VISUALIZATION OF KNOWLEDGE STRUCTURES

CHAOMEI CHEN

The VIVID Research Centre

Department of Information Systems and Computing

Brunel University, Uxbridge UB8 3PH, UK

United Kingdom

Tel: +44 1895 203080

Fax: +44 1895 251686

E-mail: chaomei.chen@brunel.ac.uk

Abstract

Tracking the growth of scientific knowledge has become increasingly challenging even in one's own specialized field due to the vast amount of new scientific publications become available. As a rapidly advancing and expanding field of computing and information technology, information visualization has focused on the discovery of interrelationships among various scientific publications. However, visualizing intrinsic structures among documents in scientific literatures can only capture some aspects of scientific knowledge. For example, the number of citations received by a scientific work is a widely accepted hallmark of its significance. This chapter describes approaches to the visualization of knowledge structures with emphasis on the role of citation-based methods. Instead of relying upon occurrence patterns of content-bearing words, visualization of knowledge structures aims to capture perceived intellectual structures of a particular knowledge domain. An ultimate goal for the visualization of knowledge structures is to provide scientists with a tool that can detect the existence of a scientific paradigm and movements of such paradigms. This chapter also includes a summary of the history of tracking the growth of scientific knowledge. The state of the art is presented to highlight the trend of future research.

Keywords: Knowledge visualization; intellectual structures; domain analysis.

1. Introduction

Dictionaries explain *knowledge* at two levels: 1) understanding, the facts, information, and skills that one has gained, and 2) what is collectively known to the human being. To philosophers, this is to ask "what is the nature of knowledge?" Epistemology, a branch of philosophy, studies the nature of knowledge. An understanding of the nature of knowledge has a profound connection to an understanding of the nature of science. Philosophy of science aims to find an answer to this question. Approaches to the visualization of knowledge structures aim to reveal insightful patterns of intellectual structures shared by scientists in a subject domain. In order to visualize and interpret structures of knowledge, we must have a clear understanding of contemporary views on the nature of knowledge in a broader context.

1.1. What is Knowledge?

Thomas Kuhn's landmark work *Structure of Scientific Revolutions* was published in 1962 [18]. He used the term *paradigm* to refer to the conceptual frameworks and worldviews of various scientific communities. In a nutshell, a paradigm represents a prevalent way of thinking. Science constitutes periods alternate between normal science, crisis, and revolution. During periods of normal science, the way of thinking is relatively stable, but in revolutionary periods a traditional way of thinking may be replaced by a new way of thinking, which could be radically different from its predecessor. This movement from one paradigm to another is widely known as a *paradigm shift*.

Widely known paradigm-shift examples in the past include Copernicus's revolutionary theory in contrast to Aristotle's model of the world and Einstein's general relativity theory in contrast to Newtonian physics. Before the Copernican revolution, it was believed the solar system and it should be perfect in many ways. For example, the earth was believed to be at the center of the solar system and the Sun was orbiting the earth, and that each planet has a perfect circular orbit. Much of these also appeared to be consistent with one's own observation. It was through several generations that the Copernican revolution started to change ones' beliefs to what we know today: the earth and other planets circle around the Sun in elliptical orbits. According to Kuhn, the two competing paradigms are incommensurable, nor are the concepts that can be used to understand and explain basic facts and beliefs. The two groups live in different worlds.

Despite critics of the notion of paradigm shifts, Kuhn's work has had enormous influence outside of the history of science. The profound and far-reaching significance of Kuhn's insight is akin to what Darwin's natural selection to evolutionary biology.

Kuhn emphasized that the paradigm determines the kinds of experiments that scientists perform, the types of questions they ask, and the problems they consider important. A shift in the paradigm alters the fundamental concepts underlying research and inspires new standards of evidence, new research techniques, and new pathways of theory and experiment that are radically incommensurate with the old ones.

Kuhn's structure of scientific revolution provides a far-reaching framework for visualizing knowledge structures. In this chapter, we will explain how knowledge structures can be visualized and interpreted in the notion of paradigm and paradigm shift.

1.1. What is Knowledge Visualization?

Information visualization has become a truly wide-ranging and interdisciplinary field of research and a vibrant global industry. There is a rapid growth in the literature of information visualization [2, 3]. Knowledge visualization shares some intrinsic characteristics with cartography – an art of making maps. The evolution of the geographic map of the world is a good starting point for us to illustrate what we might need to produce a map of knowledge structure and what we may expect given our knowledge of the nature of science. We will first look at the Greek astronomer Ptolemy's map of the world in nearly 2000 years ago, and then look at a cybermap – a map of the Internet. The evolutionary history of these maps provides us an interesting yardstick to measure where we are regarding the status of visualizing knowledge structures.

1.1.1. Mapping the Real World

The British Library has a collection of famous old world maps[1]. One of the widely known world maps is the one generated by the Greek astronomer Ptolemy in about 150 AD. Unfortunately, none of his maps survived and his work was lost to the West until the Renaissance. Scholars in the 15th century recreated Ptolemy's map following the

[1] http://portico.bl.uk/exhibitions/maps/

instructions in his work *Geography*, which explain how to project a sphere onto a flat piece of paper using a system of gridlines - longitude and latitude.

Ptolemy's map only shows the world known to him, and as he worked in Alexandria the map is most detailed round the Mediterranean. The map shows only three continents: Europe, Asia and Africa. The two red lines are the Tropics of Cancer and Capricorn; the sea is light brown, the rivers are in blue and the mountains in dark brown. The surrounding heads represent the major winds.

1.1.1. Cybermaps

The ever-growing Internet has flourished cybermaps. The idea of making a map of Internet has certainly been very popular. A wide variety of cybermaps have been produced. The following map is generated by the peacockmaps company on the millennium. The inset shows a close-up view of a region in the global map. Blue stars and lines represent *.net* nodes and backbones, which are typically non-profitable organizations, two green stars are large European nodes, and dark red lines throughout the map show North American hosts.

From the old world maps to the latest image of the Internet, human mind has been searching something profoundly in common – a big picture of an environment in which we live and work. For the same reason, a big picture of science has always been the most intuitive pursuit of a fascinating dream. Visually representing the structure of knowledge is thus a very attractive idea.

One fundamental aspect of information visualization in general, and knowledge visualization in particular, is how to identify and group various knowledge-bearing units according to certain criteria. Clustering and classification algorithms form a substantial amount of the resources of knowledge visualization. The following section summarizes some of the major theories and methodologies in measuring the relevancy of textual documents and in extracting the most significant elements from the source.

2. A Historical Summary

A number of existing information visualization systems are particularly worth mentioning since they address common issues concerning knowledge visualization, for example, selecting appropriate similarity metrics and displaying high-

dimensional structures. In 1980s SemNet produces three-dimensional graphic representations of large knowledge bases to help users grasp complex relationships involved [9]. The design of SemNet focuses on the graphical representations of three types of components: identification of individual elements in a large knowledge base, the relative position of an element within a network context, and explicit relationships between elements. SemNet represents elements of a knowledge base of Prolog rules, a logic programming language, as labeled rectangles connected by lines or color-coded arcs. A Prolog module, a set of rules, is represented as a rectangle labeled by the module's name. In SemNet, the closeness between two rectangles indicates the strength of the connection between two modules. The designers of SemNet experimented various techniques such as multidimensional scaling (MDS), simulated annealing, fisheye views, and even a sprite traveling down arcs between rectangles to show the process of the knowledge base in action. In this section, we include a number of useful techniques for detecting and extracting salient elements from unstructured text. Interrelationships between these salient elements will form the basis for the visualization of knowledge structures.

2.1. Information Retrieval Models

The vector-space model [27] originally developed for information retrieval, is a widely used framework for indexing documents based on term frequencies. In this model, each document *d* is represented by a vector *V* of terms *t*'s. Terms are weighted to indicate how important they are in representing the document. The distance between two documents can be determined by the distance between vectors in a high-dimensional space, or the angle between the two vectors. A large collection of documents can be split into a number of smaller clusters such that documents within a cluster are more similar than documents in different clusters.

The discriminative power of a term is determined by the well-known $tf \times idf$ model, term frequency (*tf*) and inverse document frequency (*idf*). Each document can be represented by an array of terms T and each term is associated with a weight determined by the above $tf \times idf$ model. In general, the weight of term T_k in document D_i, is estimated as follows:

$$w_{ik} = \frac{tf_{ik} \times log\left(\frac{N}{n_k}\right)}{\sqrt{\sum_{j=1}^{T}\left(tf_{ij}\right)^2 \times log\left(\frac{N}{n_j}\right)^2}} \quad (1)$$

where tf_{ik} is the occurrences of term T_k in D_i, N is the number of documents in a given collection, and n_k represents the

number of documents containing term T_k. The document similarity is computed as follows based on corresponding vectors $V(D_i) = (w_{i1}, w_{i2}, ..., w_{iT})$ and $V(D_j) = (w_{j1}, w_{j2}, ..., w_{jT})$:

$$sim_{ij}^{content} = \sum_{k=1}^{T} w_{ik} \times w_{jk} \tag{2}$$

2.2. Bayesian Theory

Thomas Bayes' work on mathematical probability focused on the probabilistic relationship between multiple variables and determining the impact of one variable on another. Bayesian theorem has become a central part of modern statistical probability modeling.

