match on nine of the twenty journals from that study which covered the indexing period January 1993-July 1995.

Table 1. Medical Informatics Datasets

ISI Full Journal Title	JCR 2003 Impact Factor	JCR 2003 I. F. Rank	Years Indexed in PubMed	Records In PubMed Dataset	Years Indexed in WOS	Records in WOS Dataset
Artificial Intelligence In Medicine	1.222	6	1993 -	491	1992 -	623
Cin-Computers Informatics Nursing (1)	0.217	19	1983 -	778	1992 -	249
Computer Methods And Programs In Biomedicine (2)	0.724	14	1971-	2122	1975 -	2063
IEEE Transactions On Information Technology In Biomedicine	1.274	5	1997 -	304	2000 -	210
International Journal Of Medical Informatics (3)	1.178	8	1970 -	1953	1975 -	1757
International Journal Of Technology Assessment In Health Care	0.754	12	1985-	1370	1995 -	742
Journal Of The American Medical Informatics Association (4)	2.51	1	1994-	736	1994 -	1674*
Journal Of Biomedical Informatics (5)	0.855	11	1967 -	1584	1968 -	1555
M D Computing	0.500	17	1984-2/2001	836	1984 – 02/2001	500*
Medical Decision Making	1.718	3	1981-	1164	1983 –	871*
Medical Informatics And The Internet In Medicine	0.915	10	1999 -	134	01/1999 -	136
Methods Of Information In Medicine	1.417	4	1965 -	1897	1964 -	1572*
Total			1965-2004	13369	1964-2004	11952

1: Continues Computers in Nursing; 2: Continues Computer Programs in Biomedicine; 3: Continues International Journal of Bio Medical Computing; 4: WOS has AMIA Symposium Proceedings 1994 – 2002 indexed as supplement to JAMIA; 5: Continues Computers and Biomedical Research; *: Meeting abstracts excluded.

RESULTS

Due to the limited space, only the major findings from two examples of the visualizations produced with CiteSpace II are described: a cluster view (Figure 4) and a time-zone view (Figure 5). Table 2 shows the visualization parameters, and the system used was a 1600MHz Pentium notebook with 1 GB RAM. The Burst Detection process completed running on each dataset in two to three minutes. The visualization in each figure was generated in less than one minute. The following interpretations by two of the authors of this article are based on their own experience and knowledge of medical informatics. The visualizations are qualitatively evaluated for identification of pivotal documents, areas of specialization, and research trends.

Table 2. Visualization Configuration and Metrics

View	Cluster (Figure 4)	Time-Zone (Figure 5)
Data Source	PubMed	WOS
Analysis Type	MeSH Term Co-occurrence	Document-Term Co-citation
Publication Years	2000-2004	1990-2004
Slice	1 year	5 years
Modeling	Cosine, within slices	Cosine, within slices
Thresholding (c/cc/ccv)	5/3/25	7/3/30
Pruning	Pathfinder	None
Layout	Merged	Time-Zone, Merged
Burst Terms	11,137	9,869
Document/Term Space*	9,066	136,469
Nodes & Links	151 & 148	212 & 279
Run Time (milliseconds)	35,961	42,581

