Chapter 2

MAPPING MEDICAL INFORMATICS RESEARCH

Shauna Eggers[1], Zan Huang[1], Hsinchun Chen[1], Lijun Yan[1], Cathy Larson[1], Asraa Rashid[1], Michael Chau[2], and Chienting Lin[3]

[1]Artificial Intelligence Lab, Department of Management Information Systems, Eller College of Management, The University of Arizona, Tucson, Arizona 85721; [2]The University of Hong Kong, School of Business, Hong Kong; [3]Department of Information Systems, Pace University, New York, NY 10038

Chapter Overview

The ability to create a big picture of a knowledge domain is valuable to both experts and newcomers, who can use such a picture to orient themselves in the field's intellectual space, track the dynamics of the field, or discover potential new areas of research. In this chapter we present an overview of medical informatics research by applying domain visualization techniques to literature and author citation data from the years 1994-2003. The data was gathered from NLM's MEDLINE database and the ISI Science Citation Index, then analyzed using selected techniques including self-organizing maps and citation networks. The results of our survey reveal the emergence of dominant subtopics, prominent researchers, and the relationships among these researchers and subtopics over the ten-year period.

Keywords

information visualization; domain analysis; self-organizing map; citation networks

1. INTRODUCTION

The rapid evolution of medical informatics and its subdomains makes it crucial for researchers to stay abreast of current developments and emerging trends. This task is made difficult, however, not only by the large amounts of available information, but by the interdisciplinary nature of the field. Relevant information is spread across diverse disciplines, posing a particular challenge for identifying relevant literature, prominent researchers, and research topics (Sittig, 1996, Andrews, 2002, Vishwanatham, 1998). Any attempt to understand the intellectual structure and development of the field must furthermore consider all of the contributing disciplines; as Börner et al. (2003) point out, "researchers looking at the domain from a particular discipline cannot possibly have an adequate understanding of the whole." In this chapter we report the results of an analysis of the medical informatics domain within an integrated knowledge mapping framework. We provide a brief review of the literature on knowledge mapping, then describe in detail the analysis design and results of our medical informatics literature mapping with three types of analysis: basic analysis, content map analysis, and citation network analysis.

2. KNOWLEDGE MAPPING: LITERATURE REVIEW

Domain analysis is a subfield of information science that attempts to reveal the intellectual structure of a particular knowledge domain by synthesizing disparate information, such as literature and citation data, into a coherent model (White and McCain 1997, Small 1999). Such a model serves as an overview to newcomers to the field, and reveals the field's dynamics and knowledge transfer patterns to experts.

A significant portion of domain analysis research has been focused on citation analysis. Historically, a great deal of manual effort was needed to gather citation data for this type of analysis by combining different literature resources and tracing through the citations. A manual analysis approach, however, is inherently subjective, and is impractical for the vast amounts of time-sensitive information available for most domains today (Börner et al., 2003). Digital citation indexes such as ResearchIndex (formerly CiteSeer) developed by NEC Research Institute (Lawrence et al. 1999) and ISI's Science Citation Index (SCI) eliminate the need for manual data collection, but still lead to large amounts of citation data that are difficult to analyze using traditional techniques. Recent developments in the field of domain visualization attempt to alleviate this citation information overload problem

by applying information visualization techniques to produce visual (and often interactive) representations of the underlying intellectual structure of the domain reflected in the large-scale citation data. A wide range of techniques have been applied to citation visualization, including clustering display based on co-citation (Small, 1999), the "Butterfly" display (Mackinlay et al., 1999), Pathfinder network scaling (Chen and Paul, 2001), and hyperbolic trees (Aureka, 2002).

Content, or "semantic," analysis is another important branch of domain analysis. This type of analysis relies on natural language processing techniques to analyze large corpora of literature text. Techniques ranging from simple lexical statistics to key phrase co-occurrence analysis to semantic and linguistic relation parsing are applied to reveal topic distribution and associations within the domain. To alleviate the similar information overload problem as for the citation data, many visualization techniques have been developed to produce content maps of large-scale text collections. Prominent examples include ThemeScape and Galaxies (Wise et al., 1995), the underlying techniques of which are multidimensional scaling and principle component analysis, and WebSOM (Honkela et al., 1997) and ET Map (Chen et al., 1996) which are based on the self-organizing map algorithm.

The application of visualization techniques to both citation and content analysis is consistent with the exploratory nature of domain analysis and forms the foundation of knowledge (domain) mapping. These visualization results provide valuable support for users' visual exploration of a scientific domain to identify visual patterns that may reflect influential researchers and studies, emerging topics, hidden associations, and other findings regarding the domain.

The effectiveness of domain analysis specifically in medical informatics is demonstrated by surveys by Sittig (1996) and Vishwanathan (1998), who used citation-based analyses to identify core medical informatics literature, and by Andrews (2002), who uses author co-citation analysis (ACA) to create multidimensional maps of the relationships between influential authors. We have also seen large-scale content mapping of the general medical literature (Chen et al., 2003), but not specifically of the medical informatics field.

In this study, we adopt the knowledge mapping framework proposed by Huang et al. (2003) that leverages large-scale visualization tools for knowledge mapping in fast-evolving scientific domains. Under this framework we perform three types of analysis -- basic analysis, content map analysis, and citation network analysis -- to provide a multifaceted mapping of the medical informatics literature. Through analyzing documents and citation information we identify influential researchers in the field and the

nature of their contributions, track knowledge transfer among the researchers, and identify domain subtopics and their trends of development. The results of our study present a comprehensive picture of medical informatics over the past ten years.

3. RESEARCH DESIGN

The Huang et al. (2003) framework proposes a generic set of analytical units, three analysis types, and various visualization technologies for representing the results of patent analysis. The analytical units include geographical regions, industries/research fields, sectors, institutions, individuals, and cross-units. Our medical informatics analysis focuses on individuals (authors), and research fields (subtopics) as units of analysis. We rely on two visualization techniques: self-organizing maps (SOMs) for revealing semantic grouping of topics, authors, and development trends; and citation networks for exploring knowledge transfer patterns. The details of our application of the Huang et al. three-pronged analysis are outlined below.

3.1 Basic Analysis

This first type of analysis provides "performance evaluation," namely, a measure of the level of an analytical unit's contribution to the field. Two types of measures are used for the contribution analysis, the productivity (or quantity) measures and impact (or quality) measures. We perform basic analysis at the author level to identify major researchers in medical informatics. The most prolific authors are determined by the number of publications attributed to them in our data set, with the highest-ranking authors deemed the most productive. A simple and commonly-used author impact measure is the number times an author is cited by others. The idea is that citation implies an acknowledgement of authority on the part of the citing author to the cited one, and that an author's citation level reflects the community's perceived value of their contribution to the field. This idea is supported by a substantial amount of academic literature on citation indexing. Garfield's 1955 vision of an interdisciplinary science citation index introduced the concept of citation as an impact factor indicator, and the concept has since been applied by the ResearchIndex in its citation context tool (Lawrence et al. 1999), Liu et al. (2004) in their AuthorRank indicator, and several domain analysis surveys (Andrews, 2002, Vishwanatham, 1998, Sittig, 1996, White and McCain, 1997, Chen et al., 2001, Noyons et al., 1999).

We expand on simple citation count by assigning authors an Authority score based on the HITS algorithm (Kleinberg, 1998), which was intended for identifying important web pages based on hyperlink citation structure. Following the formulation of the original HITS algorithm, two types of scores are defined for each author in our author citation analysis: an Authority score and a Hub score. An author with a high Authority score has a significant impact/influence on other authors, meaning his/her work has been extensively cited (directly and indirectly) by other authors. A high Hub score, on the other hand, indicates that an author's work has cited many influential studies. The Authority and Hub scores mutually reinforce each other: authors citing influential authors (with high Authority scores) tend to have high Hub scores; authors cited by authors who have cited influential authors (with high Hub scores) tend to be influential (with high Authority scores). With an author citation data set, we initialize the Authority scores as the number of times the authors are cited by others and the Hub scores as the number of times the authors cite others. The two scores are then computed following an iterative updating procedure:

$$\text{Authority Score}(p) = \sum_{q \text{ has cited } p} \text{Hub Score}(q)$$

$$\text{Hub Score}(q) = \sum_{q \text{ has cited } p} \text{Authority Score}(p)$$

The Authority score we use for our study is obtained with three iterations of score updating. It essentially incorporates the number of citations received by an author, the authors citing him/her, authors citing those citing authors, and so on.