A good introduction to Bayesian modeling and neural networks can be found in a book entitled *Bayesian Learning for Neural Networks* [22]. The author of the book, Radford Neal, has made software described in this book available on the Internet[2], in particular, on Bayesian regression and classification models based on neural networks and Gaussian processes, and Bayesian mixture models. The software is written in ANSI C for Unix and Linux systems. The website also includes software that supports a variety of Markov chain sampling methods, which may be applied to distributions specified by simple formulas, including simple Bayesian models defined by formulas for the prior and likelihood.

Bayesian theory can be used to judge how relevant a document is to a given query based on what we already know. More formally, $p(d|q)$ denotes the relevancy of a document d to a given query q. We may be able to work out $p(d)$ and $p(q|d)$ based on, for example, users' previous records and their profiles. We can then estimate this relevance as follows

$$p(d|q) = \frac{p(q|d) \cdot p(d)}{\sum_{d' \in \{d\}} p(q|d') \cdot p(d')} \tag{3}$$

Extensions of the theory go further than relevance information for a given query against a text. Adaptive probabilistic concept modeling analyzes correlation between features found in documents relevant to an agent

[2] http://www.cs.toronto.edu/~radford/fbm.software.html

profile, finding new concepts and documents. Concepts important to sets of documents can be determined, allowing new documents to be accurately classified.

2.3. Shannon's Information Theory

Information Theory is the mathematical foundation for all digital communications systems. Claude Shannon's innovation was described in his "Mathematical Theory of Communication" published in 1949, which shows that "information" could be treated as a quantifiable value in communications.

Consider the basic case where the units of communication (for example, words or phrases) are independent of each other. If p_i is the probability of the i^{th} unit of communication, the average quantity of information conveyed by a unit, the degree of uncertainty can be measured by Shannon's *entropy*, which is defined as:

$$H = -\sum_i p_i \cdot \log_2(p_i) \tag{4}$$

This formula reaches its maximum when the probabilities are all equal; in this case the resulting text would be random. See Section 2.4 for a measure of the occurrences of content-bearing words based on the singularity among randomness. If this is not the case the information conveyed by the text will be less than this maximum; in other words there is some redundancy. This result is then extended, by more sophisticated mathematical arguments, to when units are related.

Natural languages contain a high degree of redundancy. A conversation in a noisy room can be understood even when some of the words cannot be heard; the essence of a news article can be obtained by skimming over the text. Information theory provides a framework for extracting the concepts from the redundancy.

2.4. Condensation Clustering Value

Information retrieval systems typically distinguish between content-bearing words and terms on a stop list. But "content-bearing " is relative to a collection. For optimal retrieval efficiency, it is desirable to have automated methods for custom building a stop list. Bookstein et al. [1] developed the method for serial clustering of words in text and use such clustering as an indicator of a word bearing content.

The idea is that occurrences of a term sensitive to content are more likely to cluster, or occur in the same textual neighborhood than those of non-content-bearing terms. However, it is not easy to determine by intuitive means

which objects tend to cluster. A person may see clusters forming even among objects that are randomly distributed. Therefore, formal tests of clustering are needed. There are several possible ways to measure the strength of clustering. Condensation Clustering Value (CCV) is one of them.

If text is fed into the system as a stream, one word at a time, one would expect that occurrences of a given content-bearing word will tend to occur in clumps (just like London buses). This is an indication of the content of a textual neighborhood. If we now analyze the text in segments, i.e. strings of words, some segments should contain more of these words than expected by chance. This is in fact a serial-clustering mechanism, known as *condensation clustering* [1]. In general, consider the context of words as a container, a sentence is a container and a paragraph is a container. Containers should attract some words more effectively than others. Since non-content-bearing words are likely to appear at random across such containers, i.e. text segments, if the occurrences of a word are different from such random distributions, then the word is likely to be significant to the topic in question.

If terms are distributed at random over a number of textual units, several may land in a single unit. Thus, we expect the number of units containing at least one occurrence of the term to be less than the total number of occurrences of the term. But if the terms tend to cluster, in the condensation sense, then we expect even fewer units will contain a term. A simple measure of condensation clustering is the ratio of the actual number of units that have at least one occurrence of the term to the expected number of units having at least a single occurrence, assuming random distribution. The statistical properties of this measure can be derived using basic combinatorial arguments.

Suppose that our document is divided into D textual units. (We are currently developing effective methods for accomplishing this division. For the moment, any arbitrary segmentation method — for example, taking successive groups of five sentences, or accepting existing paragraphs — are possibilities.) Consider then the distribution of the occurrences of a term over these units.

For that term, we know:

N: the number of units containing the term; and

T: the total number of occurrences of the term.

Then *p(n, T)*, the probability of exactly *n* units containing a term can be shown to be given by

$$p(n,T) = \frac{n! \left\{ {D \atop n} \right\} \left[{T \atop n} \right]}{D^T} \tag{5}$$

The expected number of textual units containing the term, $E_1 = \Sigma_n n \times p(n, T)$, can be shown to satisfy

$$E_1 = D\left[1 - \left(1 - \frac{1}{D}\right)^T\right] \tag{6}$$

If N units actually contain the term, N/E_1 is a measure of condensation strength. If the terms do cluster within textual units, then we expect that fewer units will contain the term than predicted by the independence model — that is, that the observed N will be small compared to the value expected on the basis of the independence assumption.

As a test of statistical significance, we can compute the probability P(N, T) that N or fewer units contain the term, given the independence assumption. This value should be unreasonably small for most content-bearing terms. This observation forms the basis of Test 1 of the hypothesis that the terms are randomly distributed. This probability y is given by:

$$P(N, T) = \sum_{n=1}^{N} p(n, T) \tag{7}$$

In this manner we can assign to each term both a measure of condensation strength and the statistical significance of that value: the probability that it will occur in no more units than it actually does occur in. These quantities measure the tendency of a single term to condense into a few textual-units.

The value of Condensation Clustering Value (CCV) has been demonstrated later on by researchers at PNL in SPIRE, a suite of visualization and spatial exploration tools for information retrieval [37]. In SPIRE, topical words are extracted using the following procedure:

- remove stopwords,
- stem remaining words,
- band pass to remove both high- and low-frequency words so that remaining words appear 3~5 times in a document,
- compute condensation clustering value of each word remained, and
- select words whose CCV values are less than 1 (these are topical words).

This procedure can be easily extended to select sentences that would summarize an abstract. In the following example, we develop a light-weight algorithm based on this procedure to produce a one-liner summary for each abstract. The example is based on a keyword search in the Web of Science on “visualiz* knowledge”. The Web of

Science will automatically expand the wildcard * in the query to match visualization, visualizing, and visualized. The abstracts of the top 300 articles returned by Web of Science form the test data. Two additional steps are added to the topic-word-selection procedure as follows:

- compute the content-bearing level of every sentence in each abstract
- select the sentence with the least value as the summary sentence

If *T* is the set of all the selected topic words, *t* is a word in this set, and *s* is a sentence for all the sentences *S* from a given abstract, then the inverted condensation level of the sentence ξ(s) can be measured as follows:

$$\xi(s) = \prod_{t \in T \bigcap s} ccv(t) \tag{8}$$

The more topical words a sentence contains, the more likely that the sentence summarizes the abstract. Therefore, the best summary sentence is the sentence that has the least inverted condensation level.

$$\lambda(\{s_\alpha\}) = \min_{s \in \{s_\alpha\}} \xi(s) \tag{9}$$

For example, the following paragraph is the abstract of a real article:

> Sex-differences in route-learning past literature on map-learning tasks has generally inferred that males tend to use a geometric strategy, and females tend to use a landmark strategy to learn a map. However, none of the studies have controlled for possible effects of extra-map superior visual memory in females on their memory for landmarks, and few have probed the actual relation between accuracy of performance and geometric or landmark knowledge. This study investigated sex differences in strategies for route-learning, controlling for visual-item memory. All subjects (48 female, 49 male) were required to learn a route to criterion through a novel map. As expected, males made fewer errors and took fewer trials to reach criterion. Females remembered more landmarks both on and off the route than males, and superior memory for landmarks was not accounted for by a superior visual-item memory. Males outperformed females in knowledge of the euclidean properties of the map. However, despite the pronounced sex differences in knowledge retained from the maps, both males' and females' performance was related to spatial ability rather than to landmark recall.

Using the CCV algorithm, a number of content-bearing, or thematic words can be extracted. The lower the CCV value of a word, the stronger this indicates a thematic word. Selected thematic words are listed in the following table.

Content-Bearing Words	CCV
knowledge	0.012
map	0.148
superior	0.181
memory	0.187
landmarks	0.338
males	0.632
females	0.721

Table 1. Content-bearing words in the abstract.

Sentences in the original sequence	ξ(s)
Sex-differences in route-learning past literature on map-learning tasks has generally inferred that males tend to use a geometric strategy, and females tend to use a landmark strategy to learn a map.	0.067439
however, none of the studies have controlled for possible effects of extra-map superior visual memory in females on their memory for landmarks, and few have probed the actual relation between accuracy of performance and geometric or landmark knowledge.	0.000009
this study investigated sex differences in strategies for route-learning, controlling for visual-item memory.	0.187000
all subjects (48 female, 49 male) were required to learn a route to criterion through a novel map.	0.148000
as expected, males made fewer errors and took fewer trials to reach criterion.	0.632000
females remembered more landmarks both on and off the route than males, and superior memory for landmarks was not accounted for by a superior visual-item memory.	0.005213
males outperformed females in knowledge of the euclidean properties of the map.	0.000809
however, despite the pronounced sex differences in knowledge retained from the maps, both males' and females' performance was related to spatial ability rather than to landmark recall.	0.000809

Table 2. Sentences in the abstract and the ccv of each sentence.

