

Chapter 2. Knowledge Mapping: Foundation

Chapter Overview

Invisible college, which consists of a small group of highly productive and networked scientists and scholars, is believed to be responsible for growth of scientific knowledge. By analyzing scholarly publications of these researchers using select content analysis, citation network analysis, and information visualization techniques, “knowledge mapping” helps reveal this interconnected invisible college of scholars and their ideas. In this chapter, we discuss online resources that are often used for such analyses, including: abstracts and indexes, commercial full-text articles and digital libraries, free full-text articles and e-prints, citation indexing systems and services, electronic theses and dissertations, patents, and business and industry articles and reports. These resources can be used to identify important authors and inventors, publications and publication outlets, institutions, countries and regions, and subject and topic areas over time.

2.1 Invisible Colleges and Knowledge Mapping

In Diane Crane’s seminal book on “Invisible Colleges: Diffusion of Knowledge in Scientific Communities” (Crane, 1972), she suggests that it is the “invisible college,” which consist of a small group of highly productive scientists and scholars, that is responsible for growth of scientific knowledge. Crane shows that many scientific disciplines go through similar stages of initiation, growth, expansion, maturation, and decline. The productive scientists and scholars form a network of collaborators in promoting and developing their fields of study. The presence of an invisible college or network of productive scientists linking separate groups of collaborators within a research area has been evident in many studies (Chen, 2003) (Shiffrin & Börner, 2004).

“Knowledge Mapping” or “Science Mapping” techniques, based on content analysis, citation network analysis, and information visualization, has become an active area of research that helps reveal such an inter-connected, invisible college or network of scholars and their seminal publications and ideas.

According to Chaomei Chen in his *Mapping Scientific Frontiers* book (Chen, 2003), Science Mapping helps “depict the spatial relations between research fronts, which are areas of significant activity. Such maps can also simply be used as a convenient means of depicting the way in which research areas are distributed and conveying added meaning of their relationships... By using a series of chronically sequential maps, one can see how knowledge advances. Mapping scientific frontiers involves several disciplines, from the philosophy and sociology of science, to information science, scientometrics, and information visualization.”

In a National Academy of Sciences colloquium entitled “Mapping Knowledge Domains” (Shiffrin & Börner, 2004), the term “mapping knowledge domains” (or Knowledge Mapping) was used to “describe a newly evolving interdisciplinary areas of science aimed at the process of charting, mining, analyzing, sorting, enabling navigation of, and displaying knowledge.”

Two forces are contributing to the rapid development and the overwhelming interest in Knowledge Mapping (we will use the term to encompass Science Mapping in the rest of the book). First, the availability of online publications, from scientific Abstracts and Indexes (A&I), full-text articles, and online pre-prints, to digital dissertations, multimedia (e.g., videos and audios) magazine and journal articles, and multilingual web-accessible patent filings, has made it possible to more systematically examine the scientific outputs produced by members of the invisible colleges. Secondly, the recent advances in text mining, network analysis, and

information visualization techniques have provided more scalable and accurate methods to understand and reveal the interconnections between scientific disciplines and scholars.

Excellent research has also been performed in adopting econometric approach to improving our quantitative knowledge of the sources of economic growth (Pakes & Sokoloff, 1995; Jaffe & Trajtenberg, 2002). Some econometric researchers have used patent citations to investigate the diffusion of technological information across institutions and over time and space (Jaffe & Trajtenberg, 2002). Such works often involve economics-based theoretical modeling, econometric analysis, and parameter estimation and can help draw inferences for knowledge diffusion or spillover. Interested readers are referred to (Pakes & Sokoloff, 1995; Jaffe & Trajtenberg, 2002).

In this chapter, we provide an overview of online research resources that are increasingly available for Knowledge Mapping analysis and were used in the following chapters of this book. We then describe the units of analysis and representation issues of relevance to such resources.

2.2 Online Resources for Knowledge Mapping

Various online resources are available for mapping scientific knowledge. They vary from formal to informal publications; from text-based to multimedia presentations; and from academic literature to industry-relevant international patents.

- *Abstracts & Indexes:* A&I contains abstract and index (bibliographic) information of a given article and is used to locate articles, proceedings, and occasionally books and book chapters in various subjects. Most abstracts and indexes are available electronically. Public and university libraries often subscribe to such databases and services. Only a very few biological or scientific databases are searchable for free on the Web, primarily databases generated by the National Library of Medicine (<http://www.nlm.nih.gov/>), such as MEDLINE (medicine) or TOXLINE (toxicology). There are A&I databases in almost every subject areas, e.g., BIOSIS (biology), COMPENDEX (engineering and technology), ERIC (education), etc.
- *Commercial full-text journal articles and digital libraries:* Many commercial publishers have made their online content available on the web. The most prominent service of such type is provided by the *Web of Science* (<http://scientific.thomson.com/products/wos/>), a product of Thomson Scientific. The *Web of Science* provides seamless access to current and retrospective information from approximately 8,700 research journals in the world. More recently, many professional societies have made their articles available through various digital libraries. For example, the ACM Digital Library (<http://portal.acm.org/dl.cfm>) contains 54,000 online articles from 30 journals and 900 proceedings of the Association for Computing Machinery. The IEEE Computer Society Digital Library (<http://www.computer.org/portal/site/csdl/index.jsp>) provides online access to eighteen IEEE journals and 150 proceedings in computer science.
- *Free full-text articles and e-prints:* There is also a grass-root movement initiated by the academic community to provide free access to journals and books. For example, on the *Free Medical Journals* site (<http://www.freemedicaljournals.com/>), you can find many important academic journals made available online, free and in full-text. *HighWire Press* (<http://highwire.stanford.edu/lists/freart.dtl>), a service affiliated with the Stanford University, is believed to be the largest archive of free full-text science articles. As of December 20, 2006, it provides access to more than 1.5 million free full-text articles in many subject disciplines. In some scientific disciplines, *e-prints* (scientific or technical

documents circulated electronically to facilitate peer exchange, including preprints and other scholarly papers) are strongly encouraged and accepted by the community. For example, the *arXiv.org* service (<http://arxiv.org/>), supported by Cornell University, provides open access to about 400,000 e-prints in Physics, Mathematics, Computer Science, and Quantitative Biology.