3.2 Content Map Analysis

Content analysis is used in the Huang et al. framework to identify and track dominating themes in a field. Analyzing the content of the work produced by a specific analytical unit also provides valuable information on what subdisciplines that unit contributes to, and how the contribution changes over time. This approach augments traditional citation-based performance indicators (such as author co-citation) by operating directly on literature content, instead of inferring content from relationships between analytical units.

We use the self-organizing map (SOM) algorithm to perform content mapping of the medical informatics literature. Initially proposed by Kohonen (1990), the SOM algorithm analyzes similarities of entities with a large number of attributes and produces a map of the entities, in which the geographical distances correspond to the attribute-based similarities. In our study, we perform content mapping of papers and authors.

To generate the content maps, the text of each paper (a combination of titles and abstracts, in our study) is analyzed using the Arizona Noun Phraser, which identifies the key noun phrases based primarily on linguistic patterns (Tolle and Chen, 2000). These noun phrases, representing key concepts, are then used to represent the content of a paper by forming a binary vector, each element of which represents the occurrence of a particular noun phrase. The self-organizing map algorithm (SOM) typically produces a two-dimensional map to represent the content distribution of a set of documents. Each location in the map, that is, a node in a two-dimensional grid, is also assigned a key phrase vector, like the papers. These map node vectors are typically real-valued (for example, between 0 and 1) and initialized with random values. For each input paper, the SOM algorithm identifies a winning node that has the largest vector similarity measure to the input paper. The vector values of this winning node and its close neighbors are then updated to be more similar to the input paper vector. With all input pppapers used to perform the node vector updating process, the final configuration of the map, that is, the vector values of all map nodes, presents a content distribution of the input papers. The papers then obtain their locations in the map by finding the map nodes with the largest vector similarity measures. A map of authors is similarly generated by forming a key phrase vector for each author. The key phrase vector is created by combining the vectors for an author's papers, then used as input to the SOM algorithm in the same way as paper vectors.

We applied the multilayer SOM algorithms developed by Chen et al. (1995) to produce topic maps by adding a hierarchical topic region layer on top of a map of papers. We also perform longitudinal mapping, that is, a series of chronically sequential SOMs, to reveal the evolution of medical informatics subdisciplines. From the maps, a researcher can observe what disciplines exist at different points in time, when particular disciplines emerge, and their rate of growth and decline. A domain expert can potentially use such longitudinal maps to forecast emerging trends (Börner et al., 2003).

We also created an author map using the SOM algorithm. Based on the positions of the authors in the map, we identify groups of authors that had papers with similar contents.

3.3 Citation Analysis

Visualizing citation data as a network is a classic method for intuitively displaying knowledge transfer patterns among analytical units. Citation networks consist of nodes representing the analytical units, with directional links representing citations between them. When the analytical unit is an

author, such networks can be used to quickly identify strong communication channels in the domain, and the structure of those channels. Since citation between authors implies a human judgment that a work by the cited author is relevant to one by the citing author, frequently-occurring citations can indicate that two authors work in a similar field. Hence, citation networks can be used to identify communities of researchers. For this study, we gathered citation information from ISI's Science Citation Index for the years 1994-2003 for a core group of researchers identified by the basic analysis. We then use the freely-available graphing program NetDraw (http://www.analytictech.com/netdraw.htm) to visualize the result.

4. DATA DESCRIPTION

Andrews (2002) points out that an author co-citation analysis is only as good as the analyst's choice of authors. The same can be said for domain analysis in general. We used a number of measures to collect as comprehensive a data set for our survey as possible. First, we used NLM's expansive MEDLINE database of biomedical literature to provide source documents for our analysis. We then used four criteria to locate documents in MEDLINE relevant to medical informatics. For an article to be included in our collection, at least one of the following had to be true:

1. The article was published in one of 22 prominent journals in the medical informatics domain. These journals consist of the 18 identified by Andrews (2002) and additionally two journals and two conference proceedings that are frequently cited in (Shortliffe et al., 2000). The complete list of journal titles is given in Table 2-1.
2. The article abstract or title contains one of the selected medical informatics keywords listed in Table 2-2.
3. The article is indexed by MEDLINE under the MeSH term "Medical Informatics." MeSH is widely acknowledged to be an authoritative indexing system.
4. The article was authored by a fellow of the American College of Medical Informatics (ACMI), a group of scholars who are determined by their peers to have made “significant and sustained contributions to the field” (http://www.amia.org/acmi/acmi.html).

The use of ACMI fellows as a test set on which to perform domain analysis is supported by Andrews (2002), who also cites the use of ACMI by Greenes and Siegel (1987).

Using the above criteria, we identified 24,495 medical informatics articles in MEDLINE, as of August 2004. Restricting our data set to articles published during our ten-year test bed, 1994-2003, yielded 16, 964 articles.

Table 2-1. Prominent medical informatics journals included in our study.

Journal Name
Artificial Intelligence in Medicine
Biomedizinische Technik (Biomedical Engineering)
Computer Methods and Programs in Biomedicine
Computers, Informatics, Nursing: CIN
IEEE Engineering in Medicine and Biology Magazine
IEEE Transactions on Information Technology in Biomedicine
International Journal of Medical Informatics
International Journal of Technology Assessment in Health Care
Journal of Biomedical Informatics
Journal of Cancer Education: The Official Journal of the American Association for Cancer Education
Journal of Evaluation in Clinical Practice
Journal of the American Medical Informatics Association (JAMIA)
M.D. Computing: Computers in Medical Practice
Medical and Biological Engineering and Computing
Medical Informatics and the Internet in Medicine
Medical Decision Making
Methods of Information in Medicine
Proceedings of the American Medical Informatics Association (AMIA) Annual Fall Symposium
Proceedings of the Annual Symposium on Computer Applications in Medical Care
Statistical Methods in Medical Research
Statistics in Medicine

Table 2-2. Keywords used to identify MEDLINE documents relevant to medical informatics.

Keyword
Medical informatics
Clinical informatics
Nursing informatics
Health informatics
Bioinformatics
Biomedical informatics

As White and McCain (1997) state, "we wished to let 'the field' dictate its top authors rather than choosing them ourselves." This means that in addition to using ACMI fellows for our analysis, we allowed our document set to determine the remainder of our author set: anyone identified as an author of an article in the medical informatics collection was included in our collection of authors. A count of the most frequently-occurring names in the collection determined the most prolific authors in the field, as listed in Table

2-3. These authors comprise the “core” set used to gather citation data from the Science Citation Index (SCI). As of this study, SCI is only searchable through the online Web of Science. A "citation search" was manually performed in the Web of Science for each author in our core set, to gather information on who has cited them, and who they cite. This search yielded some commonly-cited names that are not included in our core set, which can be seen in Tables 2-4 and 2-5. Together the core set and frequently-cited names list some of the most recognizable and influential researchers in the field, and citation information for all of these authors was used for our citation analysis.

5. RESULTS

5.1 Basic Analysis

Our basic analysis focused on authors as the analytical unit, with the results presented in Tables 2-3, 2-4, and 2-5. These tables offer different perspectives - productivity and impact factor, respectively - on the most highly contributing researchers in the domain. Table 2-3 lists the 96 most prolific authors, that is, those with the most publications attributed to them in our data set. James J. Cimino at Columbia University tops the list with 62 publications, followed closely by Arie Hasman at the University of Maastricht in the Netherlands, Robert A. Greenes of Harvard Medical School, and Perry L. Miller at Yale University. The citation search described in Section 4 above yielded some frequently cited authors that do not appear in the core set shown in Table 2-3. Citation counts were gathered for these authors in addition to those in the core set, and the most frequently cited of the combined list are shown in Table 2-4. Some authors of note in the list that do not appear among the core authors in Table 2-3 are Lucian L. Leape at the Harvard School of Public Health, Mor Peleg at Stanford University, and Suzanne Bakken at Columbia University.