Figure 1 shows the results of the simple summarization algorithm. It is currently a challenging task to assess to what extent the automatically generated summaries match the ones that human experts would choose.

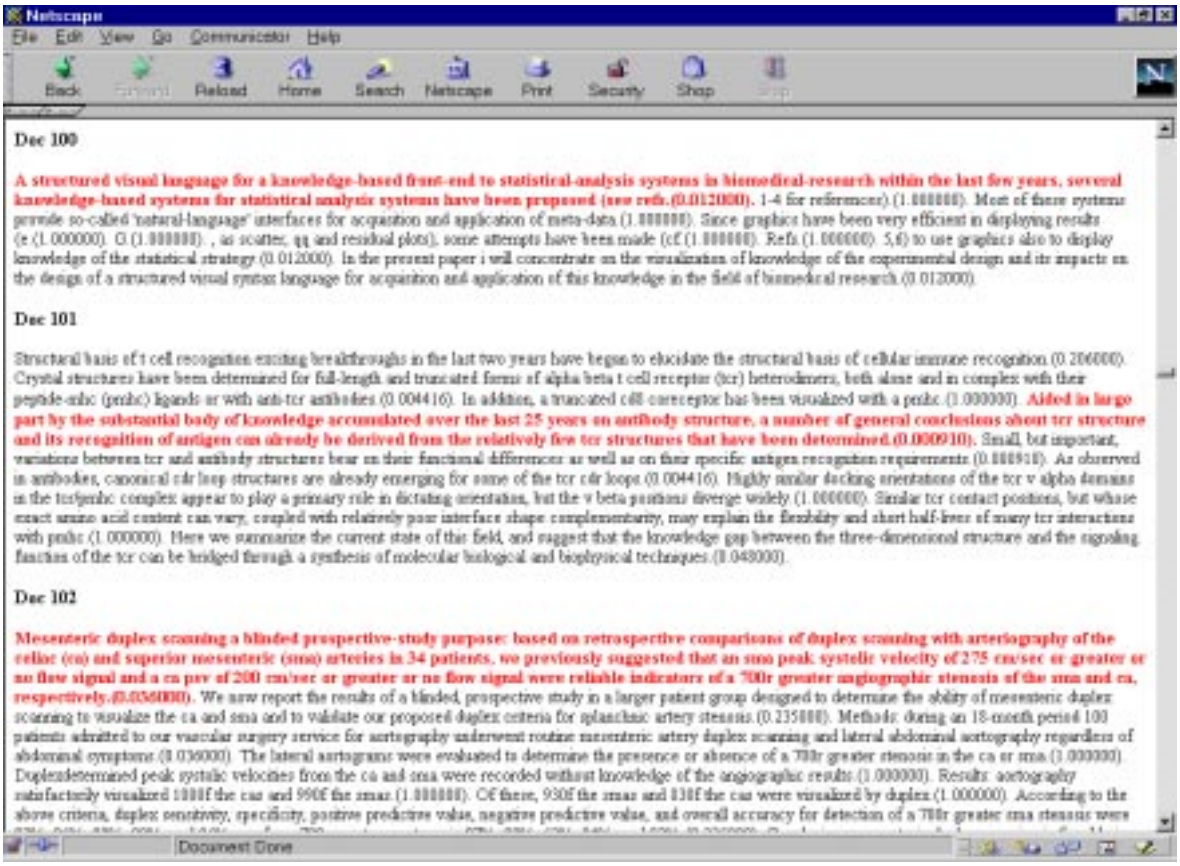

Figure 1. Summary lines extracted from abstracts.

2.5. Self-Organized Feature Maps

Self-organized feature maps are generated based on artificial neural network techniques. It is essentially a classification process through a neural network. Xia Lin is the first to use self-organized maps to visual information retrieval [19]. ET-Map[3] is another example of site maps on the Web. ET-Map is a prototype Internet homepage categorization system developed at the University of Arizona [7]. ET-Map aims to demonstrate a scalable, automatic, and concept-based approach to Internet homepage categorization and search.

WEBSOM[4] uses self-organized maps to organize textual documents for exploration and search. WEBSOM has been used to map discussion groups. A click in any area on the map will lead to a zoomed view. Color denotes the density or the clustering tendency of the documents. Lighter areas are clusters and darker areas are empty space between the clusters. The most specific discussions are mostly found in the clearest "clusters", i.e. light regions surrounded by

[3] http://ai2.BPA.Arizona.EDU/ent/

darker color. Near the edges of the map you typically find the most "different" documents represented on the map. In the central areas, the discussions are more "typical", or they may concern many different topics found on the map. Viscovery SOMine is a recently launched commercial product[5]. It can generate self-organized maps. The standard version costs $289, but can only deal with 50 variables. The enterprise version is $1990.00, which can handle a lot more variables.

2.6. Citation Indexing

Today's widely available of citation index databases was originated in 1950s. Indexing in 1950s was inconsistent and uncoordinated. There was widespread dissatisfaction with the array of traditional discipline-oriented indexing and abstracting services. In 1955, Eugene Garfield published his pioneering paper in *Science* [10], aiming to improve the retrieval of science information.

Garfield presented a history on the sociological and historical uses of citation data in 1975[6]. He acknowledged that the most important paper on sociometric analysis with citation data was the one by Sher and Garfield on patterns for Nobel Prize winners [30]. The following historical account of citation indexing is drawn from his description.

One of the pioneering science mapping based on citation data is the creation of the historical map of research in DNA, which was done manually more than 30 years ago in early 1960's. [12]. The same data was soon used by Derek Price in his classic work of mapping scientific networks [23].

Mapping of science is also known as scientography [11]. Because in maps of science interrelationships between research fronts are spatialised, users can navigate the scientific literature according to the depicted spatial relationships.

One of the most central objectives of science mapping is to identify trend associated with a field of study. As the maps of science cover the literature year by year, the map created through citation analysis provides a series of historical views. Traditionally, from a global viewpoint, these maps show latent semantic connections among fields or disciplines. The maps reveal which realms of scholarship are being investigated today and the individuals, publications, institutions, regions, or nations currently eminent in these areas. This is echoed in more recent work of Hjorland [16].

---

[4] http://websom.hut.fi/websom/
[5] http://www.unisoftwareplus.com/products/somine.html
[6] http://www.garfield.library.upenn.edu/papers/retrospectivey1975.html

Longitudinal mapping is another concept discussed by Garfield [11]. By using a series of chronologically sequential maps, one can easily see how knowledge advances. By observing changes from year to year, one can detect trends. In this way, these maps become forecasting tools. Since some co-citation maps include core works, even a novice can instantly identify those articles and books used most often by members of the "invisible college", or specialties.

Citation-based mapping evolves into a distinct discipline known as scientometrics, which applies bibliometrics to scientific literature. Scientometrics is the quantitative study of scientific communications. The "father of scientometrics" was Derek Price (1922-1983), as regarded by Robert Merton and Eugene Garfield in their foreword[7] to the second edition of Price's landmark work *Little Science, Big Science, and Beyond* [24]. Price's most important contribution to information science is his another seminal work Networks of Scientific Papers, published in science [25].

In 1981, Institute for Science Information (ISI) published the Atlas of Science in biochemistry and molecular biology. The Atlas was constructed based on co-citation index associated with publications in the field over a limited period of one year. 102 distinct clusters of articles were identified, which were called research front specialties, in order to give researchers a snapshot of significant research activities in biochemistry and molecular biology.

More than 100 people were involved in the project for several months. In the Atlas, journal articles were clustered according to associated co-citation index. The atlas provides a clear, distinct snapshot of the scientific network and how it was structured. More recently, ISI developed Sci-Map software for users to navigate the citation network. Garfield and Small [13] explained how they visualised the changing frontiers of science based on citation relationships.

Henry Small described ISI's SCI-Map in a case study of mapping AIDS research in [31]. SCI-Map creates maps of individual research areas specified by the user. Given an author, paper, or keyword as a starting point, one can seed a map and then grow the map by specifying various desired connections at different thresholds of co-citation strength or distance. The network of connected nodes is formed by a series of iterations of clustering, including additional core papers with each successive node. The nodes are selected according to the strength of their links, and positioning is determined by the *geometric triangulation method*. A map of AIDS research by SCI-Map (Small, 1994) is available at http://www.isinet.com/isi/hot/essays/.

[7] http://www.garfield.library.upenn.edu/lilscibi.html

In 1977, Henry Small published his longitudinal study of co-citation linkages [32]. He identified the most important advances in collagen research over a period of five years. The results were validated by questionnaires. In 1998, the American Society for Information Science (ASIS) awarded him the ASIS Award of Merit[8] for his outstanding contributions to the field of information science and for his groundbreaking work on co-citation as a dynamic measure of the scientific literature.

In his most recent work, Small explored the notion of a passage through science [33, 34]. Passages linking the literature of different disciplines are likely to import or export a method established in one discipline into another. This has been known as cross-disciplinary fertilization. As Small has noted, this reaching out or stretching can import or export methods, ideas, models, or empirical results from the author's field to the other field. This requires scientists to have not only a broad awareness of literature, but also the creative imagination to foresee how the outside information fits with the problem at hand. He developed algorithms to blaze a magnificent trail of more than 300 articles across the literatures of different scientific disciplines. This is a pioneering development that brings Bush's concept of associated information trails to life.