- *Citation indexing systems and services*: In addition to accessing bibliographic and full-text content of scientific articles, aggregated and individualized citation information is critical in the assessment of highly-cited, influential papers and authors. The *Science Citation Index* (<http://scientific.thomson.com/products/sci/>), a product of Thomson Scientific, provides access to bibliographic information, abstracts, and cited references in 3,700 world's scholarly science and technical journals covering more than 100 disciplines. A recent service provided by *Google Scholar* (<http://scholar.google.com/intl/en/scholar/>) also supports broad access to scholarly literature. A user can search across many disciplines and sources: peer-reviewed papers, theses, books, abstracts, and articles. The service features many advanced search functionalities, including ranking articles based on how often an article has been cited in other scholarly literature. *CiteSeer* (<http://citeseer.ist.psu.edu/citeseer.html>) is another example of advanced search system (for computer and information science literature) that is built upon citation information. It was one of the first digital libraries to support automated citation indexing and citation linking.
- *Electronic Theses and Dissertations (ETD)*: In addition to formal literature published in journals, magazines, and conference proceedings, Ph.D. and Master's theses and dissertations constitute a significant part of scientific knowledge generated. University Microfilms (UMI) was founded in 1938 to collect, index, film, and republish doctoral dissertations in microfilm and print. Currently UMI's dissertation abstract database has archived over 2.3 million dissertations and master's theses. Some two million of them are available in print, microfilm, and digital format, via its *ProQuest* system (<http://il.proquest.com/brand/umi.shtml>). More recently, the Networked Digital Library of Theses and Dissertations (NDLTD, <http://www.ndltd.org/>) was formed to promote the adoption, creation, use, dissemination, and preservation of electronic analogues to the traditional paper-based theses and dissertations. Via ETD, graduate students learn electronic publishing as they engage in their research and submit their own work, often in a rich multimedia format. Universities learn about digital libraries as they collect, catalog, archive, and make ETDs accessible to scholars worldwide.
- *Patents*: Patent publications have often been used in evaluating science and technology development status worldwide (Narin, 1994). While academic literature represents fundamental scientific knowledge advancement; patents reveal scientific and technological knowledge that has a strong potential for commercialization. There are several governmental or intergovernmental patents offices that control the granting of patents in the world. United States Patent and Trademark Office (USPTO, <http://www.uspto.gov/>), European Patent Office (EPO, <http://www.european-patent-office.org/index.en.php>), and Japan Patent Office (JPO, <http://www.jpo.go.jp/>) issue nearly 90 percent of the world's patents (Kowalski et al, 2003). USPTO handles over 6.5 million patents with 3,500 to 4,000 newly granted patents each week. EPO handles over 1.5 million patents with more than 1,000 newly granted patents each week. JPO handles

over 1.7 million patents with 2,000 to 3,000 newly granted patents each week. All three patent offices provide search systems for web-based access.

- *Business and industry articles and reports*: Critical science and technology knowledge eventually flows from academic literature and patents to various industries and companies. At the other end of the knowledge mapping resources are various business and industry articles and reports; some are reported in general-interest science and technology magazines and newspapers, while others can be purchased from industry-specific consulting firms. For example, timely, in-depth industry-specific or technology-specific reports are available at sites such as: Forrester (<http://www.forrester.com>), IDC (<http://www.idc.com>), and Gartner, (<http://www.gartner.com>), among others.

2.3 Units of Analysis and Representations

For knowledge mapping analysis, pre-processing of raw online resources is needed. Each article, patent, or report needs to be processed to identify key indicators for further analysis and comparison. Among the most common units of analysis for knowledge mapping are: authors or inventors, publications and publication outlets, institutions (companies or universities), countries or regions, subject and topic areas (broad categories or specific topics), and timeline (publication date). These units of analysis are used in the study presented in this book.

- *Authors or inventors*: The most critical unit of analysis for knowledge mapping consists of the researchers, authors, and inventors who are the productive members in the invisible college. Extracting author or inventor field from various knowledge sources is a non-trivial task. Although html, XML, and structured database representations have made automatic name identification easier (than in the paper-based format), author name extraction and identification is difficult in different cultural contexts (e.g., recognizing Chinese names), especially when a publication does not contain complete first and last names. For example, how many different researchers by the name of “W. Zhang” or “L. Liu” are there in the Chinese Academy of Sciences (one of the most productive and largest academic research institutions in the world)?
- *Publications and publication outlets*: Different academic publications have different levels of prestige, most are measured based on their Impact Factor (an aggregate, normalized number based on citation counts). For example, the Impact Factor of *Science* is 30.927 in 2005; while the *Journal of Computational Biology* Impact Factor is 2.446. There are many other publications that do not even have an Impact Factor score. In order to determine the value and impact of a researcher’s work, quality is more important than quantity. Quality is often determined based on the prestige of a publication outlet. In addition, the number of citations is also a major determinant. A seminal or landmark paper can often help define a person’s career or a particular field. For example, while many good academic articles are cited hundreds of times, Albert Einstein’s seminal paper on “Can quantum-mechanical description of physical reality be considered complete?” that appeared in *Physical Review* in 1935 was cited 3,753 times (based on a search on *Google Scholar*). Based on analysis result reported by *ScienceWatch* (<http://www.sciencewatch.com/>), the most cited paper of the past two decades (1983-2002) was: Chomczynski, N. Sacchi, "Single-step method of RNA isolation by acid guanidinium thiocyanate phenol chloroform extraction," *Analytical Biochemistry*, 162(1): 156-9, 1987. The paper received a citation count of 49,562 (based on data from *Thomson Scientific’s* Web of Science). However, correctly parsing and identifying unique

publication names is a difficult task as many databases record those names in cryptic, short-hand forms, e.g., *Analyt. Biochem*, *Proc. natn. Acad. Sci.*, *J. biol. Chem.*, *J. gen. Physiol.*, *Physiol, Lond.*, etc. While many are easily recognizable by domain scientists, a computer program would have difficulty parsing them correctly.

- *Institutions*: While researchers publish their research, it is often the institutions (companies or universities) that they reside own their intellectual property. An analysis based on institutional output and productivity can help depict an institution's relative strength and position in the competitive knowledge landscape. Knowledge mapping can help reveal not just the invisible college of researchers, but the "invisible college of institutions." A comparison between basic university research and applied industry invention can also help understand the progression and impact of knowledge creation.
- *Countries or regions*: Similar to institutional analysis, it is often important to analyze publications (especially patents) based on their countries or regions (e.g., Europe vs. Asia) of origin. Such kind of analysis is useful for depicting a competitive international landscape and is often relied upon for governmental research policy and funding decisions. For example, the US National Nanotechnology Initiative (NNI) has performed excellent cross-regional analyses for worldwide nanotechnology research, development, and funding.
- *Subject and topics areas*: Academics are often defined by their traditional academic boundaries in colleges or departments. However, a researcher could work in several (often related) subject or topic areas. Academic publication outlets are also defined by their fields of interest and focus. While most academic journals provide a list of interested topics; some information resources are more comprehensive in their listing. For example, the USPTO provides a detailed patent classification scheme (USPC), which consists of two levels. The first level contains about 450 categories; while the second contains about 160,000 categories. In addition to these predefined subject categories, important topic-specific keywords, phrases, and concepts can be extracted from the title, abstract, and text body of an article. However, advanced Natural Language Processing (NLP) techniques are needed for such topic identification purposes.
- *Timeline*: All scientific disciplines evolve over time. Most of the online resources for mapping scientific knowledge contain explicit publication dates. Dynamic analysis and visualization of changes in research topics and citation networks could help reveal the advancement in scientific knowledge.