Table 2-5 ranks the authors in the combined list by their citation-based Authority scores. James Cimino is again among the five highest scoring in this table, along with Mark A. Musen at Stanford University, Edward H. Shortliffe at Columbia University (formerly at Stanford), George Hripcsak at Columbia, and Paul D. Clayton, who was at Columbia until 1998 and is currently Chief Medical Informatics Officer at Intermountain Health Care in Salt Lake City. The latter four authors are shown in Table 2-3 to have approximately half the number of publications as the most prolific author, yet their Authority scores indicate the significant impact of their publications.

Table 2-3. Publication counts for prolific authors.

Author name	Number of publications in collection	Author name	Number of publications in collection
Cimino, James J.	62	Van der Lei, J.	22
Hasman, A.	52	Kahn, Michael G.	22
Greenes, Robert A.	45	Friedman, Carol	22
Miller, Perry L.	44	Rector, Alan L.	22
Haux, Reinhold	42	Whitehead, J.	21
Musen, Mark	39	Cerutti, S.	21
Patel, Vimla L.	38	Tierney, William M.	21
Safran, Charles	37	Warner, Homer R.	21
Barnett, Octo G.	35	Habbema, J. D.	20
Stefanelli, Mario	35	Friedman, Charles P.	20
Miller, Randolph A.	31	Beck, J. Robert	20
Shortliffe, Edward	31	Royston, P.	19
Van Bemmel, J. H.	30	Zhou, X. H.	19
Haug, Peter	29	McDonald, Clement	19
Hripcsak, George	29	Wigton, Robert S.	19
Fagan, Larry	29	Shahar, Y.	18
Kohane, Issac	28	Fieschi, M.	18
Weinstein, M. C.	27	Lui, K. J.	18
Degoulet, Patrice	27	Haynes, R. Brian	18
Bates, David W.	27	Brinkley, James	18
Lenert, Leslie A.	27	Brennan, Patricia F.	18
Durand, L. G.	26	Kuperman, Gilad J.	18
Timpka, T.	26	Stead, William W.	18
Chute, Christopher	26	Tuttle, Mark S.	18
Clayton, Paul D.	26	Pinciroli, F.	17
Johnson, Stephen B.	26	Bolz, A.	17
Sittig, Dean F.	26	Spiegelhalter, D. J.	17
Greenland, S.	25	Simon, R.	17
Pfurtscheller, G.	25	Mitchell, Joyce A.	17
Hersh, William R.	25	Ohno-Machado, Lucila	17
Donner, A.	24	Tang, Paul C.	17
Thompson, S. G.	24	Tu, Samson W.	17
Huff, Standley M.	24	Van Ginneken, A.M.	16
Gardner, Reed M.	24	Dössel, O.	16
Dudeck, Joachim	24	Freedman, L. S.	16
Nadkarni, Prakash	24	Groth, T.	16
Teich, Jonathan M.	24	Meinzer, H. P.	16
Bellazzi, R.	23	Altman, Russ B.	16
Cooper, Greg	23	Reggia, James A.	16
Scherrer, Jean-Raoul	23	Slack, Warner V.	16
Wigertz, Ove	23		

Table 2-4. Citation counts for frequently cited authors.

Author name	Times cited by authors in medical informatics collection	Author name	Times cited by authors in medical informatics collection
Bates, D. W.	989	Greenes, R. A.	142
Cimino, J. J.	691	Lui, K. J.	137
McDonald, C. J.	359	Giuse, D. A.	135
Patel, V. L.	356	Neuper, C.	134
Hripcsak, G.	331	McCray, A. T.	131
Pfurtscheller, G.	306	Hersh, W. R.	129
Friedman, C.	301	Rind, D. M.	128
Miller, R. A.	289	Riva, A.	127
Musen, M. A.	287	Montani, S.	123
Greenland, S.	280	Huff, S. M.	123
Bellazzi, R.	243	Kuhn, K. A.	123
Overhage, J. M.	225	Johannesson, M.	122
Leape, L. L.	219	Kaplan, B.	120
Peleg, M.	215	Baud, R. H.	119
Hasman, A.	206	Lenert, L. A.	119
Bakken, S.	196	Combi, C.	117
Campbell, K. E.	188	Fox, J.	117
Chute, C. G.	183	Zeng, Q.	114
Shahar, Y.	180	Das, A. K.	114
Haux, R.	175	Degoulet, P.	113
Kushniruk, A. W.	167	Perl, Y.	113
Elkin, P. L.	167	Spackman, K. A.	112
Zhou, X. H.	164	Johnston, M. E.	112
Kuperman, G. J.	162	Safran, C.	112
Boxwala, A. A.	157	Owens, D. K.	111
Simon, R.	155	Andreassen, S.	111
Evans, R. S.	152	Friedman, C. P.	111

Table 2-5. Authority score ranking for frequently cited authors.

Author name	Authority score	Author name	Authority score
Clayton, P. D.	4.06	Tierney, W. M.	1.93
Cimino, J. J.	4.00	Tuttle, M. S.	1.89
Hripcsak, G.	3.86	Johnston, M. E.	1.84
Musen, M. A.	3.66	Hasman, A.	1.80
Shortliffe, E. H.	3.58	Brennan, P. F.	1.77
Safran, C.	3.54	McDonald, C. J.	1.63
Barnett, G. O.	3.33	Miller, P. L.	1.58
Greenes, R. A.	3.31	Shea, S.	1.57
Campbell, K. E.	3.01	Stefanelli, M.	1.56
Hersh, W. R.	2.95	Overhage, J. M.	1.49
Stead, W. W.	2.90	Ohnomachado, L.	1.42
Gardner, R. M.	2.90	Haynes, R. B.	1.37
Bates, D. W.	2.87	Friedman, C.	1.36

continued

Author name	Authority score	Author name	Authority score
Chute, C. G.	2.82	Lobach, D. F.	1.38
Kuperman, G. J.	2.76	Humphreys, B. L.	1.34
Friedman, C. P.	2.73	Haux, R.	1.33
Rector, A. L.	2.68	Rind, D. M.	1.29
Teich, J. M.	2.67	Evans, R. S.	1.25
Sittig, D. F.	2.64	Zielstorff, R. D.	1.21
Shahar, Y.	2.47	Peleg, M.	1.20
Warner, H. R.	2.45	McCray, A. T.	1.18
Slack, W. V.	2.41	Kohane, I. S.	1.16
Haug, P. J.	2.23	Dolin, R. H.	1.11
Tang, P. C.	2.19	Leape, L. L.	1.10
Patel, V. L.	2.12	Tu, S. W.	1.09
Miller, R. A.	2.09	Owens, D. K.	1.02
Shiffman, R. N.	2.00	Spackman, K. A.	1.02
Huff, S. M.	1.98	Van Bemmel, J. H.	1.01

5.2 Content Map Analysis

5.2.1 Topic Map Analysis

The content map analysis uses time-series topic maps to present development trends in medical informatics over the ten years. For this temporal analysis we created topic maps of three periods, 1994-1997, 1998-2000, and 2001-2003. By breaking the medical informatics papers published over the past decade into three periods, we hope to glean the recent evolution and topic changes of the field. To generate the maps, the abstracts and titles of 5,837 papers in our collection were processed for 1994-1997, 5,755 for 1998-2000, and 5,375 for 2001-2003.

In these topic maps clusters of papers are represented by shaded regions and labeled by representative noun phrases appearing in those papers. The medical noun phrases were extracted using the Arizona Noun Phraser as described previously. These noun phrases were extracted from the original text and the capitalization varies. However, phrases with capitalization variations were treated as the same phrases for the phrase vector representation. Numbers of papers within each cluster are presented in parentheses after the topic labels. As described previously in Section 3.2, neighboring topic regions have high content similarities. Users can click on the map regions to browse the papers.