The latest idea of mapping the tracks of science is explained by Garfield in [11]. The aim of such work is to identify research fronts in science. We use a simple timeline to highlight the development of science mapping.

2.7. Author Co-Citation Analysis

Author co-citation analysis (ACA) aims to find out how scientists in a particular subject domain are intellectually interrelated as perceived by authors in their scientific publications. White and Griffith are the first to introduce author co-citation analysis (ACA) in 1981 as a literature measure of intellectual structure [35]. McCain [20] produced a comprehensive technical review of mapping authors in intellectual spaces. In 1998, White and McCain applied ACA on information science in their award-winning paper - the best JASIS paper award of the year [36]. In ACA, the unit of analysis is authors and their intellectual relationships as reflected through scientific literatures.

Typically, the process of an ACA starts with sampling representative publications from a literature of a given field of study. Science Citation Index (SCI) and Social Science Citation Index (SSCI) are among the most widely used sources of citation data - both from Institute for Scientific Information (ISI). An intellectual structure of prominent authors in the field provides a good candidate for knowledge visualization. Normally the predominance of an author

[8] http://www.asis.org/Bulletin/Jan-99/small.html

is determined by citations he or she has received or by other criteria, such as the membership of a scholarly institution. Author co-citation frequencies among these selected authors are then calculated. If a pair of authors X and Y are cited by the same scientific publication, the author co-citation counts of the pair will increment by one.

ACA traditionally relies on a range of data analysis methods in order to identify emergent patterns in the co-citation data. Commonly used methods include cluster analysis, factor analysis, and multidimensional scaling (MDS).

Step	Operation
1	Draw citation data from bibliographic databases
2	Select bibliographic records corresponding to the range of citation analysis
3	Clean up data
4	Compute co-citation index
5	Convert co-citation index to Pearson correlation coefficients
6	Identify specialties using factor analysis
7	Generate overall co-citation maps
8	Add specialty information into co-citation maps
9	Add citation history into co-citation maps
10	Repeat above steps for longitudinal mapping
11	Interpret the results
12	Repeat above steps for longitudinal mapping

Table 3. Procedure of citation analysis.

2.7.1. Factor Analysis

Factor analysis has been a very powerful step in author co-citation analysis. Factor analysis, notably Principal Component Analysis (PCA), has been used to identify the latent dimensionality of given co-citation data. It has been routinely used to identify specialties of a subject domain.

In addition to cluster authors into mutual exclusive groups, or specialties, the use of factor analysis allows a multiple-specialty membership for each author. For example, White and McCain demonstrated in their recent author co-citation analysis of the information science field that some authors indeed belong to several specialties simultaneously.

Factor analysis can be conducted on standard statistical packages such as SPSS. For large datasets, the size of the corresponding author co-citation matrix can be very big and the analysis becomes computationally expensive. Following [36], first, the raw co-citation counts should be transformed into Pearson's correlation coefficients using the factor analysis. These correlation coefficients measure the proximity between authors' co-citation profiles. Self-citation counts should be replaced with the mean co-citation counts for the same author. In factor analysis, principal component analysis (PCA) with varimax rotation is a recommended option to extract factors. The default criterion, eigenvalues greater than one, is normally chosen to determine the number of factors extracted. Missing data should also be replaced by mean co-citation counts for corresponding authors.

The following SPSS script illustrates how to conduct factor analysis in SPSS. This example is based on our author co-citation of ACM Hypertext data. The input file is "hypertext.dat", which represents the lower triangle of author co-citation matrix. Elements on the diagonal positions will be ignored because they will be regarded as missing values and to be replaced by the mean for corresponding authors. 367 authors identified in the matrix as `author1`, `author2`, …, `author367`. The factor extraction uses principal component analysis with varimax rotation.

```
MATRIX DATA /VARIABLES author1 to author367 /FILE'hypertext.dat'
  /CONTENTS N_MATRIX /FORMAT LIST
LOWER NODIAGONAL.
SET LISTING 'hypertext' RESULTS LISTING.
FACTOR /VARIABLES author1 to author367 /MISSING MEANSUB
  /PRINT EXTRACTION /EXTRACTION PC /ROTATION VARIMAX.
```

The following SPSS script can be used to generate factor loading data, which can be used subsequently to enhance the visualization. Factor loading indicates the strength that an author belongs to a particular specialty. There are many possible ways that domain visualization can utilize this information.

```
FACTOR /VARIABLES author1 to author367 /MISSING MEANSUB
  /PRINT FSCORE /EXTRACTION PC /ROTATION VARIMAX.
```

Pearson's correlation coefficient *r* can be used as a measure of similarity between pairs of authors. According to [36], it registers the likeness in shape of their co-citation count profiles over all other authors in the set.

2.7.2. Multidimensional Scaling

Multidimensional scaling (MDS) is a multivariate statistical technique, which is often used to map high-dimensional numerical data onto a spatial structure in lower dimensions. A classic MDS algorithm was given by Kruskal [17], known as KYST. KYST places N points in a space of dimension LDIM so as to minimize STRESS, which measures the "badness-of-fit" between the configuration of points and the data. If it finds the minimizing configuration by starting with some configuration perhaps found by the classical scaling procedure and moving all the points a bit to decrease the stress, then this procedure will be iterated over and over again until the stopping criteria are reached. KYST users the iterative numerical method of gradients, the method of steepest descent, with a step-size procedure based primarily on the angles between successive gradients.

The most common measure of the stress is used to evaluate how well a particular configuration reproduces the observed distance matrix. The raw stress value ϕ of a configuration is defined by:

$$\phi = \left(d_{ij} - \delta_{ij}\right)^2 \tag{10}$$

In this formula, d_{ij} stands for the reproduced distances, given the respective number of dimensions, and δ_{ij} stands for the observed distances. The expression f (δ_{ij}) indicates a non-metric, monotone transformation of the observed distances. There are several measures that are commonly used, including the sum of squared deviations of observed distances from the reproduced distances. The smaller the stress value, the better is the fit of the reproduced distance matrix to the observed distance matrix. The greater the stress, the greater the distortion.

Interpreting the essence of a MDS map focuses on clusters and dimensions. Clusters are groups of items that are closer to each other than to other items. Within a cluster, one should bear in mind not over interpret local patterns because such patterns within a tight cluster are not reliable. Dimensions may exist if items appear to be placed in the MDS map along a continuum. The underlying dimensions provide the basis to explain the perceived similarity between items.

The most common complaint about MDS maps is probably about the ambiguity in the nature of each dimension involved. The interpretation of MDS solutions has been mainly subjective in nature. Figure 2 shows a 2D MDS map of a network of top 100 most cited authors derived from their co-citation relationships in ACM Hypertext proceedings [5].

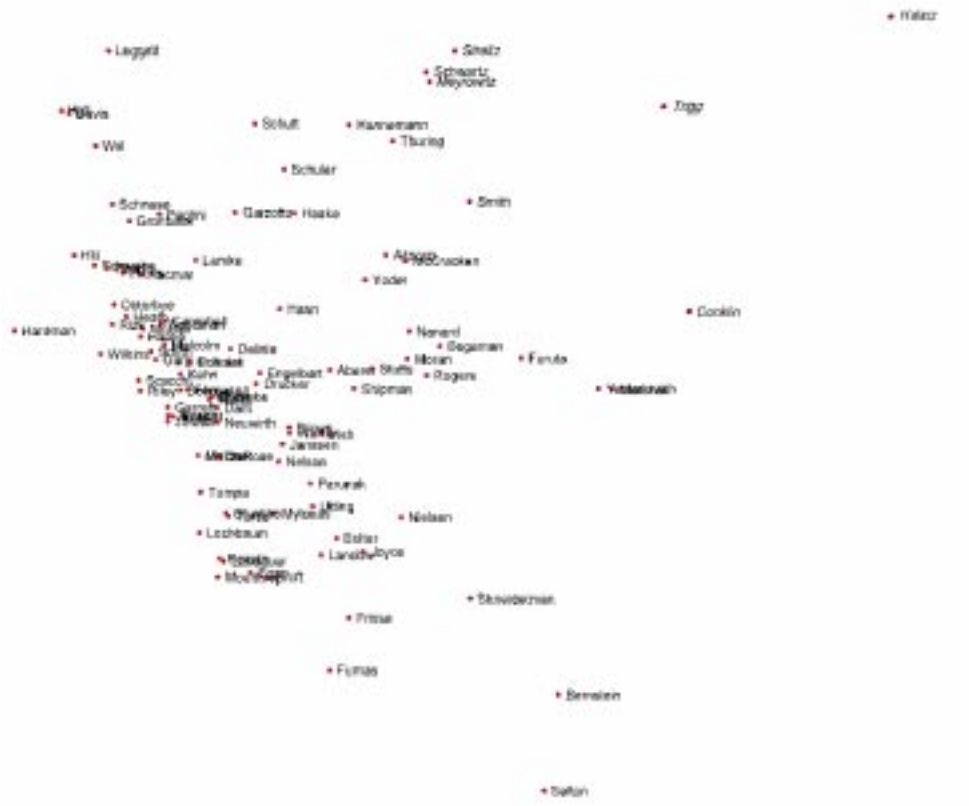

Figure 2. Top 100 authors in the field of hypertext in a 2D MDS map [5]**.**

2.7.3. Minimum Spanning Trees

Although MDS has been a widely used method in many disciplines, interpreting MDS maps is often subject to one's experience and the familiarity with the data. There is a need for an alternative representation that has more accurate local details and more explicit interpretation of dimensionality.