2.4 Questions for Discussion

1. What are other additional informal online resources that could be used for knowledge mapping?
2. How can the Web and social networking sites help promote the "invisible college" of scholars and researchers?
3. How can multimedia contents (e.g., tables, images, photos, audios, and videos) produced by scholars be analyzed to reveal knowledge generated in a scientific discipline?

Chapter 3. Knowledge Mapping: Analysis Framework

Chapter Overview

Three types of analysis are often adopted in knowledge mapping research: *text mining*, *network analysis*, and *information visualization*. Text mining consists of two significant classes of technique: Natural Language Processing (NLP) and content analysis. In NLP, we describe automatic indexing and information extraction techniques that are effective and scalable for concept extraction. In content analysis, clustering algorithms, self-organizing map, multidimensional scaling, principal component analysis, co-word analysis, and PathFinder network are techniques often adopted for knowledge mapping analysis. Network analysis is reviewed based on research in social network analysis (SNA) and complex networks. In SNA, we review research that detects subgroups, discovers patterns of interactions, and identifies roles of individuals. In complex networks, we summarize research in network models, topological properties, and evolving networks. The last section reviews information visualization research of relevance to knowledge mapping. Seven information representation methods are discussed: 1D, 2D, 3D, multi-dimensional, tree, network, and temporal. Two useful user-interface interaction methods, overview + detail and focus + context, are also presented. We believe these knowledge mapping analysis and visualization methods can be applied to most of the online resources presented earlier.

3.1 Text Mining

Text mining, sometimes alternately referred to as *text data mining*, refers generally to the process of deriving high quality [information](#) from text (according to Wikipedia, http://en.wikipedia.org/wiki/Text_mining). Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent integration into a [database](#)), deriving patterns within the structured data, and finally evaluation and interpretation of the output (Chen & Chau, 2004). Typical text mining tasks include entity and relation extraction, [text categorization](#), text clustering, sentiment analysis, and document summarization (Chen, 2001).

For knowledge mapping research, text mining can be used to identify critical subject and topic areas that are embedded in the title, abstract, and text body of published articles. While most structured fields (such as authors, publication outlets, dates of publication, institutions, etc.) can be parsed from online resources, extracting meanings or semantics from multimedia publications requires advanced computational techniques. Different processing algorithms are needed for different media types, e.g., text (Natural Language Processing), image (color, shape, and texture based segmentation), audio (indexing by sound and pitch), and video (scene segmentation). Non-text, multimedia content extraction techniques are still under active research and development. Our discussion will be limited to text-based techniques.

3.1.1 Natural Language Processing:

Automatic indexing: Automatic indexing (Salton, 1989) is a method commonly used to represent the content of a document by means of a vector of keywords or terms. The Bag of Words (BOW) representation has often been used as a baseline implementation for information retrieval and text mining research. When implemented using multi-word matching, a Natural Language Processing noun-phrasing technique can capture a richer linguistic representation of

document. Most noun phrasing techniques rely on a combination of part-of-speech-tagging (POST) and grammatical phrase-forming rules. This approach has potential to improve precision over other word-based document indexing techniques. Examples of noun-phrasing tools include MIT's Chopper, Nptool (Voutilainen, 1997), and Arizona Noun Phraser (Tolle & Chen, 2000).

Information extraction: Information extraction is another computationally effective method to identify important concepts from text documents. It can extract entities of interest (also referred to as entity extraction), such as persons (e.g., "John Doe"), locations (e.g., "Washington, D.C."), and organizations (e.g., "National Science Foundation") and identify relationships between entities. Other entities that are often extracted from unstructured textual narratives include: dates, times, number expressions, dollar amounts, email addresses, and Web addresses (URLs). Such information can be extracted based on either human-created rules or statistical patterns occurring in text. Most existing information extraction approaches combine machine learning algorithms such as neural networks, decision tree, Hidden Markov Model, and entropy maximization with a rule-based or a statistical approach. The best systems have been shown to achieve more than 90% accuracy in both precision and recall rates when extracting persons, locations, organizations, dates, times, currencies, and percentages from newspaper articles (Chinchor, 1998).

While automatic indexing and information extraction techniques are computationally scalable and feasible for large-scale knowledge mapping research, more advanced and fine-grained computational linguistics techniques are being developed in the NLP community. Sentence-level analysis including context-free grammar and transformational grammar can be performed to represent grammatically correct sentences. In addition, semantic analysis based on techniques such as semantic grammar and case grammar can be used to represent semantic (meaning) in sentences and stories. However, most of these full-scale linguistic and semantic analysis techniques lack scalability across different domains and are not yet suitable for large-scale knowledge mapping research (Chen, 2001).

3.1.2 Content Analysis:

Based on automatic indexing or information extraction techniques, documents are often represented as a vector of features (i.e., keywords, noun phrases, or entities). Articles that are collected and grouped based on authors, institutions, topic areas, countries, or regions can be analyzed to identify the underlying themes, patterns, or trends. Popular content analysis techniques include: *Clustering Algorithms*, *Self-Organizing Map (SOM)*, *Multidimensional Scaling (MDS)*, *Principal Component Analysis (PCA)*, *Co-Word Analysis*, and *PathFinder Network*.

Clustering Algorithms: Everitt (Everitt, 1974) defines a cluster as "a set of entities which are alike, and entities from different clusters are not alike." Clustering algorithms are used to organize (group) similar documents or topics in a hierarchical structure. There are two types of hierarchical cluster analysis: agglomerative and divisive. The agglomerative approach starts with each point as a separate cluster. Each point is merged successively into a larger cluster based on their degree of similarity. Conversely, divisive hierarchical cluster analysis begins with only one large cluster of points. The large cluster is divided successively into smaller clusters based on their degree of similarity. Both approaches produce a dendrogram to represent a hierarchy of points and their associated clusters. Hierarchical agglomerative clustering (HAC) algorithms are the most commonly used method for document clustering (Willett, 1988). One of the most popular HAC algorithms is Ward's clustering (Ward, 1963). Over time, it has been widely used

in various domains: astrophysics, pattern recognition, applied statistics, etc. in 1984, Murtagh proposed the Reciprocal Nearest Neighbor (RNN) approach (Murtagh, 1995), which is significantly faster than the original Ward's implementation. The time complexity of the algorithm was reduced from $O(N^3)$ to $O(N^2)$. Roussinov and Chen (Roussinov & Chen, 1999) performed a systematic comparison of the RNN-based Ward's algorithm with the SOM technique for document clustering. Both techniques are computationally efficient for large-scale knowledge mapping applications.