The first topic map (Figure 2-1) displays an assortment of dominating themes for the first time period. There are many prominent but general medical information topics that occupy large regions, including: “Electronic Medical Records,” “Computer-Based Patient Record,” “Health Care,”

"Information Technologies," "Computer Programs," "Medical Students," etc. A few specific medical informatics applications also occupy large regions, including: "Hospital Information Systems" and "Clinical Information Systems." In addition, we also notice several small but distinct topic regions that are related to data analysis and mining, e.g., "Decision Support Systems," "Statistical Analysis," "Regression Models," "Artificial Neural Networks," and "Neural Networks." It appears that data mining and knowledge discovery research had already begun to emerge in 1994-1997, the first era of our analysis.

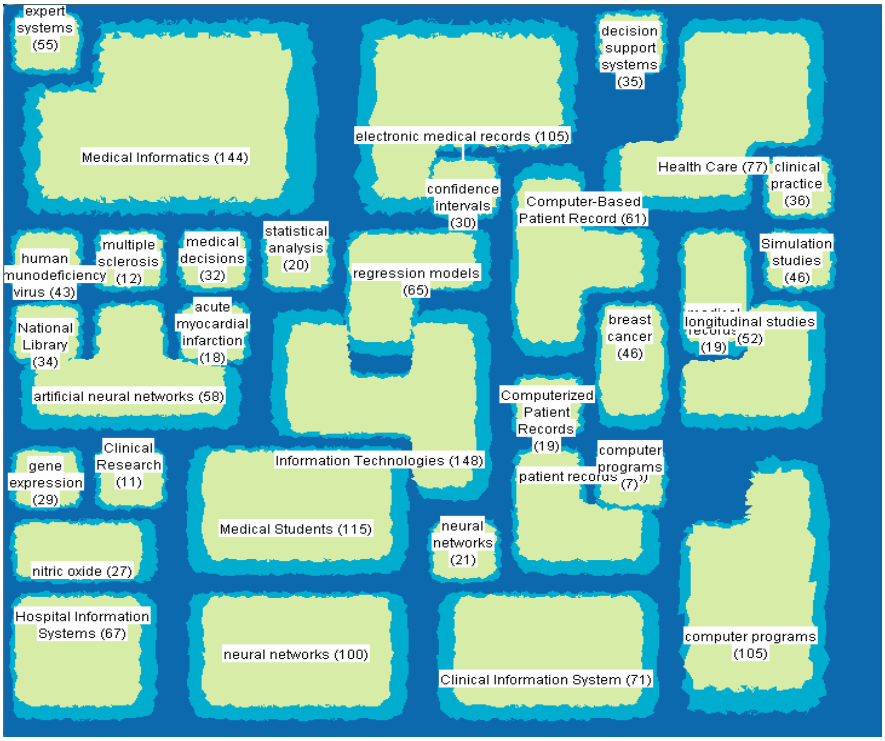

Figure 2-1. Top level content map for 1994-1997.

The topic regions in the second and third time periods were colored to reflect the growth rate of the topic compared with the previous time period (not shown here due to production reasons), which is computed as the ratio between the number of papers in the region for the current time period, and the number of papers in the region of the same topic label in the previous time period. The color legend of the growth rate is presented as well below these two content maps. In Figure 2-2, regions such as "Human Genome" and "Medical Imaging" correspond to the right end of the color legend,

which represents newly emerged topic regions, while regions with lighter colors such as "Hospital Information System" corresponds to color legends close to the left end, which represent topic regions that had a slow or average growth rate.

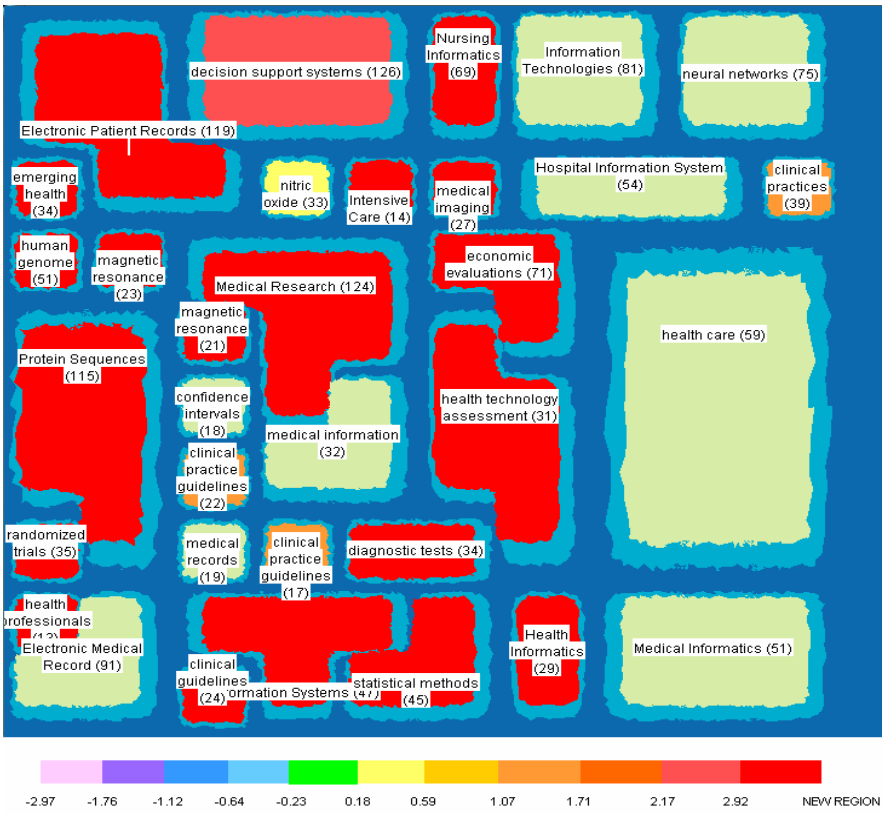

Figure 2-2. Top level content map for 1998-2000.

In the second map (Figure 2-2), we see the continued presence of several important, but general medical informatics topic regions, including: "Health Care," "Information Technologies," "Electronic Medical Records," "Hospital Information Systems," etc. Several data analysis and mining topics began to occupy larger regions than in 1994-1997, e.g., "Decision Support Systems" and "Neural Networks." In addition, "Protein Sequence" and "Human Genome" topics emerged the first time, increasing the scope of biomedical data. There is also an increased diversity of applications and methodologies such as: "Nursing Informatics," Medical Imaging," "Economic Evaluation," and "Health Technology Assessment."

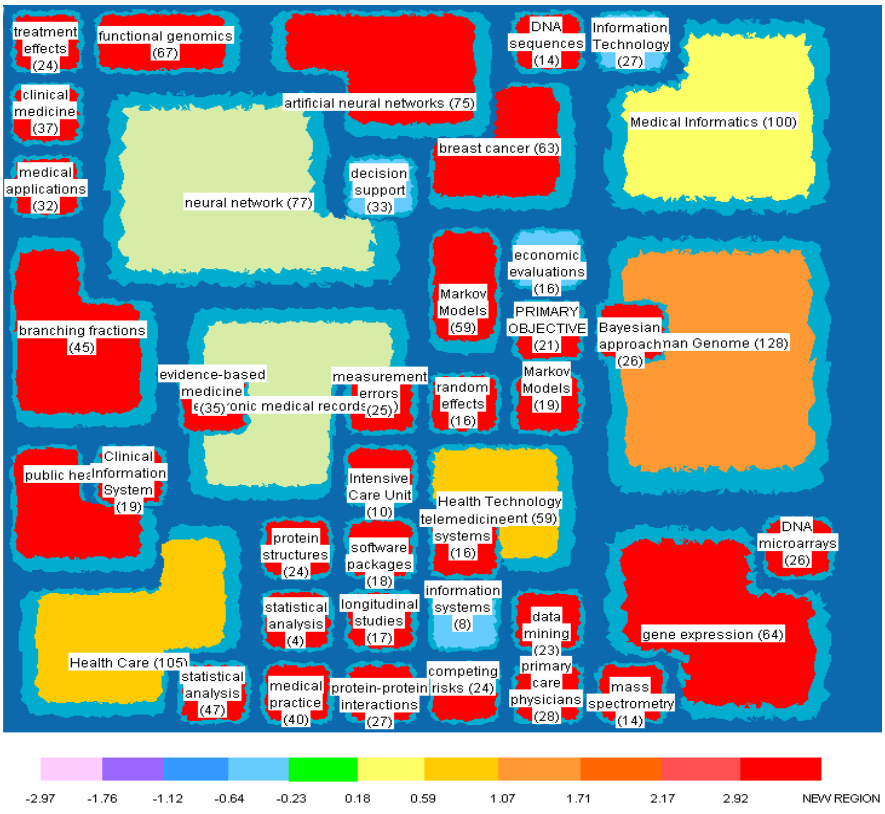

Figure 2-3. Top level content map for 2001-2003.