A wide range of real-world problems can be transformed into a network problem in one way or another. A network representation is generally a good candidate for capturing the essence of a structure. From a graph-theoretical perspective, a large number of widely available graph-handling algorithms have been a great gold mine.

In a complete graph, there exists a link between each pair of vertices. So for a given graph of *N* vertices, there could be as many as $N \times (N\text{-}1)$ edges. It is often desirable to reduce the number of edges while the most significant ones can be preserved. Minimum spanning trees (MSTs) are widely used solutions, which include all the vertices from the original graph, but only consist of $N - 1$ edges at most. MST algorithms are widely available. Many visualization systems rely on MSTs to represent complex structures, such as the well-known cone trees [26] and hyperbolic views [21].

Figure 3 shows a minimum spanning tree (MST) solution of the same author co-citation network derived from the ACM Hypertext conference proceedings [4]. This is a network of 367 prominent scientists in the field of hypertext. The interconnection between a pair of authors is determined according to the co-citation frequencies between the two authors. The original network consists of 61,175 links among these authors. As a graph, the maximum number of possible links in a fully connected graph would be

$$\frac{N(N-1)}{2} = \frac{367 \times 366}{2} = 67161 \tag{11}$$

In this case, 91% of the maximum possible co-citations are observed. The number of links in an MST is $N - 1$, i.e. 366 in our example, which is only 0.6% of the actual observed co-citation links. It leads to a much simplified picture of the patterns.

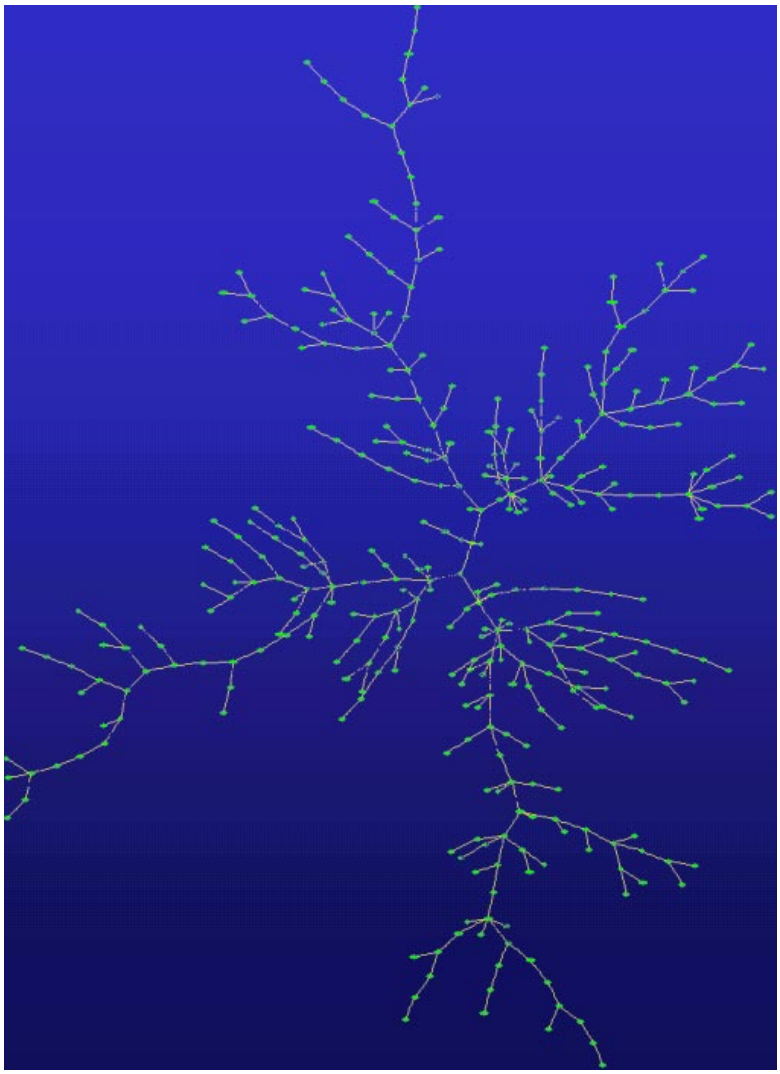

Figure 3. A MST solution to the same author co-citation network of 367 prominent scientists in the field of hypertext (Nodes=367, Links=366).

Hierarchical structures are one of the most commonly used data structures in many application domains. Since MST solutions can simplify a complex network into a hierarchical structure, MST is regarded as a standard method to deal with the visualization of complex structures. A relatively recent visualization technique for displaying a large hierarchical structure is called hyperbolic views. Hyperbolic views are originally developed at Xerox PARC. It is now available from its spin-off company *Inxight*. Hyperbolic visualization is also licensed to Microsoft's SiteMap. Much of the work in hyperbolic views is motivated to balance both the focus and its context in each screen display.

A hyperbolic view is specified by a mathematical model, especially suitable to display a large, unbalanced hierarchy. The greatest advantage of a hyperbolic display is that nodes higher up in the hierarchy are initially displayed in the

center of the view, while nodes further down in the hierarchy are displayed with an increasingly reduced screen estate. Inxight is a spin-off company from Xerox. It commercializes a tree visualization interface based on hyperbolic views.

Figure 4. A hyperbolic visualization from Inxight.

The explicit display of edges in a network representation has an advantage over spatial-proximity designs such as multidimensional scaling (MDS), where interrelationships are implicitly conveyed by spatial arrangements of objects. Explicit links tell a more detailed story about the latent structure. Pathfinder network scaling is a method that can select salient relationships from proximity data and maintain the semantic integrity over the entire graph.

2.7.4. Pathfinder Network Scaling

Pathfinder network scaling is a method originally developed by cognitive psychologists for structuring modeling [29]. Pathfinder network scaling relies on a criterion known as the triangle inequality condition to select the most salient relations from proximity data. Results of Pathfinder network scaling are called Pathfinder networks, consisting of all the vertices from the original graph. The number of edges in a Pathfinder network, however, is determined by the intrinsic structure of semantics. On the one hand, a Pathfinder network with the least number of

edges is identical to an MST. On the other hand, additional edges in a Pathfinder network indicate salient relationships that might have been missed from an MST solution.

The topology of a PFNET is determined by two parameters q and r and the corresponding network is denoted as PFNET(r, q). The q-parameter controls the scope that the triangular inequality condition should be imposed. The r-parameter refers to the Minkowski metric used for computing the distance of a path. The weight of a path with k links is determined by weights w_1, w_2, ..., w_k of each individual link as follows:

$$W(P) = \left(\sum_{i=1}^{k} w_i^r \right)^{\frac{1}{r}} \tag{12}$$

The Minowski distance (geodetic) depends on the value of the r-metric. For $r = 1$, the path weight is the sum of the link weights along the path; for $r = 2$, the path weight is computed as Euclidean distance; and for $r = \infty$, the path weight is the same as the maximum weight associated with any link along the path.

$$W(P) = \left(\sum_{i=1}^{k} w_i^r \right)^{\frac{1}{r}} = \begin{cases} \sum_{i=1}^{k} w_i & r = 1 \\ \left(\sum_{i=1}^{k} w_i^2 \right)^{\frac{1}{2}} & r = 2 \\ \max_i w_i & r = \infty \end{cases} \tag{13}$$

The q-parameter specifies that triangle inequalities must be satisfied for paths with $k \le q$ links:

$$w_{n_i n_i+1} = \left(\sum_{i=1}^{k-1} w_{n_i n_{i+1}}^r \right)^{\frac{1}{r}} \qquad \forall k \le q \tag{14}$$

When a PFNET satisfies the following three conditions, the distance of a path is the same as the weight of the path:

1. The distance from a document to itself is zero.
2. The proximity matrix for the documents is symmetric; thus the distance is independent of direction.
3. The triangle inequality is satisfied for all paths with up to q links.

If q is set to the total number of nodes less one, then the triangle inequality is universally satisfied over the entire

network. The number of links in a network can be reduced by increasing the value of parameter *r* or *q*. The geodesic distance between two nodes in a network is the length of the minimum-cost path connecting the nodes. A minimum-cost network (MCN), PFNET(*r*=∞, *q*=n-1), has the least number of links. See [3, 28] for further details.

Figure 5 is a Pathfinder network solution of the author co-citation matrix described earlier in our MST example. Red circles highlight the additional links with reference to the MST solution. A total of 398 links are included in the network – the pathfinder network contains additional 32 links compared with the MST solution, and these links are semantically non-trivial.