Self-Organizing Map (SOM): The Self-Organizing Map, developed by Kohonen (Kohonen, 1989; Kohonen, 1995), is an unsupervised, two-layered neural network used for clustering and dimension reduction. An advantage of SOM over other clustering algorithms is its ability to visualize high dimensional data using a two-dimensional grid while preserving similarity between data points. It is a technique similar to Multi-Dimensional Scaling (MDS) (Jain & Dubes, 1988). In SOM, each input node corresponds to a dimension (e.g., a keyword). Each output node corresponds to a node in a two-dimensional grid. The network is full connected in that every output node is connected to every input node with some connect weight. During the training phase, the inputs (e.g., documents represented as vectors of keywords) are presented several times to the SOM to decide the proper placement of the inputs on the output grid. Connection weights associated with the input and output nodes are adjusted (learned) to ensure that similar inputs are grouped in a close proximity on the two-dimensional grid. After training, all inputs (e.g., documents) can be grouped and placed on a two-dimensional map. Topics (of similar documents) are often represented as regions on the map, where larger regions represent more important topics. The SOM-generated categories were found to be comparable to those generated by human subjects (Orwig, et al., 1997). Chen and his team (Chen et al., 1996) developed a multi-layered SOM (ET-Map) to categorize 110,000 Internet web pages according to their content. Kohonen and his colleagues (Kohonen, et al., 2000) adopted SOM to map 6.8 million patent abstracts onto a one million-node SOM. Several SOM implementations have taken advantage of the sparse input feature vector representation to improve the speed and scalability of their algorithms.

Multidimensional Scaling (MDS): Multidimensional Scaling and Principal Component Analysis (PCA) are two classical and widely used techniques for dimension reduction. They are well known, easy to implement, and computationally efficient. One way to apply MDS is to take a set of p -dimension vectors and to approximate them on the two- or three-dimensional Cartesian coordinate space. Beginning with the matrix of distance or dis-similarity between objects, we first compute an initial configuration using Singular Value Decomposition (SVD) (Cox & Cox, 1994; Forsythe, et al., 1977; McQuaid, et al., 1999). We then measure the information loss between the original matrix and the initial configuration using the STRESS (standardized residual error sum of squares) metric of Kruskal (Kruskal, 1964). It then finds a new configuration with smaller information loss than the initial configuration by using an isotonic regression algorithm (Grotzinger & Witzgall, 1984) to obtain fitted distances and a conjugate gradient descent algorithm to optimize. The algorithm repeats until a threshold information loss is reached or until a threshold number of iterations is performed. Similar to SOM for knowledge mapping, MDS produces a two-dimensional display which agrees with human perception of document similarity (McQuaid, et al., 1999). Also based on SVD, Latent Semantic Indexing (LSI) uses a [term-document matrix](#), which describes the occurrences of terms in documents (Deerwester, et al., 1990). It then transforms the original matrix into a relationship between the terms and (latent) *concepts*, and a relation between the documents and the same concepts. The

terms and documents are now indirectly related through the concepts. LSI can be used effectively for concept extraction and association for knowledge mapping.

Principle Component Analysis (PCA): Principal component analysis is central to the study of multivariate data. Although one of the earliest multivariate techniques, it continues to be the subject of much research, ranging from new model-based approaches to algorithmic ideas from neural networks. It is extremely versatile with applications in many disciplines (Jolliffe, 2002). The central idea of PCA is to reduce the dimensionality of a data set in which there are a number of interrelated variables, while retaining as much as possible the variation present in the data set. The reduction is achieved by transforming to a new set of variables, the principal components, which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables. Computation of the principal components reduces to the solution of an eigenvalue-eigenvector problem for a positive-semidefinite symmetric matrix (Jolliffe, 2002). In knowledge mapping, PCA can be used to extract principal components that represent interrelated keywords or topics.

Co-word Analysis: Word-occurrence patterns in text originated in the co-word analysis method developed in the 1980s (Callon et al., 1986). The outcome of co-word analysis is typically depicted as a network of concepts. Given a corpus of N documents, each document can be indexed by a set of unique keywords or terms. If two terms, t_i and t_j , appear together in multiple documents, their probability of co-occurrence can be computed using different formulas. A matrix of term co-occurrence defines a network of concepts. In some past research, such a matrix is referred to as a *concept space* (Chen et al., 1996; Chen et al., 1997). The original co-word analysis prunes a concept using a triangle inequality rule on conditional probabilities. If a shorter path can be found from term t_i to t_j than the direct path, then the shorter one is chosen. By raising or lowering a threshold, the number of valid links in a network can be decreased or increased.

PathFinder Network (PFNET): According to Börner et al., (Börner, et al., 2003), “PathFinder algorithms take estimates of the proximities between pairs of terms as input and define a network representation of the items that preserves only the most important links.” The input is pairs of terms, and the pairs are linked as output only if their co-occurrence weights are the highest in their respective vectors. By emphasizing only the most prominent links, PFNET reduces the user’s cognitive overload in browsing a network of interrelated concepts. (White, et al., 2003). In PFNET, paths are required not to violate the triangle inequality $d(a,c) \leq d(a,b) + d(b,c)$, where d is the distance between points a , b , and c . A spring embedder algorithm (Kamada & Kawai, 1989) is often used to display the network of terms based on the relative strengths between any pairs of terms. The algorithm aims to provide a layout that avoids crossed links and overlapping nodes. White et al. (White et al., 2003) adopted both SOM and PFNET in creating “localized” mapping of 24 most relevant terms given a single input term, a medical subject heading, a co-cited author, or a co-cited journal from the *Proceedings of the National Academy of Sciences (PNAS)*, 1971-2002. Mane and Börner (Mane & Börner, 2003) adopted Klienberg’s burst detection algorithms, PFNET, and graph layout techniques to generate maps that support the identification of major research topics and trends in *PNAS*, 1982-2001. Both the general co-word analysis and the PFNET algorithm implementation have been shown to be valuable for mapping scientific knowledge.

3.2 Network Analysis

Recent advances in *social network analysis* and *complex networks* have provided another means for studying the network of productive scholars in the invisible college.

3.2.1 Social Network Analysis:

A collection of methods that are recommended in literature for studying networks is Social Network Analysis (SNA) techniques (McAndrew, 1999; Sparrow, 1991; Xu & Chen, 2005a). Because SNA is designed to discover patterns of interactions between social actors in social networks, it is especially apt for co-authorship network analysis. Specially, SNA is capable of detecting subgroups (of scholars), discovering their pattern of interactions, identifying central individuals, and uncovering network organization and structure. It has also been used to study criminal networks (Xu & Chen, 2005a; Xu & Chen, 2005b).

Subgroup Detection: A collaboration or co-authorship network can be partitioned into subgroups consisting of individuals who closely interact with each other. Given a network, traditional data mining techniques such as cluster analysis may be employed to detect underlying groupings that are not otherwise apparent in the data. Burt (Burt, 1976) applied hierarchical clustering methods based on structural equivalence measure (Lorrain & White, 1971) to detect subgroups in a social network. Two nodes are structurally equivalent if they have identical links to and from all other nodes in the network. Since perfectly equivalent nodes rarely exist in reality, this measure is relaxed to be an indicator of extent to which two nodes are equivalent. With structural equivalence measures between nodes, a hierarchical clustering algorithm partitions a network into subgroups so that members within a group are more similar to each other and members belonging to different groups are more different from each other. Cliques whose members are fully or almost fully connected can also be detected based on clustering results.