In addition to some of the general medical informatics topics ("Health Care," "Medical Informatics," etc.), the third map (Figure 2-3) shows a strong presence of data mining and knowledge discovery topics in 2001-2003 including: "Neural Networks," "Artificial Neural Networks," "Bayesian Approach," "Data Mining," "Markov Models," etc. Most interestingly, we see an explosion of biological and genomic data types and applications, including: "DNA Microarrays," "DNA Sequences," "Gene Expression," "Mass Spectrometry," "Protein-Protein Interactions," "Functional Genomics," etc.

The pattern of mixed topics observed between maps is consistent with the observation that medical informatics is a fast-growing, multidisciplinary field (Andrews, 2002). Sittig (1996) and Greenes and Siegel (1987) recount the difficulty of defining the boundaries of the medical informatics domain, and the resulting diversity of subfields attributed to it. Despite such challenges, we observed a consistent focus on health care, electronic medical records, and information technologies topics in general in the three eras of

analysis. In addition, we also see overwhelming evidence of the presence of many emerging and exciting data mining and knowledge discovery research applications, especially those which leverage the opportunities presented by a wide spectrum of new, diverse, and large-scale biological and genomic data and problems.

5.2.2 Author Map Analysis

The author map in Figure 2-4 attempts to group individual researchers in the domain space, based on their common research interests. For this analysis we used the core author set from Table 2-3 as the input data. The result presents five major clusters of authors who had papers with similar contents. Each resulting cluster has been assigned a label indicating the common concept(s) that the cluster represents. The labels were manually selected from the keywords extracted by the SOM algorithm, a process which requires human judgment, but as Andrews (2002) points out, consistent with other cluster analysis methods. The keywords used to determine each label are listed in Table 2-6, and the individual groups are shown in detail in Figures 2-5 through 2-8 (with the exception of Group 3, which was decided not to be dense enough to require a zoomed in view).

Table 2-6. Top keywords generated from authors' texts and used to label author map groups.

Group 1	Group 2	Group 3
Decision support system	Clinical trials	Clinical applications
Decision support	Breast cancer	Clinical information system
Expert system	Risk factors	
Knowledge-based system	Cardiovascular disease	
	Coronary heart disease	
Group 4	**Group 5**	
Patient care	Clinical trials	
Medical record	Cohort study	
Electronic medical record	Confidence intervals	
Unified medical language system	Multivariate analysis	

The largest group in the center of the author map, Group 1, is labeled "Decision support and knowledge-based systems." This group contains 37 of the 96 authors, including W.R. Hersh, C.G. Chute, and M.A. Musen. Author proximity on the map indicates a degree of similarity between the research interests. Group 2, "Clinical trials for diseases," contains 15 authors, including R.A. Miller, Y. Shahar, and M. Stefanelli. Group 3, "Clinical applications and information systems," contains 6 authors, among them D.W. Bates and P.D. Clayton. Group 4, labeled "Patient care and electronic medical records," is comprised of such prolific authors as J.J. Cimino, D.F. Sittig, R.A. Greenes, C. Friedman, and E.H. Shortliffe.

Finally, Group 5, "Clinical trials and analysis," contains 8 authors, among them A. Donner and K.J. Lui. Authors in our original 96 that are not included in a group can be seen in the overall map in Figure 2-4.

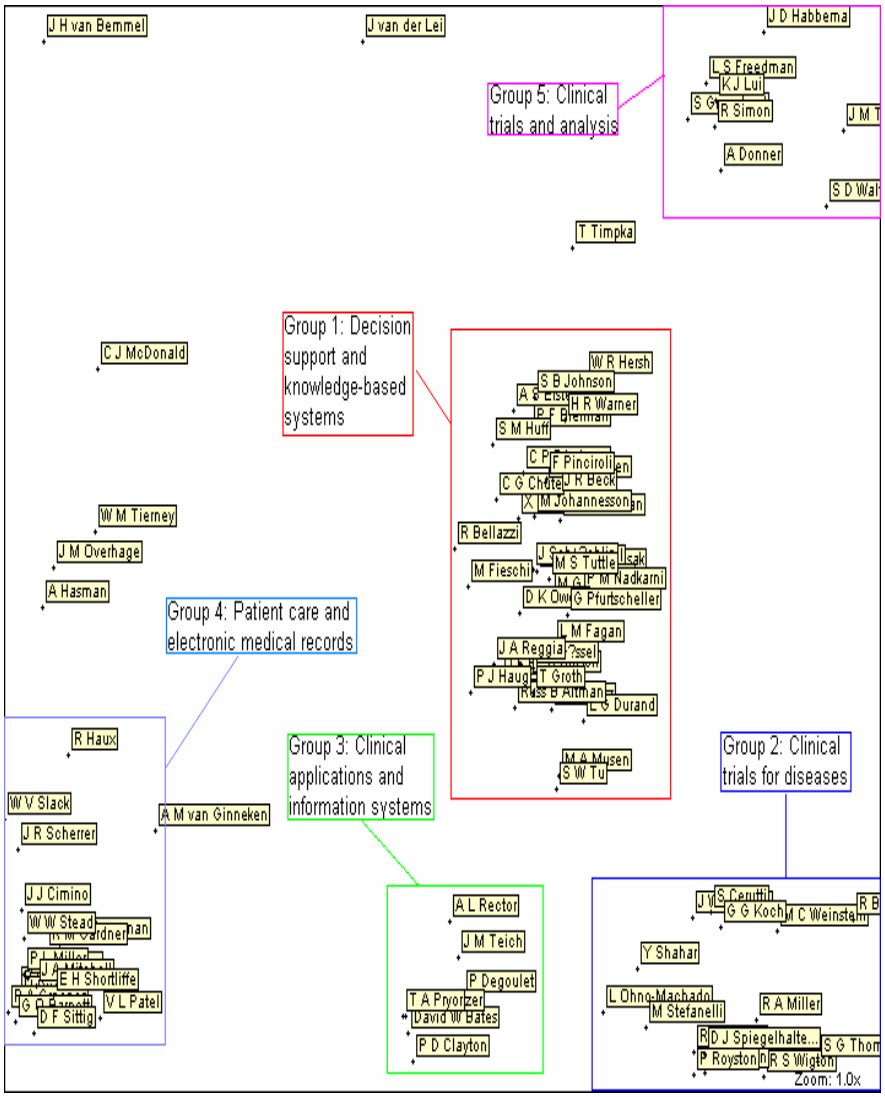

Figure 2-4. Overall author similarity map.

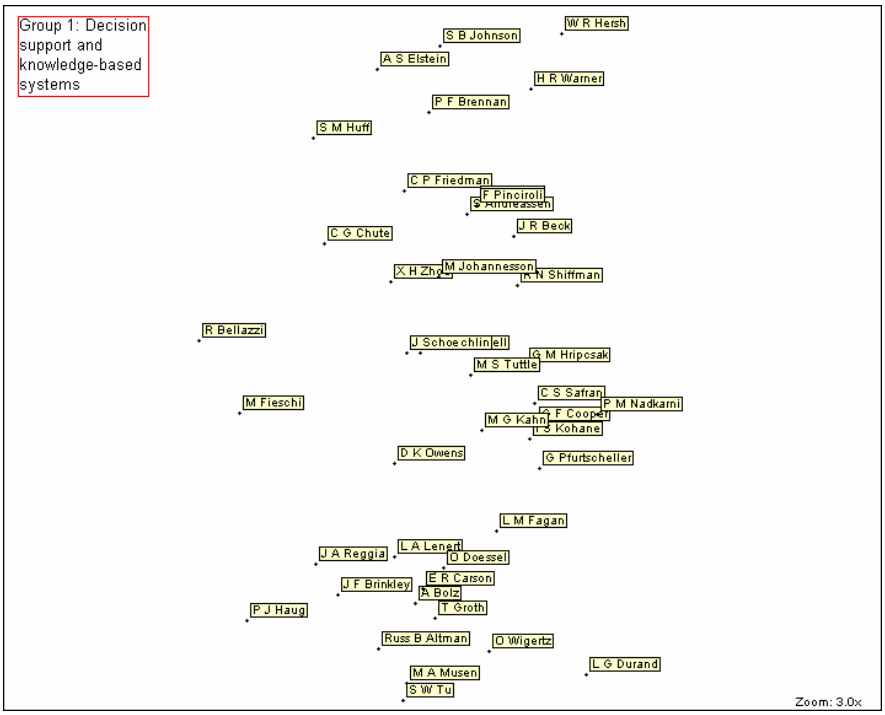

Figure 2-5. Author map - Group 1.