*WOS data includes cited references

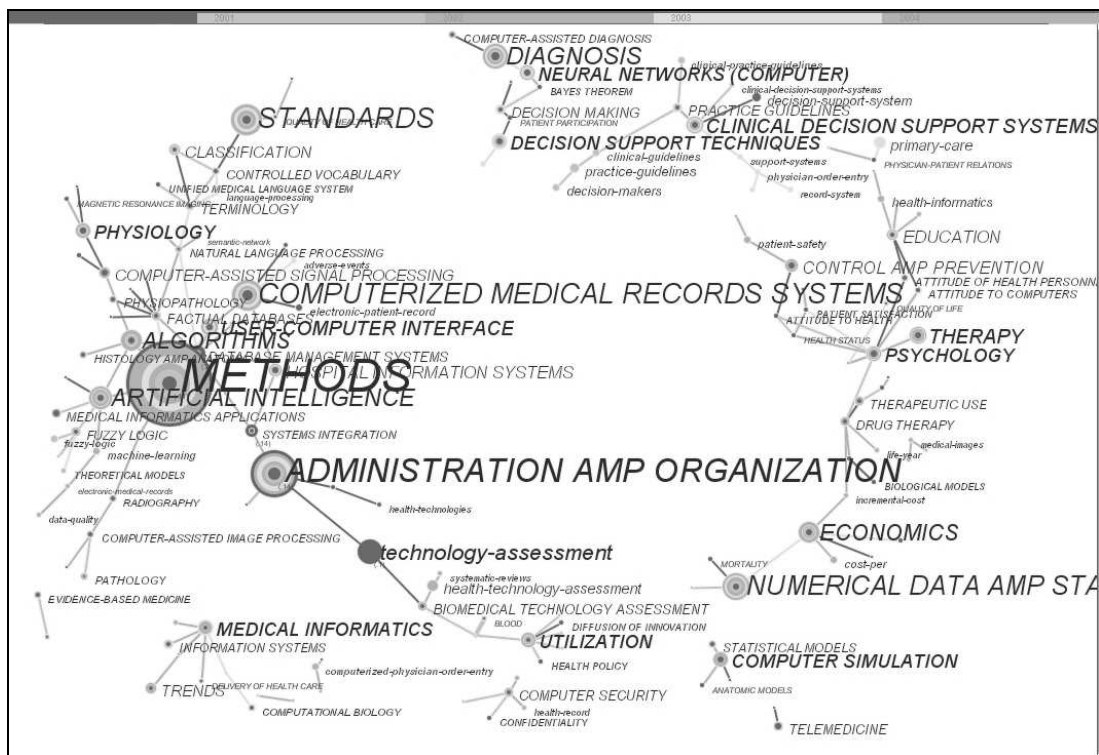


Figure 4. Cluster view of Medical Informatics 2000 - 2004.