Discovery of Pattern of Interactions: Patterns of interactions between subgroups can be discovered using an SNA approach called *blockmodel analysis* (Wasserman & Faust, 1994; Xu & Chen, 2005b). This approach was originally designed to interpret and validate theories of social structures. When used in collaboration or co-author network analysis, it can discover patterns of inter-group interactions and associations, and help reveal the overall structure of networks under study. Blockmodeling usually follows clustering, and then determines the presence and absence of an association between a pair of subgroups by comparing the density of the links between these two subgroups with a predefined threshold value. When the link density is greater than the threshold value, an inter-group association presents, indicating that these two subgroups interact with each other constantly and thus have a strong association. Using blockmodeling, a complex network is reduced to a simpler structure by summarizing individual interaction details into interactions between groups, so that the overall structure of the network becomes more salient.

Roles of Individuals: Centrality deals with the roles of individuals in a network. Several measures, such as *degree*, *betweenness*, and *closeness*, are related to centrality (Wasserman & Faust, 1994). The degree of a particular node is the number of direct links it has; its betweenness is the number of geodesics (shortest paths between any two nodes) passing through it; and the closeness is the total number of all the geodesics between that particular node and every other node in the network. Although these three measures are all intended to describe the importance or centrality of a node, they have different interpretations for the roles network members play.

An individual with high degree, for instance, may implies his or her leadership; while an individual with high betweenness may be a gatekeeper or connector in the network.

3.2.2 Complex Networks:

Complex networks of individuals and other entities have been traditionally studied under the random graph theory (Albert & Barabasi, 2002). However, later studies suggested that real-world complex networks (such as collaboration or co-authorship networks) may not be random but may be governed by certain organizing principles. This prompted the study of real-world networks. These studies have explored the topology, evolution and growth, robustness and attack tolerance, and other properties of networks.

Network Models: Three broad models of network topologies have emerged (Albert & Barabasi, 2002): *random graphs*, *small-world* networks, and *scale-free* networks. Random graphs are networks in which any two nodes are connected with a fixed probability p , thus edges are distributed randomly among nodes of the network. Small-world networks are not random networks and have relatively small path lengths despite their often large size (Watts & Strogatz, 1998). In scale-free networks the degrees (number of edges) of nodes follow a power law distribution (Barabasi & Albert, 1999). Some of the networks that have been studied include the World Wide Web (Albert et al., 1999; Kumar et al., 2000), citation networks (Jeong, et al., 2000), and co-authorship networks (Newman, 2001a). These networks were found to have similar topological, evolutionary and robustness characteristics (Albert & Barabasi, 2002). They were found to be predominantly small-world and scale-free.

Topological Properties: Topological properties of networks help us study the network as a whole instead of studying the individual constituents. Three concepts dominate the statistical study of the topology of networks: *small-world*, *clustering*, and *degree distribution* (Albert & Barabasi, 2002).

Small-world: The small-world concept is based on the fact that large networks often have small path lengths between their nodes. This concept is an important one as it can depict the ease of communications within a network. Communications can range from the spread of disease in human populations to the spread of ideas in a collaboration network. A widely cited example of a small-world network study is the “six degrees of separation” study by psychologist Stanley Milgram, who concluded that there was a path of acquaintances with a typical length of about six between most pairs of people in the United States (Kochen, 1989). The small world property is measured by the average shortest path length that is obtained by averaging the shortest paths between all pairs of nodes in a network (Albert & Barabasi, 2002). For instance, the average shortest path length between two actors in a network of movie actors (225,226 nodes) was found to be 3.65 (Watts & Strogatz, 1998). The average shortest path length between co-authors in the MEDLINE collection (1.5 million nodes) was found to be 4.6 (Newman, 2001a). There has been research on the phenomenon that leads to the short path lengths in real world networks. It has been suggested that shortcuts between nodes that otherwise may not be connected reduce the average path length in small world networks (Watts, 1999). This is especially true in social networks where people are likely to have friends with other individuals outside their immediate friend circle.

Clustering: Cliques that represent circles of friends and acquaintances often form in social networks. For instance, authors often collaborate with the same set of people in a co-authorship network. Cliques also form in networks that do not involve people, for example, related websites on the Web often point to each other through hyperlinks. This inherent tendency to cluster is quantified by the clustering coefficient (Watts & Strogatz, 1998). The clustering coefficient is

measured by the ratio of the number of edges that exist in a network to the total number of possible edges (Albert & Barabasi, 2002). Real-world networks tend to have relatively high clustering coefficients as compared to random graphs. The movie actor network had a clustering coefficient of 0.79 (Watts & Strogatz, 1998) and the MEDLINE co-authorship network had a coefficient of 0.066 (Newman, 2001), both values are several orders of magnitude higher than their random counterparts.

Degree distribution: Nodes in a network have different number of edges connecting them. The number of edges connected to a node is called its degree. The spread of node degrees is given by a distribution function $P(k)$, which gives the probability that a randomly selected node has exactly ' k ' edges (Albert & Barabasi, 2002). The distribution functions of most real world networks follow power law scaling with exponents ranging from 1.0 to 3.0 (Albert & Barabasi, 2002). The movie actor network has a power law degree distribution with an exponent of 2.3 (Watts & Strogatz, 1998). The MEDLINE co-authorship network was found to have an exponent of 1.2 (Newman, 2001a). The degrees of nodes are also used to study the growth and evolution of a network.

Using data from three bibliographic databases in biology, physics, and mathematics, networks are constructed in which nodes are scientists, and two scientists are connected if they have coauthored a paper (Newman, 2003). Newman uses various network topological properties to answer a board variety of questions about collaboration patterns, such as the number of papers authors write, how many people they write them with, what the typical distance between scientists is in the network, and how patterns of collaboration vary between subjects and over time.

Evolving Networks: Most real-world networks are not static and grow due to the addition of nodes and/or links. For instance, the World Wide Web grows exponentially by the addition of new web pages and a co-authorship network grows by the addition of collaborators. The growth leads to changes in the topological characteristics of the networks. Albert and Barabasi (Albert & Barabasi, 2002) identified two factors in the evolution of a scale free network: (1) growth: networks expand continuously by adding new nodes and, (2) preferential attachment: new nodes attach preferentially to nodes that are already well connected, an effect called "rich-get-richer." The preferential attachment concept assumes that the probability that a new node will connect to an existing node i depends on the degree of the node i . The higher the degree of i , the higher the probability that new nodes will attach to it. The functional form of preferential attachment ($\Pi(k)$) for a network can be measured by observing the nodes present in the network and their degrees at a particular time, t . After adding new nodes (time = $t+1$), plotting the relative increase as a function of the earlier degree gives the $\Pi(k)$ function (Jeong, et al., 2003). Preferential attachment has been studied for citation and co-authorship networks, actor network and the Internet and has been found to follow the power law distribution (Jeong, et al., 2003; Newman, 2001b). In other cases $\Pi(k)$ may grow linearly till a point and then fall off. This usually happens at high degrees implying that high degree nodes are unable to attract new nodes. For instance, Newman (Newman, 2001b) found that individuals with a large number of collaborators in a co-authorship network did not attract many new ones.