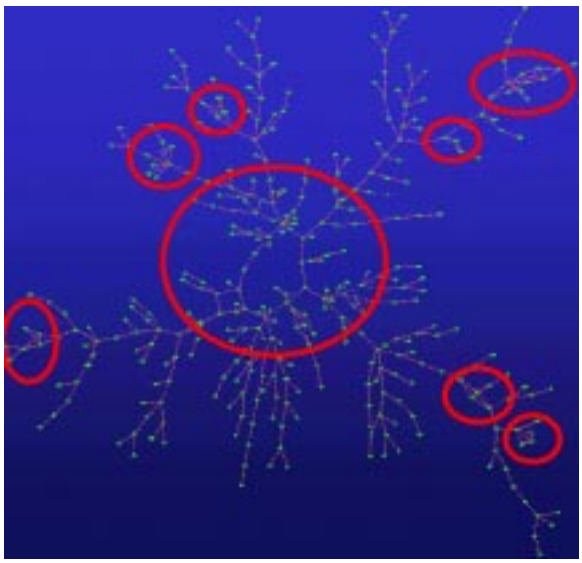

Figure 5. A Pathfinder network of the Hypertext author co-citation network, consisting of the same 367 predominant scientists as in the MST, but a total of 398 links.

3. Practical Approaches

This section collects several examples of practical approaches to knowledge extraction from unstructured documents. These examples highlight the state of the art in areas relevant to the visualization of knowledge structures. Our focus in this section is on how the key techniques discussed earlier in this chapter have been commercialized in various companies.

Company	Software	Further Information

Autonomy	Portal-in-the-Box	http://www.autonomy.com/
Cartia	Themescape	http://www.cartia.com/
Excalibur	RetrievalWare	http://www.excalibur.com/
Inxight	TableLense	http://www.inxight.com/
Institute for Scientific Information	SCI-Map	http://www.isi.net/
Pacific National Laboratories	ThemeView	http://www.pnl.gov/
Sandia National Laboratories	VxInsight	http://www.sandia.gov/
TextWise	DR-LINK	http://www.textwise.com/

Table 4. Relevant techniques for knowledge extraction.

3.1. Autonomy

Autonomy is a company specialized in knowledge management. It was founded in 1996. Autonomy's software is based on a combination of Bayesian inference and Shannon's Information theory to automatically extract the key concepts from unstructured information.

As explained earlier in this chapter, Bayesian inference is a mathematical technique for modeling the significance of ideas based on how they occur in conjunction with other ideas. Shannon's information theory provides the heuristics of extracting the most meaningful ideas in these documents.

Autonomy takes one step further from traditional search engines by actively matching information to user's interest. Furthermore, the interest of a user is constantly monitored and updated by observing and analyzing actions of the user. The user's topical interests and expertise are modeled by key concepts extracted from the articles the user reads online. The system updates the user profile by replacing concepts that are no longer important with newly emerged concepts. For example, UK experienced a petrol crisis in September 2000. The word *petrol* may be best characterized the user's interest at the time. When the crisis was over, petrol is no long an issue. The word *petrol* will disappear from the user's more recent reading and writing habits. Therefore, it will also disappear from the updated user profile. There is no need for users to update their profiles explicitly.

Autonomy's software supports not only individual users, but also a group of users. For example, one can borrow someone else' profile. This could be useful, if one needs to find information outside one's own expertise, but has access to the profile of an expert in that area.

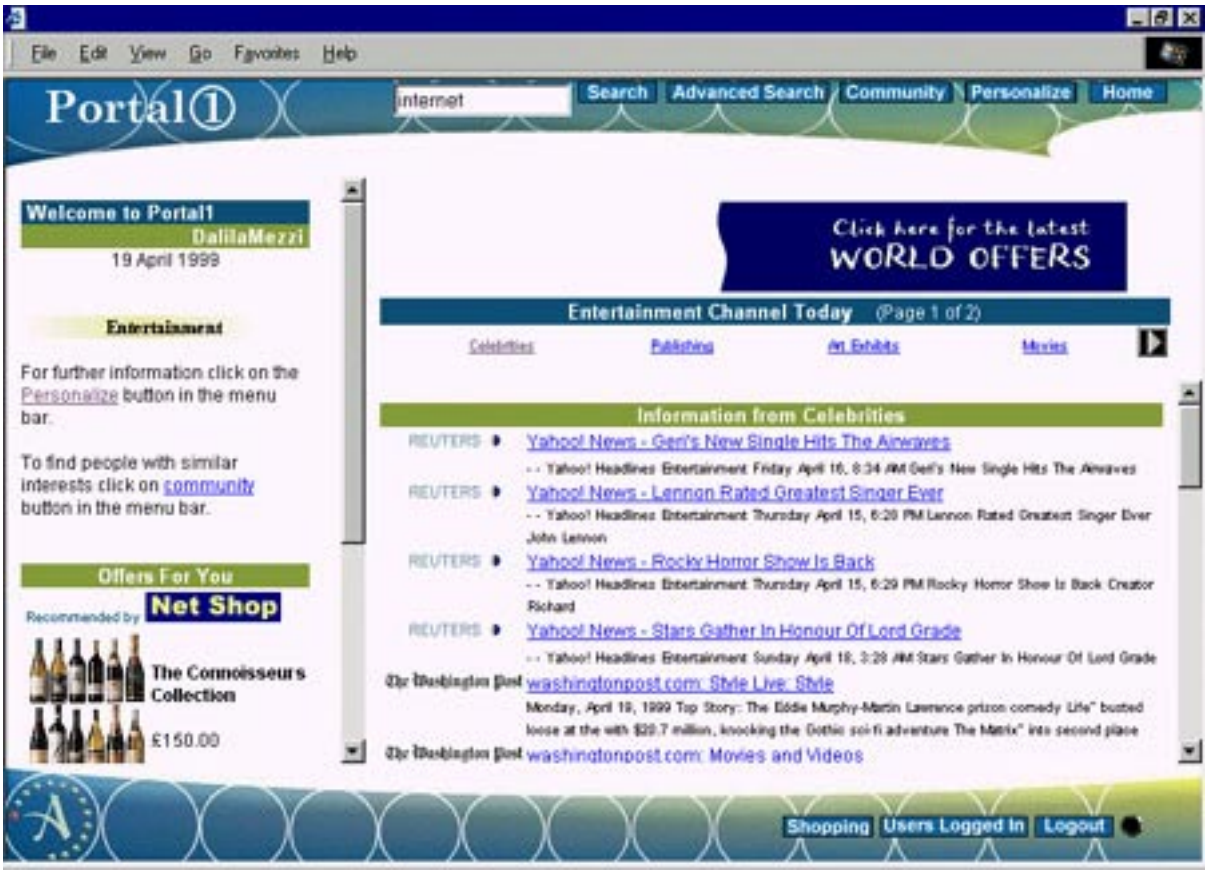

Figure 6. Autonomy's Portal interface.

3.2. TextWise

KNOW-IT is a commercial product from TextWise. It generates a rich conceptual representation of textual content and provides a suite of tools for mining the information content of the material. KNOW-IT supports information exploration and knowledge discovery by exploiting in-depth semantic analysis of textual information. KNOW-IT has been developed with funds from DARPA and the US Air Force's Rome Lab. Manning & Napier Information Services (MNIS) is a company established to commercialize information retrieval and text processing products (http://www.mnis.com). MNIS-TextWise Labs was founded in 1993 at Syracuse University as a wholly owned subsidiary of MNIS.

The system can take raw text, such as newspaper articles, as input and create a knowledge base of concepts and relations automatically. These concepts and relations can then be used to help users locate desired information. Users can formulate questions in ordinary English, for example, "What is the effect of the prolonged U.S.

presidential election on economics?" Visually presented answers allow users to explore documents of interest.

KNOW-IT is in fact an integrated suite of four different knowledge discovery tools:

- Topic Miner
- Link Miner
- Answer Miner
- Information Extractor

Information processing, knowledge visualization, and knowledge management will increasingly rely on automated archiving and knowledge extraction techniques. It is a rapidly growing market for tools that can extract relevant segments from unstructured text and organize them in a way that one can easily retrieve them.

3.3. ThemeView

SPIRE™, the Spatial Paradigm for Information Retrieval and Exploration is a classic example of information visualization developed by Pacific Northwest Laboratory [38]. SPIRE in fact is a suite of visualization tools for browsing and selecting text documents from large corpora, including a visualization view called Themescapes, which is later known as ThemeView. The SPIRE project is funded by the Department of Energy and the U.S. intelligence agencies. ThemeView provide a fine example of information landscapes, which have become one of the most widely known metaphors in information visualization. Wise provided a detailed description of the procedure they used to generate the relief map in his recent paper published in the *Journal of the American Society for Information Science* [37].

ThemeView forms abstract, three-dimensional landscapes of information that are constructed from a large set of documents, also known as a document collection, or corpus. A thematic terrain simultaneously communicates both the primary themes of the underlying document collection and a measure of their relative prevalence in the collection. Thematic peaks and valleys in ThemeView produce a simplified representation of the complex content of a document corpus.

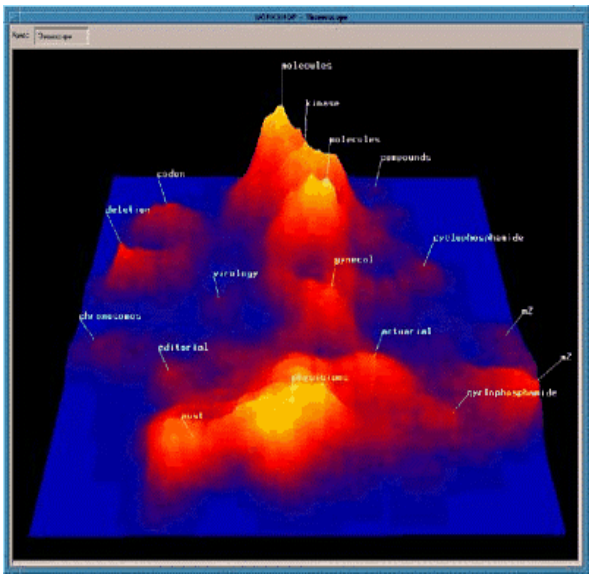

Figure 7. Valleys and peaks in ThemeView. (© Pacific Northwest National Laboratory).