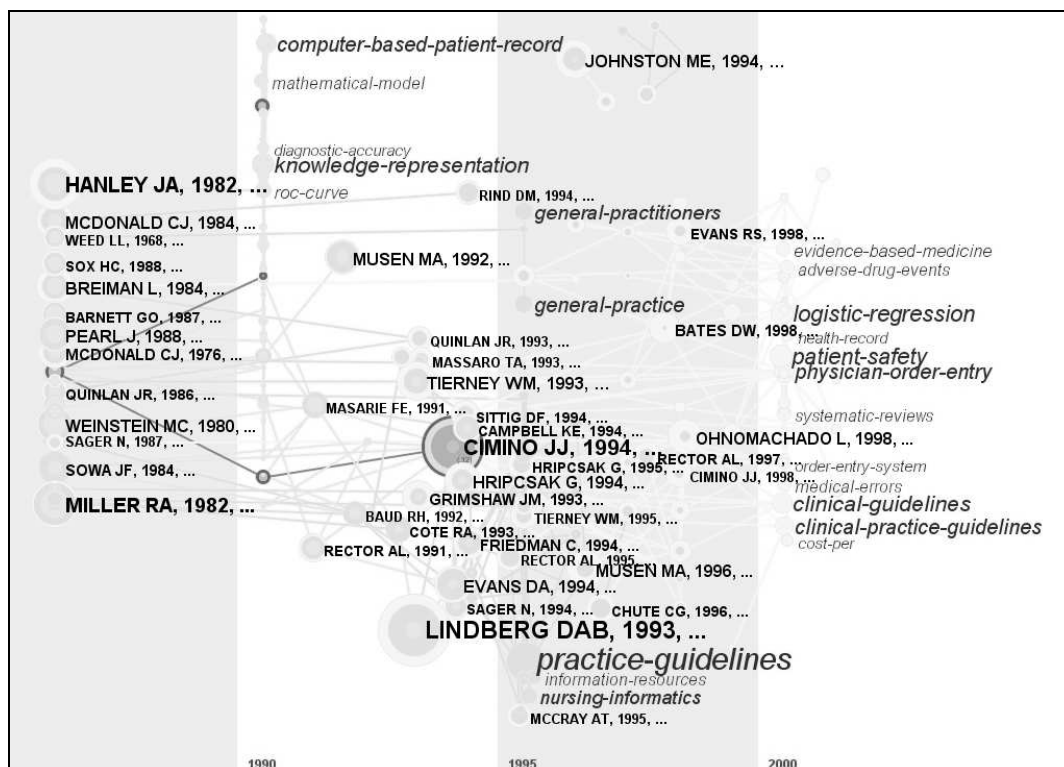


Figure 5. Time-zone view of Medical Informatics 1990 - 2004

The cluster view (Figure 4) provides an overview of research areas within the field of medical informatics during the years from 2000 to 2004. In this visualization the node size represents the overall frequency of occurrence of keyword terms and the colored rings of the nodes represents yearly time-slices. A trail of several pink rimmed nodes (those with a high measure of “betweenness centrality”) highlights a transition from the early decrease in “technology assessment” to the growth then decrease in “administration amp(&) organization” to the recent increase in the frequency of the term “methods”. In comparison to previous journal co-citation multidimensional scaling displays [13], the specialties are automatically labeled at the level of detail of MeSH headings and keyword terms as opposed to manual assignment of labels at the level of clusters of journals. This affords insight into the structure of a knowledge domain without requiring prior domain or journal knowledge, but does still require conceptualizing labels for clusters of terms. The time-slicing feature of CiteSpace also provides information on the relative activity of research areas within time periods.

The time-zone view (Figure 5) adds additional insights by mapping the highly cited and pivotal documents that constitute the knowledge base of medical informatics and the timing of emergence of new topics. Figure 5 depicts the evolution of themes that could be considered central to medical informatics research and practice over time. There are a number of particularly prominent themes, such as ROC curve analysis and decision making in the early 1990s, giving way to practice guidelines and patient safety by the turn of the century. Concomitantly, there is a shift in the centrality of certain authors, that largely parallels the focal areas, and this is to be expected.

DISCUSSION AND CONCLUSION

CiteSpace II is a system that could be potentially used by a wide range of users, notably scientists, clinicians, science policy researchers, and medical librarians. For example, clinical researchers would find CiteSpace II particularly useful in creating domain-specific ontologies for use in developing evidence-based knowledge bases for decision support. Information scientists and librarians would find it indispensable for tracking the growth of new areas, virtually in real-time, which in turn could aid in collection development. However, there are several limitations to using CiteSpace II, the most important of which is the learning curve required to set accurate visualization parameters. In addition,

some maps and clusters may be highly complex, requiring specialized domain knowledge for interpretation. Even with these limitations in mind, CiteSpace II should prove to be a very valuable tool for a variety of users.

Notes. CiteSpace II is available for download from: <http://cluster.cis.drexel.edu/~cchen/citespace>.

References

1. Price, DD. Networks of scientific papers. *Science*. 1965 Jul 30;149:510-5
2. Fernández-Cano A, Torralbo M, Vallejo M. Reconsidering Price's model of scientific growth: An overview. *Scientometrics*. 2004 Jan; 61(3):301 – 321.
3. Smyth P. Data mining: data analysis on a grand scale? *Stat Methods Med Res*. 2000 Aug;9(4):309-27. Review.
4. Han J, Kamber J. *Data Mining: Concepts and techniques*. San Francisco:Morgan Kaufmann Publishers; 2001.
5. Babic A. Knowledge discovery for advanced clinical data management and analysis. *Stud Health Technol Inform*. 1999;68:409-13. Review.
6. Synnestvedt M, Chen C. Visualizing AMIA : a medical informatics knowledge domain analysis. *AMIA Annu Symp Proc*. 2003;:1024.
7. Chen C. Searching for intellectual turning points: progressive knowledge domain visualization. *Proc Natl Acad Sci U S A*. 2004 Apr 6;101 Suppl 1:5303-10.
8. Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JASIST*. In press 2005.
9. Persson O. The intellectual base and research fronts of Jasis 1986-1990. *JASIST*. 1994; 45(1):31-38.
10. Kleinberg J. (2002). Bursty and hierarchical structure in streams. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2002;:91-101.
11. Freeman LC. Centrality in social networks: Conceptual clarification. *Social Networks*. 1979; 1:215-239.
12. American Medical Informatics Association [homepage on the Internet]. Resource Center - Publications of Interest - Journals. [updated 2003 Jan 1; cited 2005 Mar 16]. Available from: <http://www.amia.org/resource/pubs/f3.html>
13. Morris TA, McCain KW. The structure of medical informatics journal literature. *J Am Med Inform Assoc*. 1998 Sep-Oct;5(5):448-66.