Constraints on the number of links that a node can attract may be due to aging or cost (Amaral, et al., 2000). Since the growth of the network may be over time, some high degree nodes might become too old to participate in the network (e.g., actors in a movie network). It might also become too costly for a node to attach to a large number of nodes (e.g., a router in a network slows down when it has too many connections). Constraints on the growth may be

domain specific and have been studied in many domains. For instance, in plant-animal pollination networks, some animals cannot pollinate certain plants; hence a link cannot be established (Jordano, et al., 2003). This is an example of a cost constraint. In criminal networks, trust may restrict the growth of networks. Criminal and terrorists do not include many people in their inner trust circle (Klerks, 2001). In addition, disruption might restrict growth in criminal networks. Individuals may get jailed, wounded or killed and thus not contribute to the growth (Xu & Chen, 2005b).

3.3 Information Visualization

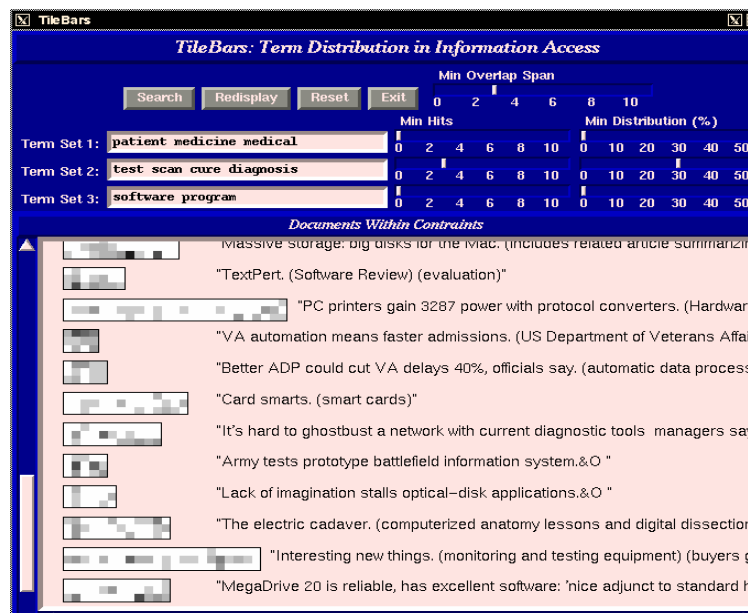
The last step in the knowledge “mapping” process is to make knowledge transparent through the use of various information visualization (or mapping) techniques. *Information representation* and *user-interface interaction* are two dimensions often considered in information visualization research (Zhu & Chen, 2005).

3.3.1 Information Representation:

Shneiderman (1996) proposed seven types of information representation methods including the *1D (one-dimensional)*, *2D*, *3D*, *multi-dimension*, *tree*, *network*, and *temporal* approaches. We use this framework as the basis to review related research and present selected examples.

1D representation: The 1D approach represents abstract information as one-dimensional visual objects and displays them on the screen in a linear or a circular manner (Eick et al, 1992; Hearst, 1995). 1D representation has been applied to display either the contents of a single document (Hearst 1995) or an overview of a collection of documents (Eick et al., 1992). Colors usually represent some attributes of each visual object. For instance, colors indicate type of document in the SeeSoft system (Eick et al., 1992) and depict the location in a document of search terms in Tilebar (Hearst, 1995). Figure 3-1 displays an interface of the tile bar that shows the occurrence of search terms in documents. The darkness of each tile indicates the frequency of a search term in a document.

Figure 3-1 TileBar uses 1D representation to show the term-document relevance (http://www.acm.org/sigchi/chi95/Electronic/documnts/papers/mah_fg4.gif @ 1995 ACM, Inc.)



2D representation: A 2D approach represents information as two-dimensional visual objects. Visualization systems based on 2D output of a self-organizing map (SOM) (Kohonen, 1995; Chen et al., 1996; Huang et al., 2003; Huang, et al., 2004) belong to this category. Such systems display categories created over a large collection of textual documents, with the layout of each category based on its location in the two-dimensional output of the SOM. Spatial proximity on the interface represents the semantic proximity of the categories created. The challenge in this approach is to help users deal with a large number of categories that have been created for the large volume of textual data. The CancerMap system described in Chen et al. (2003) adopted the SOM and the Arizona Noun Phraser (Tolle and Chen, 2000) to generate a subject map automatically. Figure 3-2 presents two consecutive screen shots, displaying the top-level categories and sub-categories under the category of “Liver Neoplasm.” The empirical study described in Chen et al. (2003) indicates that this approach generated a meaningful subject hierarchy to supplement or enhance human-generated hierarchies in digital libraries.



Figure 3-2. Example for 2D representation: The Interface of CancerMap. Category “Liver Neoplasm” was selected at the top level and the sub-map of “Liver Neoplasm” was displayed

3D representation: A 3D approach represents information as three-dimensional visual objects. One example is the WebBook system (Card, et al., 1996) that folds web pages into three-dimensional books. Realistic metaphors such as rooms (Card, et al., 1996), bookshelves (Card, et al., 1996), or buildings (Andrews, 1995) are employed to depict abstract information. Visualization systems using 3D version of a tree or network representation also belong to this

category. One example is the 3D hyperbolic tree created by (Munzner, 2000) to visualize large-scale hierarchical relationships. Figure 3-3 shows screenshot of WebBook, where the book metaphor is applied to organize web pages from the same web site.

Figure 3-3. Example for 3D representation: The WebBook
(<http://acm.org/sigchi/chi96/proceedings/papers/Card/skc1txt.html>,
©1996 ACM, Inc.)



Multi-dimensional representation: The multi-dimensional approach represents information as multi-dimensional objects and projects them into a three-dimensional or a two-dimensional space. This approach often represents textual documents as a set of key terms that identify the theme of a textual collection. A dimensionality reduction algorithm such as multidimensional scaling (MDS), hierarchical clustering, principal components analysis (PCA), or self-organizing map (SOM) is used to project document clusters or themes into a two-dimensional or three-dimensional space. The SPIRE (Spacial Paradigm for Information Retrieval and Exploration) system presented in Wise et al. (1995) and the VxInsight system in (Boyack et al., 2002) belong to this category. Figures 3-4 and 3-5 display two types of visualization developed for the SPIRE system. The Galaxy (Figure 3-4) clusters 567,437 abstracts of cancer literature based on the semantic similarity, whereas the ThemeView (Figure 3-5) visualizes relationships among topics of a collection of document. Glyph representation, another type of multi-dimensional representation, uses graphical objects or symbols to represent data through visual parameters that are spatial (positions x or y), retinal (color and size), or temporal (Chernoff, 1973). It has been applied in various social visualization techniques to describe human behavior during computer-mediated communication (CMC) (Zhu & Chen, 2001; Donath 2002).