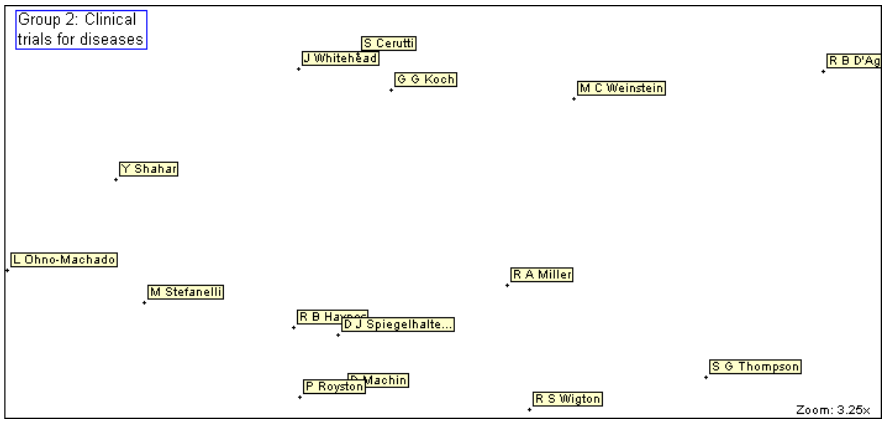

Figure 2-6. Author map - Group 2.

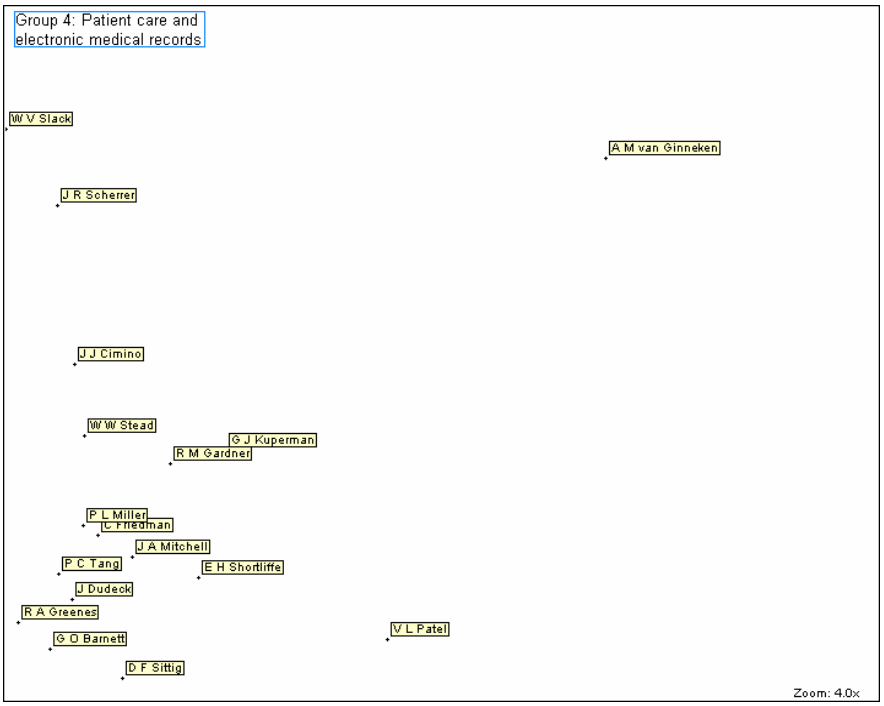

Figure 2-7. Author map - Group 4.

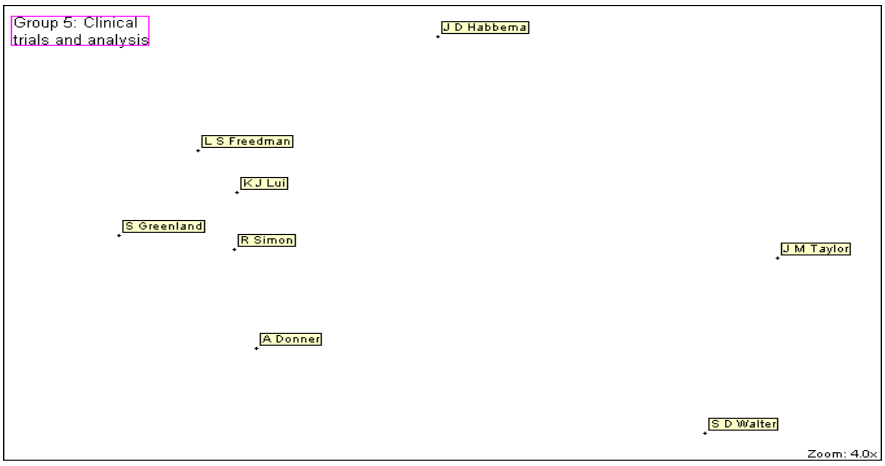

Figure 2-8. Author map - Group 5.

5.3 Citation Network Analysis

Using the data gathered from SCI, we created two citation networks of the most prominent researchers in medical informatics, as identified by our basic analysis. Both networks present views of the same data with different levels of filtering. A link from author A to author B indicates that A frequently cites B. In the visualization results, triangles indicate "core" authors (presented in Table 2-3) and circles represent "non-core" authors. In order to reveal only the strongest communication patterns, links associated with a small number of citations are filtered from the networks. Figure 2-9 is filtered by a link threshold of 10, that is, only links associated with 10 or more citations are shown. The result is a rather dense cluster, but hubs can still be observed around the major players from our basic analysis results: Edward H. Shortliffe, Paul D. Clayton, George Hripcsak, David W. Bates, James J. Cimino, and William R. Hersh, to name a few. These authors are not only frequently published and cited, they are cited repeatedly by consistent sets of other authors. Figure 2-10 is a view of the same citation data, filtered by a threshold of 20. In this view, clearer subgroups of citations emerge. One distinct subgroup of eight authors is disconnected from the larger graph. This group appears in the upper right-hand part and consists of four "core" authors from Table 2-3, and four "non-core" authors from Table 2-4. In the larger graph itself, hubs from Figure 2-9 begin to pull apart into subgroups. The most distinct group clusters around David Bates and William M. Tierney, and includes high-ranking authors from the basic analysis, such as Dean F. Sitting and Jonathan M. Teich. Other subgroups of the larger graph can be observed but are much less distinct. Obvious hubs are James Cimino, George Hripcsak, and Edward Shortliffe. Tightly connecting these are Carol Friedman, Vimla L. Patel, and Robert A. Greenes.

It should be noted that as a result of filtering by link strength, the citation networks do not reflect an overall qualitative performance measure of the authors, but rather the nature of their communication channels. That is, the graphs do not show who is the most cited, but who most frequently cites whom. It can be observed, for example, that there are no links to William Hersh in the 20-threshold network; however, our basic analysis indicates that Hersh is highly influential in the field, and is cited by numerous other authors. According to Figure 2-10, he is simply not cited more than 19 times by the same author. In contrast, there are two incoming links to Christopher G. Chute (from James Cimino and Peter L. Elkin). Chute is only slightly below Hersh in Authority ranking, but frequently cites and is cited by two specific authors, so is connected to the main graph.