The SPIRE used the classic vector space model [27]. The greatest advantage of the ThemeView approach over the use of traditional high-dimensional vector spaces is that the user is able to establish connections easily between the construction and the final visualization. In particular, their procedure usually results in 300~500 nouns to be visualized, which represent a significant improvement on the readability of the final result.

In a nutshell, the Galaxies visualization is based on clustering documents and projecting them into a two-dimensional plane. ThemeView extends this document space into a three-dimensional landscape visualization. The third dimension is used to convey the probability of finding a particular word in the document underneath. In terms of the thematic landscape, there are peaks and valleys of thematic terms.

3.4. Themescape of Cartia

Cartia recently launched a commercial product for visualization. The product is called ThemeScape. Cartia's ThemeScape has stirred up a great deal of enthusiasms in knowledge extraction. It represents a new generation of visualization software, which is now regarded not only as a text visualization tool, but also as a knowledge extraction tool.

The predecessor of ThemeScape is called Spyr, which is an experimental visualization technology developed at Batelle Memorial Institute for the CIA. U.S. intelligence agents used Spyr to analyze patterns in Iraqi message traffic in order to determine that Saddam Hussein's threat to re-enter Kuwait after the Gulf War was a hoax[9]. Spyr eventually evolved into ThemeScape as a commercial product launched by Cartia.

The user model of ThemeScape is to look at information from 30,000 feet. One can instantly recognize important patterns and relationships without ever opening a document. ThemeScape creates a visual landscape of information - a topographical map - that shows what's inside large collections of documents and web pages. With a quick scan of the landscape, one can locate the major topics within thousands of documents, and how different topics relate to one another.

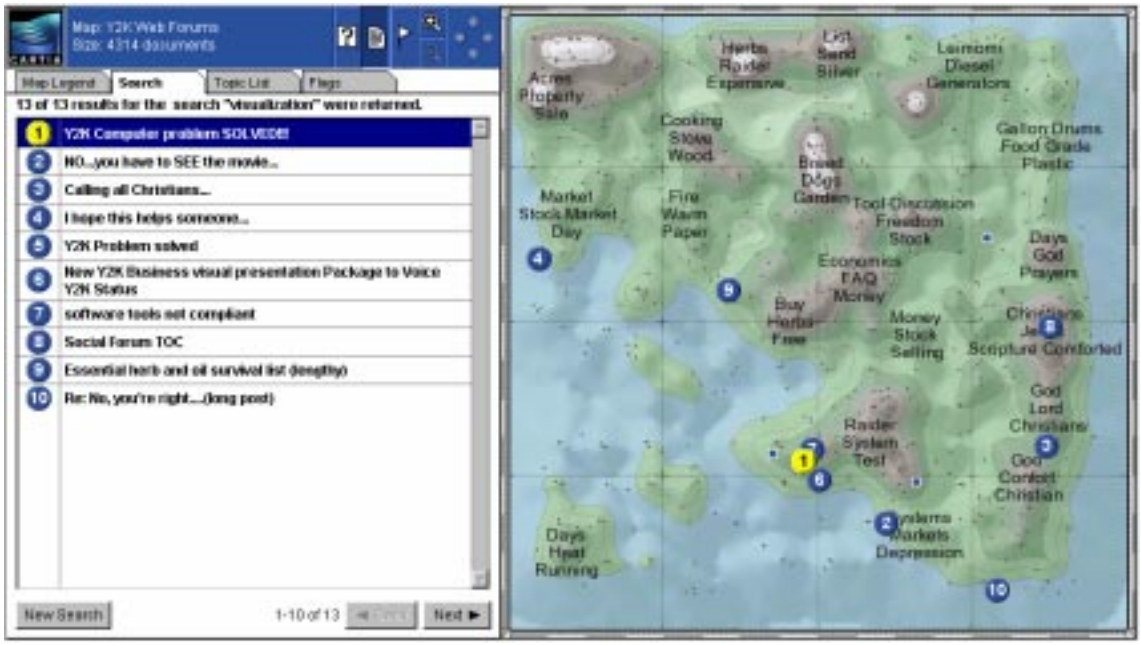

Figure 8. Themescape (©1999 Cartia, Inc.)

http://www.cartia.com/

As with any map, ThemeScape shows points of interest and the distance between them. The concept is simple: the greater the similarity between any two documents, the closer together they appear on the map. Concentrations of documents about a similar topic literally "pile up" to form peaks - points of interest - and the distance between peaks

[9] http://www.topsecretnet.com/knowmag.htm

shows how closely the topics are related. For example, a map about the Millenium Bug problem, also known as the Y2K problem, is shown in Figure 8. Prices for a ThemeScape publishing system vary depending on configuration, beginning at about $20,000 in 2000.

3.5. VxInsight

VxInsight [8] is another example that uses an information landscape metaphor. It is a knowledge visualization tool developed by Sandia National Laboratories. In VxInsight, data elements from very large datasets are grouped by similarity. VxInsight uses the height of a mountain in a three-dimensional virtual landscape to portrait the density of data elements distributed underneath. One application of VxInsight is the visualization of nuclear physics based on a subset of the Science Citation Index (SCI), a citation database from the Institute for Scientific Information (ISI). The similarity between two documents is proportional to the extent that the two documents have common citation links. VxInsight generates visualizations using a combination of eigenvector-based and force-directed placement solutions.

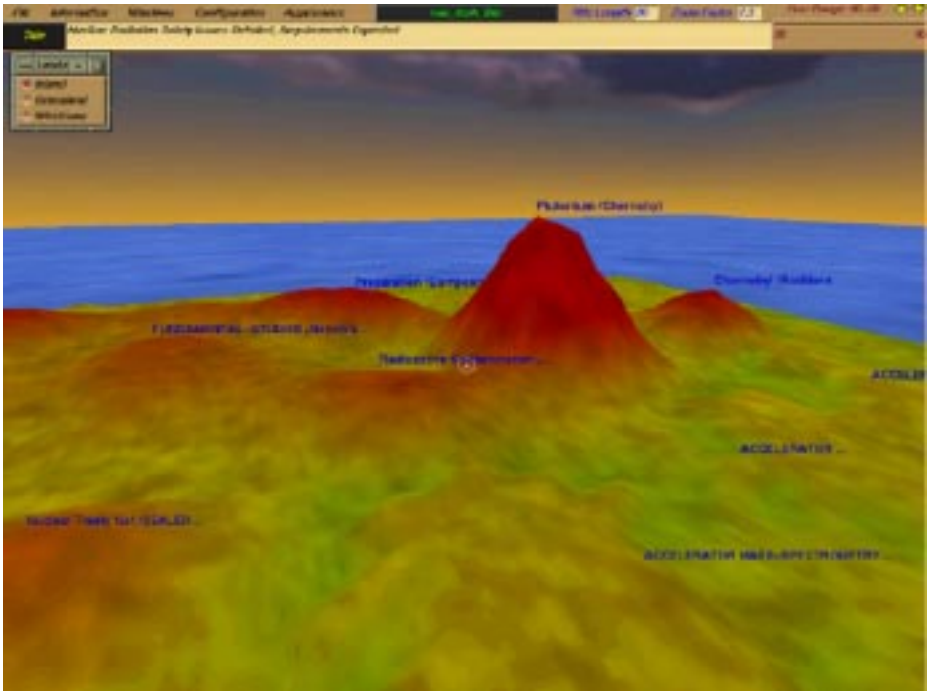

Figure 9. The mountain terrain of nuclear physics (© Sandia National Lab).

3.6. Visualizing Knowledge Structures

Figure 10 illustrates the four-step procedure of our approach [6]. First, select authors who have received citations above a threshold. Intellectual groupings of these authors represent snapshots of the underlying knowledge domain. Co-citation frequencies between these authors are computed from a citation database, such as ISI's SCI and SSCI. ACA uses a matrix of co-citation frequencies to compute a correlation matrix of Pearson correlation coefficients. According to [36], such correlation coefficients best capture the citation profile of an author.

Second, apply Pathfinder network scaling to the network defined by the correlation matrix. Factor analysis is a standard practice in ACA. However, in traditional ACA, MDS and factor analysis rarely appear in the same graphical representations. In order to make knowledge visualizations clear and easy to interpret, we overlay the intellectual groupings identified by factor analysis and the interconnectivity structure modeled by the Pathfinder network scaling. Authors with similar colors essentially belong to the same specialty and they should appear as a closely connected group in the Pathfinder network. Therefore, one can expect to see the two perspectives converge in the visualization. This is the third step. Finally, display the citation impact of each author on top of the intellectual groupings. The magnitude of the impact is represented by the height of a citation bar, which in turn consists of a stack of color-coded annual citation sections.