Figure 3-4. Example for multi-dimensional representation: Galaxy visualization of text documents (http://www.pnl.gov/infviz/gal_cancer800.gif, reprinted with permissions)

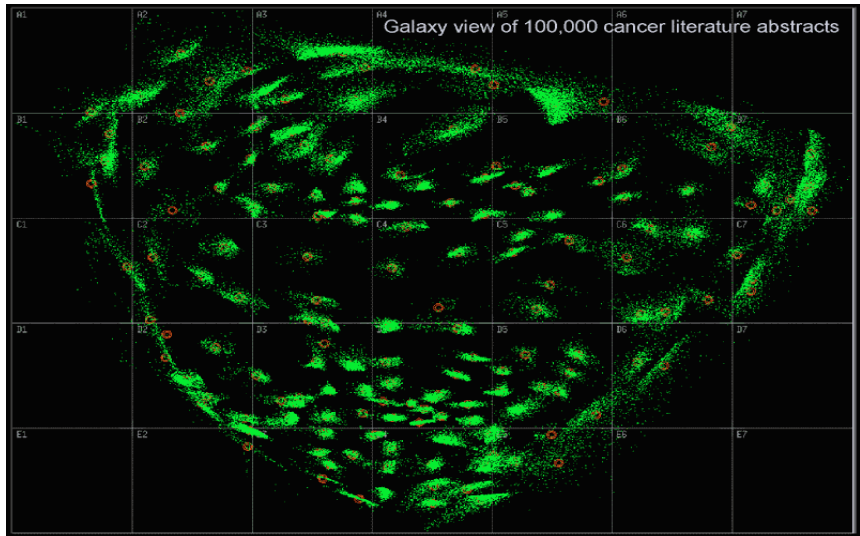
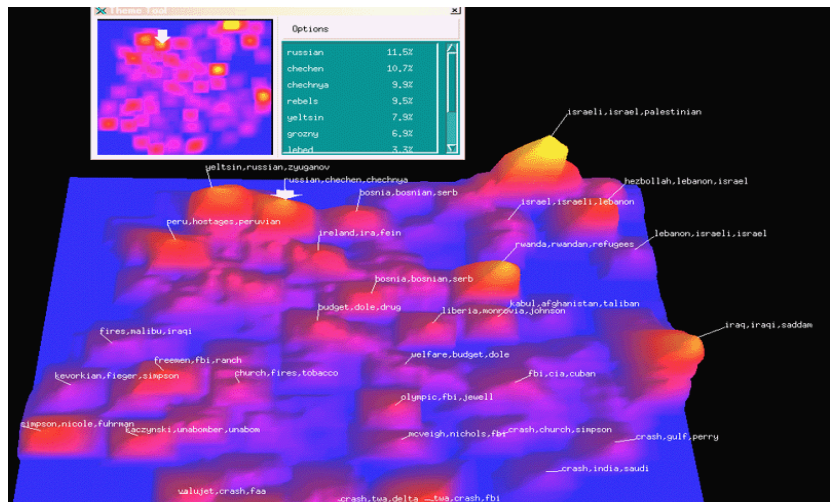


Figure 3-5. Example for multi-dimensional representation: ThemeView - The height of a peak indicates the strength of a given topic in the collection of documents. (http://www.pnl.gov/infviz/theme_cnn800.gif, reprint with permission).



Tree representation: The tree approach is often used to represent hierarchical relationships. The most common example is an indented text list. Other tree-based systems include the Tree-Map (Johnson & Shneiderman, 1991), the Cone Tree (Robertson et al., 1991), and the Hyperbolic Tree (Lamping et al., 1995). One crucial challenge to this approach is that the number of nodes grows exponentially as the number of tree levels increases. As a consequence, different layout algorithms have been applied. For instance, the Tree-Map (Johnson & Shneiderman, 1991) allocates space according to attributes of nodes, while the Cone Tree

(Robertson et al., 1991) takes advantage of the 3D visual structure to pack more nodes on the screen. Figure 3-6 displays the visual interface of the Cat-a-Cone system (Hearst and Karadi, 1997) that applies the 3D Cone Tree to visualize hierarchies in Yahoo. The Hyperbolic Tree (Lamping et al., 1995), on the other hand, projects sub-trees on a hyperbolic plane and puts the plane into the range of display. 3D version of the hyperbolic tree has also been developed by Munzner (2000) to visualize large-scale hierarchies (Figure 3-7).

Figure 3-6. Example for tree representation: Cat-a-cone tree that displays hierarchies in Yahoo. (<http://www.sims.berkeley.edu/~hearst/cac-overview.html>, @ 1997 ACM, Inc.)