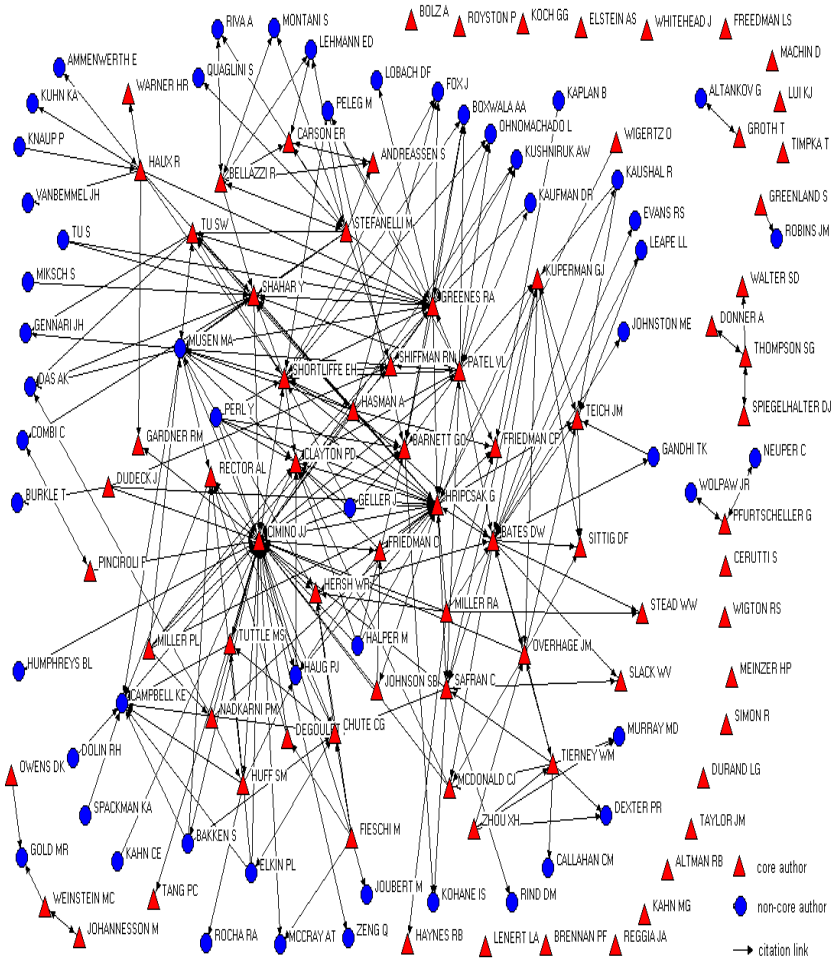

Figure 2-9. Author citation network (minimum cites per link: 10).

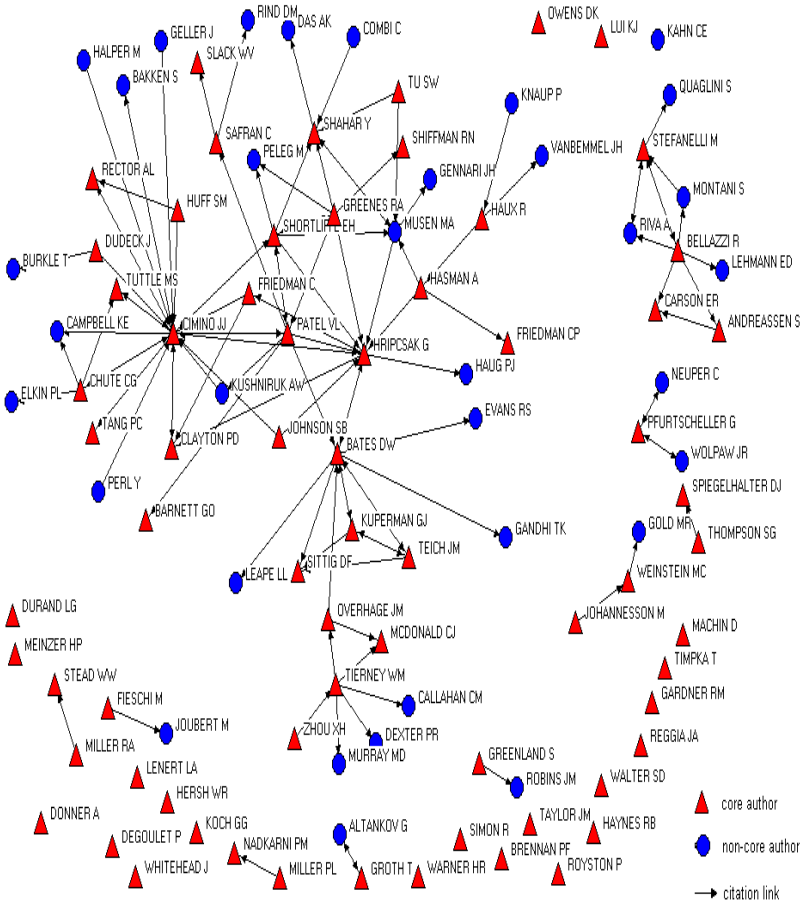

Figure 2-10. Author citation network (minimum cites per link: 20).

6. CONCLUSION AND DISCUSSION

For a fast-growing, interdisciplinary knowledge domain such as medical informatics, it is valuable to be able to create a picture of the state of the research from a variety of perspectives. Such a picture helps organize the vast amounts of information available in order to determine past and current (and possibly future) directions of the field, as well as prominent researchers, their relationships to each other, and the parts of the domain to which they contribute. Automatic information visualization techniques can perform these knowledge tasks efficiently and systematically. In this study

we augment classic domain analysis techniques with visualization tools to create a variety of views of medical informatics over the past ten years. The results of our study present development trends of subtopics of the field, a performance evaluation of the prominent researchers, and graphs of knowledge transfer among researchers.

This study was designed in the context of the analysis framework developed by Huang et al. (2003), and implements the three types of analysis presented in that work: basic analysis, content maps, and citation networks. Based on the data set extracted from widely-used data sources such as the MEDLINE database and SCI, we believe our analysis helps reveal the coverage and evolution of the field. It would be interesting to compare the particular findings from our analysis with the pictures of the field in the minds of the domain experts. Such evaluation would help determine how accurate our analysis results are and reveal interesting discrepancies between automatic analysis results and expert knowledge that might enhance our understanding of the state of the field.

7. ACKNOWLEDGEMENTS

This research was supported by the following grants: (1) NIH/NLM, 1 R33 LM07299-01, 2002-2005, "Genescene: A Toolkit for Gene Pathway Analysis" and (2) NSF, IIS-9817473, 1999-2002, "DLI - Phase 2: High Performance Digital Library Classification Systems: From Information Retrieval to Knowledge Management."

REFERENCES

Andrews, J. (2002). "An Author Co-citation of Medical Informatics," *Journal of the Medical Library Association*, 91(1), 47-56.

Borgatti, S. and Chase, R. (n.d.) NetDraw Network Visualization Tool. Version 1.39, Retrieved Aug. 30, 2004, http://www.analytictech.com/netdraw.htm

Börner, K., Chen, C., and Boyack, K. (2003). "Visualizing Knowledge Domains," in Blaise Cronin (Ed.), *Annual Review of Information Science and Technology*, 37: 179-255. Information Today, Inc. and American Society for Information Science and Technology, 2003.

Boyak, K. W. and Börner, K. (2003). "Indicator-assisted Evaluation and Funding of Research: Visualizing the Influence of Grants on the Number and Citation Counts of Research Papers," *Journal of the American Society for Information Science and Technology*, 54(5), 447 - 461.

Chen, C. and Paul, R. J. (2001). "Visualizing a Knowledge Domain's Intellectual Structure," *IEEE Computer*, 34(3), 65-71.

Chen, C., Paul, R. J., and O'Keefe, B. (2001). "Fitting the Jigsaw of Citation: Information Visualization in Domain Analysis," *Journal of the American Society for Information Science and Technology*, 52(4), 315-330.