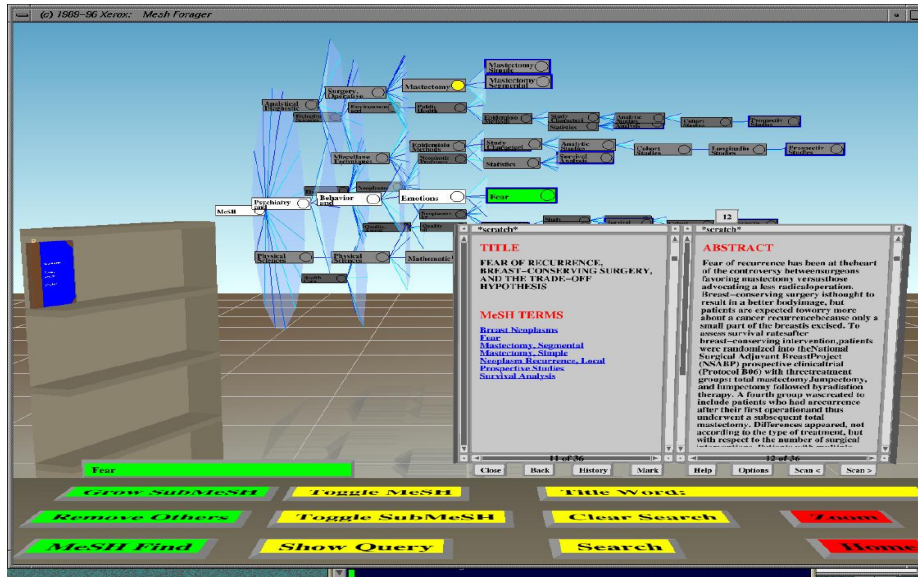
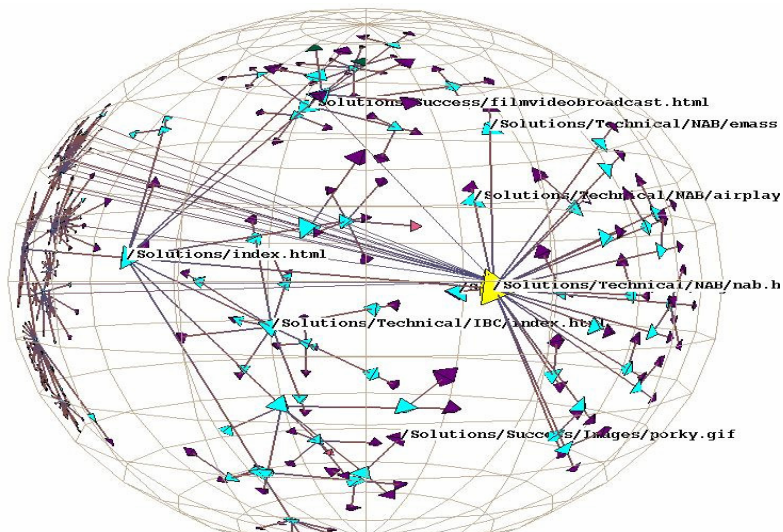
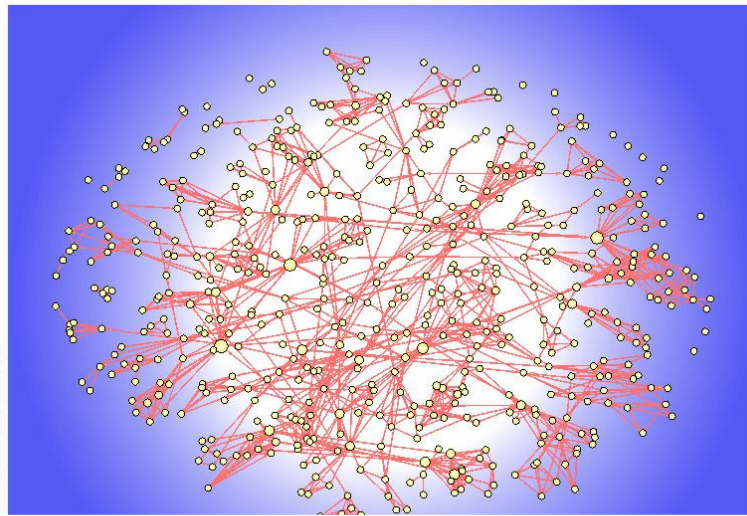


Figure 3-7. Example for network representation: 3D hyperbolic space (http://graphics.stanford.edu/papers/munzner_thesis/hyp-figs.html, reprint with permission of Tamara Munzner)



Network representation: The network representation method is often applied when a simple tree structure is insufficient for representing complex relationships. Complexity may stem from citations among many academic papers (Mackinlay et al. 1995; Chen & Paul, 2001) or from interconnected web pages on the Internet (Andrews, 1995). Among various network visualizations that have been created, the spring embedder algorithm, originally proposed by Eades (1984), and its variants (Kamada & Kawai, 1989), have become the most popular graph drawing algorithms. Figure 3-8 presents the visualization of co-authorship relationships among 555 scientists using a spring embedder algorithm.

Figure 3-8. Example of network representation: Visualization of a large co-authorship network (<http://www.mpi-fg-koeln.mpg.de:80/~lk/netvis/Huge.html>, reprint with permission of Lothar Krempel)



Temporal representation: The temporal approach visualizes information based on temporal order. Location and animation are two commonly used visual variables to reveal the temporal aspect of information. Visual objects are usually listed along one axis according to the time when they occurred, while the other axis may be used to display the attributes of each temporal object (Eick et al., 1992; Robertson, et al., 1993). For instance, the Perspective Wall (Robertson, et al., 1993) lists objects along the x-axis based on time sequence and presents attributes along the y-axis. Animation is another effective way to display temporal information.

The seven types of representation method turn abstract textual documents into objects that can be displayed. A visualization system usually applies several methods at the same time. For instance, the multi-level ET-Map system created by Chen et al. (Chen, et al., 1998) combines both 2D and the tree structure, where a large set of web pages are partitioned into hierarchical categories based on their content. While the entire hierarchy is organized in a tree structure, each node in the tree is a two-dimensional SOM, on which the sub-categories are graphically displayed.

Many powerful visualization methods also require advanced analysis techniques. For example, the TileBar system (Hearst, 1995) employs the text-tiling analysis algorithm to segment a document, while ThemeView and Galaxy (Wise et al., 1995) use multidimensional scaling (MDS) to cluster and lay out documents on the screen.

3.3.2 User-Interface Interaction:

The “small screen problem” (Robertson et al., 1993) is common to representation methods of any type. To achieve effectiveness, an effective information representation method needs to be integrated with user-interface interaction. Recent advances in hardware and software allow quick user-interface interaction, and various combinations of representation methods and user interface interactions have been employed.

Interaction between an interface and its users not only allows direct manipulation of visual objects displayed, but also allows users to select what is to be displayed and what is not (Card et al., 1999). The two commonly used interaction approaches are: *overview + detail* and *focus + context* (Card et al., 1999).

Overview + Detail: Overview + detail provides multiple views, with the first being an overview, providing overall patterns to users. Details about only the part of user’s interest can then be displayed. These two views can be displayed at the same time or separately. When a detailed view is needed, two types of zooming are usually involved (Card et al., 1999): spatial zooming and semantic zooming. Spatial zooming refers to the process of enlarging selected visual objects spatially to obtain a closer look, whereas semantic zooming provides additional content about a selected visual object by changing its appearance.

Focus + Context: The focus + context technique provides detail (focus) and overview (context) dynamically on the same view. One example is the 3D perception approach adopted by systems like Information Landscape (Andrews, 1995) and Cone Tree (Robertson et al., 1991), where visual objects at the front appear larger than those at the back. Another commonly used focus + context technique is the fisheye view (Furnas, 1986), a distortion technique that acts like a wide-angle lens to amplify part of the focus. The objective is to simultaneously provide neighboring information in reduced detail and supply greater detail on the region of interest. In any focus + context approach, users can change the region of focus dynamically. A well-known visualization system that applies the fisheye technique is the Hyperbolic Tree (Lamping et al., 1995), in which users can scrutinize the focus area and scan the surrounding nodes for a big picture. Other focus + context techniques include filtering, highlighting, and selective aggregation (Card et al., 1999).

The overview + detail and focus + context user-interface interaction approaches could support knowledge navigation in various knowledge mapping applications.

The presented analysis framework is applied in this book to nanotechnology knowledge mapping.

3.4 Questions for Discussion

- 1 What are other new analysis and visualization techniques that are promising for large-scale knowledge mapping research?
- 2 How can we adopt selected multimedia indexing techniques (for images and audios) in content analysis and knowledge mapping?
- 3 How can we adopt selected knowledge mapping techniques to study the informal invisible college of scholars using informal online resources (e.g., via forums and web blogs)?
- 4 How can we develop an interactive, user-controlled, highly-visualized system for knowledge mapping using different online resources?