Chen, H., Houston, A. L., Sewell, R. R., and Schatz, B. R. (1998). "Internet Browsing and Searching: User Evaluation of Category Map and Concept Space Techniques," *Journal of the American Society for Information Science*, 49(7), 582-603.

Chen, H., Lally, A., Zhu, B., and Chau, M. (2003). "HelpfulMed: Intelligent Searching for Medical Information over the internet," *Journal of the American Society for Information Science and Technology (JASIST)*, 54(7), 683-694.

Chen, H., Schuffels, C., and Orwig, R. (1996). "Internet Categorization and Search: A Self-organizing Approach," *Journal of Visual Communication and Image Representation*, **7(1),** 88-102.

Garfield, E. (1979). *Citation Indexing: Its Theory and Application in Science, Technology and Humanities*. New York, NY: John Wiley.

Garfield, E. (1995). "Citation Indexes for Science: a New Dimension in Documentation Through Association of Ideas," *Science*, 122, 108-111.

Greenes, R.A. and Siegel, E.R. (1987). "Characterization of An Emerging Field: Approaches to Defining the Literature and Disciplinary Boundaries of Medical Informaticcs," in *Eleventh Annual Symposium on Computer Applications in Medical Care*, 411–415.

Honkela, T., Kaski, S., Lagus, K., and Kohonen, T. (1997). "WebSom - Self-Organizing Maps of Document Collections," in *Proceedings of the Workshop on Self-Organizing Maps*, 310-315.

Huang, Z., Chen, H., Yip, A., Ng T. G., Guo, F., Chen, Z. K., and Roco, M. C. (2003). "Longitudinal Patent Analysis for Nanoscale Science and Engineering: Country, institution and Technology Field," *Journal of Nanoparticle Research*, 5, 333-386.

Kleinberg, J. (1998). "Authoritative Sources in a Hyperlinked Environment," in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 668-677

Kohonen, T. (1990) "The Self-Organizing Map," in *Proceedings of the IEEE*, 78(9), 1464-1480.

Lawrence, S., Giles, C. L., and Bollacker, K. (1999). "Digital Libraries and Autonomous Citation Indexing," *IEEE Computer*, 32(6), 67-71.

Lin, C., Chen, H., and Nunamaker, J. F. (2000). "Verifying the Proximity Hypothesis for Self-organizing Maps," *Journal of Management Information Systems*, 16(3), 57-70.

Liu, X., Bollen J., Nelson M. L., and Van De Sompel, H. (2004). "All in the Family? A Co-authorship Analysis of JCDL Conferences (1994 - 2003)," http://lib-www.lanl.gov/~xliu/trend.pdf

Mackinlay, J. D., Rao, R., and Card, S. K. (1999). "An Organic User Interface for Searching Citation Links," in *Proceedings of the CHI'95, ACM Conference on Human Factors in Computing Systems*, 67-73.

Noyons, E. C. M., Moed, H. F., and Luwel, M. (1999). "Combining Mapping and Citation Analysis for Evaluative Bibliometric Purposes: A Bibliometric Study," *Journal of American Society for Information Science*, 50(2), 115-131.

Shortliffe, E. H, Fagan, L., Perreault, L. E., and Wiederhold, G. (Eds.) (2000). *Medical Informatics: Computer Applications in Health Care and Biomedicine* (2nd Edition). New York, NY: Springer Verlag.

Sittig, D. F. (1996). "Identifying a Core Set of Medical Informatics Serials: An Analysis Using the MEDLINE Database," *Bulletin of the Medical Library Association*, 84(2), 200-204

Small, H. (1999). "Visualizing Science by Citation Mapping," *Journal of the American Society for Information Science*, 50(9), 799-812.

Tolle, K. and Chen, H. (2000). "Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools," *Journal of the American Society for Information Science, Special Issue on Digital Libraries*, 51(4), 518-22.

Vishwanatham, R. (1998). "Citation Analysis in Journal Rankings: Medical Informatics in the Libary and Information Science Literature," *Bulletin of the Medical Library Association*, 86(4), 518-22.

White, H. D. and McCain, K. (1998). "Visualizing a Discipline: An Author Co-citation Analysis of Information Science, 1972 - 1995," *Journal of the American Society for Information Science*, 49(4), 327 - 355.

Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., and Crow, V. (1995). "Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents," in *Proceedings of the IEEE Information Visualization 95 (InfoViz'95)*, 51-58.

SUGGESTED READINGS

Andrews, J. (2002). "An author co-citation of medical informatics," *Journal of the Medical Library Association*, 91(1), 47-56.
Andrews applies multivariate analyses and visualization techniques to map relationships between the fifty most-cited ACMI fellows for the years 1994 to 1998.

Cronin, B. (Ed). (2003). *Annual Review of Information Science and Technology*, Vol 37. Medford, NJ: Information Today, Inc./American Society for Information Science and Technology.
Number 37 in a series that offers a comprehensive overview of information science and technology. This volume contains chapters on indexing and retrieval for the web, and visualizing knowledge domains in general.

Chen, C. (2003). *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. Secaucus, NJ: Springer-Verlag.
A thorough investigation of the effectiveness of using visualization tools to reveal shifts in scientific paradigms, and of the need for interdisciplinary research in information visualization and information science.

Chen, C., Paul, R. J. (2001). Visualizing a knowledge domain's intellectual structure. *IEEE Computer*. 34(3), 65-71.
Introduces Pathfinder network scaling to produce a 3D knowledge landscape from science citation patterns. The authors propose a four-step approach to "extends and transform" traditional author citation and co-citation analysis.

Garfield, E. (1979). Citation Indexing: Its theory and application in science, technology and humanities. John Wiley, New York.
Garfield's influential review of the creation and usefulness of citation indexes for understanding knowledge domains, especially since his seminal 1955 paper on the subject (*Science*, 122, 108-111).

Honkela, T., Kaski, S., Lagus, K., Kohonen, T. (1997). WebSom - Self-Organizing Maps of Document Collections. *Proceedings of the Workshop on Self-Organizing Maps*. 310-315.
Introduces WEBSOM, a well-known application of the SOM algorithm to organize high dimensional text documents according to similarity, and to present the results in an intuitive user interface.

Kohonen, T. (1990) The Self-Organizing Map, *Proceedings of the IEEE*. 78(9), 1464-1480.
Influential review and demonstration of various applications of the SOM algorithm.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*. 50(9), 799-812.
Demonstrates the use of associative trails and virtual reality software to create and navigate spatial representations of a sample of multidisciplinary science citation data. The author also provides a nice overview discussion and justification for applying information visualization techniques to science.

White, H. D., McCain, K. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972 - 1995. *Journal of the American Society for Information Science*. 49(4), 327 - 355.
The authors use author co-citation data to map the field of information science.

ONLINE RESOURCES

ISI Science Citation Index, through the Web of Science
ISI Journal Citation Reports
http://isi6.isiknowledge.com/portal.cgi

ResearchIndex (also known as CiteSeer)
http://citeseer.ist.psu.edu/
http://www.neci.nec.com/~lawrence/researchindex.html

Entrez PubMed, from NLM
Access to NCBI's MeSH, MEDLINE, and journal databases:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

American College of Medical Informatics
http://www.amia.org/acmi/facmi.html

NetDraw, network visualization tool
http://www.analytictech.com/netdraw.htm

Information analysis and visualization demos
SOM and GIS: http://ai.bpa.arizona.edu/go/viz/index.html
SOM: http://www.cis.hut.fi/research/som_pak/
CiteSpace: http://www.pages.drexel.edu/~cc345/citespace/
SPIRE and Themescape: http://nd.loopback.org/hyperd/zb/spire/spire.html

QUESTIONS FOR DISCUSSION

1. What analytical units in addition to authors and documents can be used to examine the state of medical informatics research? What kind of perspectives on the field would these analytical units provide?

2. What is the relationship between citation data and the topology of a knowledge domain? What is the motivation for using such data for domain analysis?

3. What are the advantages of using content analysis over citation analysis for identifying domain subtopics? What are the advantages of using citation analysis over content analysis?

4. How effective are the results of visualization technologies (such as citation networks and self-organizing maps) at presenting domain knowledge in an intuitive way? Are the results informative, easy to understand?