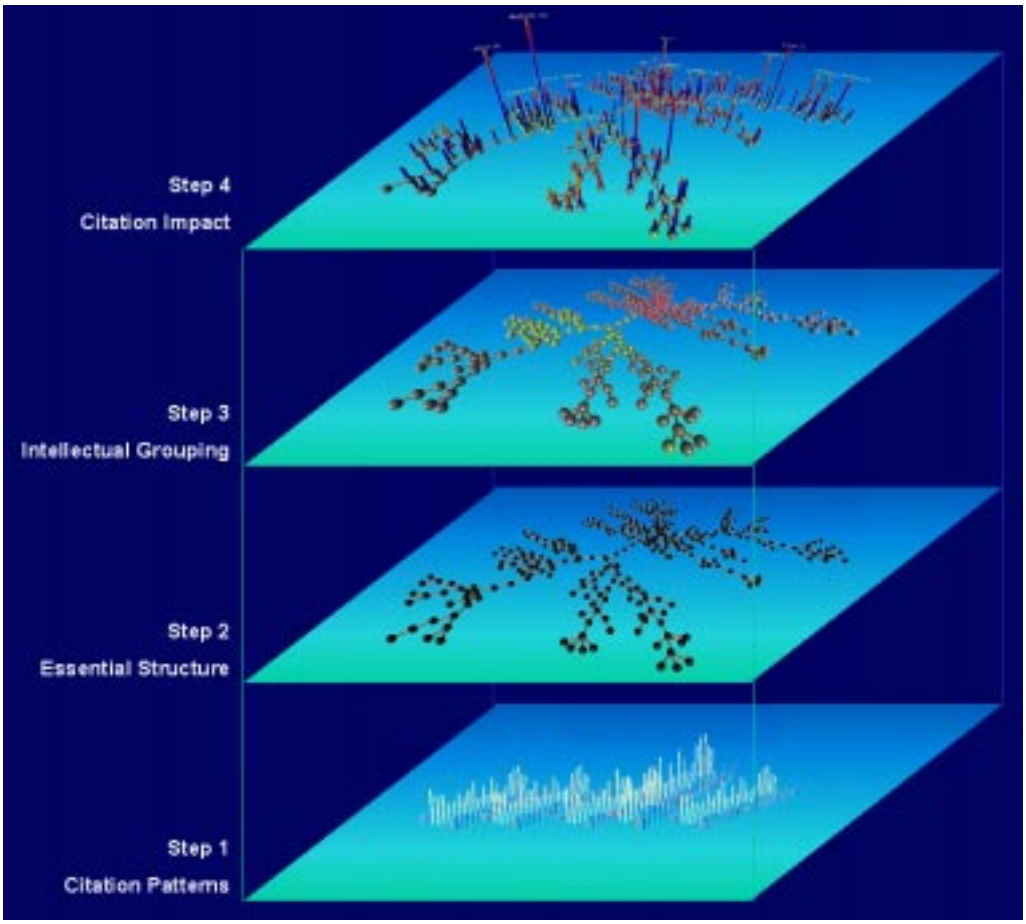

Figure 10. The procedure for visualizing intellectual structures [6]**.**

3.6.1. Intellectual Maps

In order to incorporate multiple aspects of author co-citation networks, we have developed the following design. An author co-citation network is represented as a Pathfinder network, the specialty membership of an author is denoted by color coding based on corresponding factor loading for the author, and the citation history of the author is depicted by a stacked bar.

Figure 11 shows an author co-citation map rendered by this method. The author co-citation map is derived from all the citations appeared in the 1998 ACM Hypertext conference proceedings. A factor analysis on the data found 39 factors, which correspond to 39 specialties in the field of hypertext. In the map, nodes, i.e. authors, are colored by the first three of the 39 factors. The strongest specialty is in red. The next two strongest ones are in green and blue. The strongest specialty branches out from the top of the ring structure, whereas the second strongest specialty

appears to concentrate around the lower left-hand corner of the ring. The color map provides additional information on the context of a particular author's specialty.

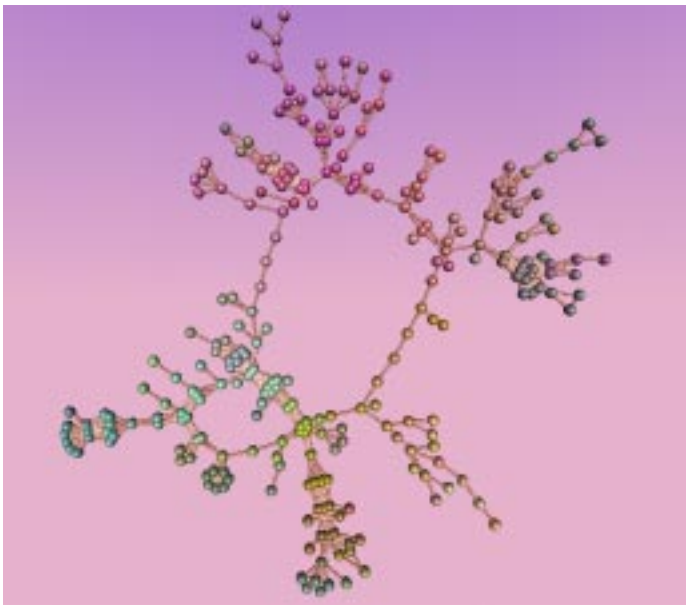

Figure 11. The author co-citation map of hypertext in 1998.

Figure 12 shows an author co-citation map based on author co-citation patterns derived from the proceedings of 1998. It contains 259 authors. Citation indices over three periods are displayed as stacked bars to provide additional cues for understanding the implications of the overall structure. The higher the stacked bar, the more frequently the author has been cited in this sample of the literature. We used the factor loading of the first three of the 18 factors extracted from the co-citation patterns to color the network. The first three factors together explain 42.2% of the variance. Color-coded nodes and links allow users to identify and distinguish authors from different specialties, or invisible colleges.

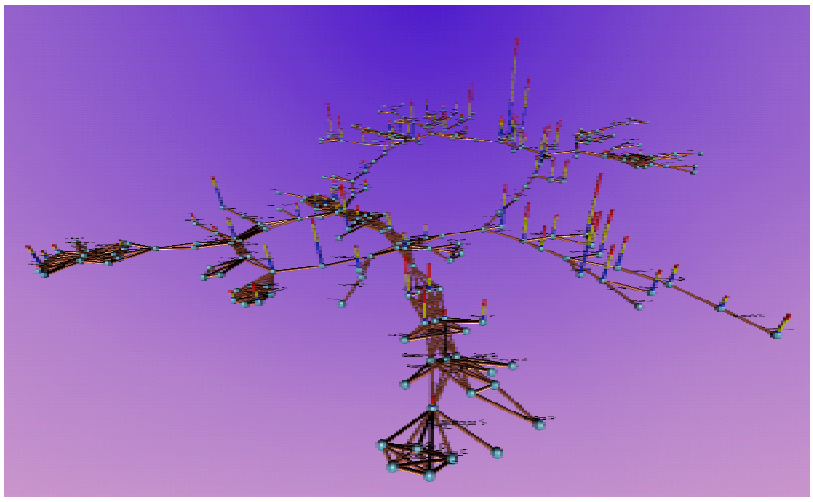

Figure 12. The author co-citation map in 1998 with citation indices displayed as stacked bars to provide additional cues for understanding the implications of the structure. (© 1999 IEEE, reproduced with permission)

We have learned a lot of how to interpret the resultant Pathfinder networks by using different ways of triangulation. It is clear that when it is adapted to the visualization of semantic spaces, there are two types of documents in terms of their distribution in a Pathfinder network: those tend to appear in the center or a relatively fully connected area, and those tend to appear towards the tips of the Pathfinder network. The more central the location, the more likely that the document is of generic nature. The more peripheral the location, the more likely that the document is about a rather specified topic or a perspective.

Longitudinal Maps

A single snapshot of a scientific discipline reveals the structure of the domain at a particular time point. A series of periodical snapshots have the advantage of uncovering the dynamics and trends in the changing structure over a long period of time. Henry Small refers the generation of such periodical snapshots as longitudinal mapping.

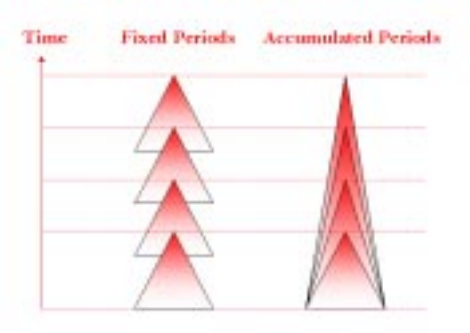

Figure 13. Examples of two different types of periodical snapshots.

If S denotes the set of years of publication corresponding to the source domain of citations to be analyzed and T denotes the set of years of publication for the target domain of corresponding citations, one can think of a variety of ways to arrange a series of periodical snapshots. For example, one can choose to fix the length of each period to 3 years or 5 years to have a moving window of 3 or 5 years across the timeline of the literature. Alternatively, one can choose to accumulate periods one over another so that an increasingly enlarged window of citations will be analyzed. A moving, fixed-size window is more appropriate for revealing the contemporary views at each time point when snapshots are taken, whereas an accumulating, comprehensive window is more suitable to produce a combined big picture.

The following example is based on all the papers appeared in nine ACM Hypertext conferences over a period of ten years (1989-1998). In order to discover significant advances and trends in the history of the field, the series of nine conferences are divided into 3 sub-periods: 1989-1991, 1992-1994, and 1996-1998. Author co-citation analysis was conducted on each of the three individual detests. The domain snapshot in the first period includes 196 authors, the second period reveals 195 authors, and the third period identifies 195 authors.

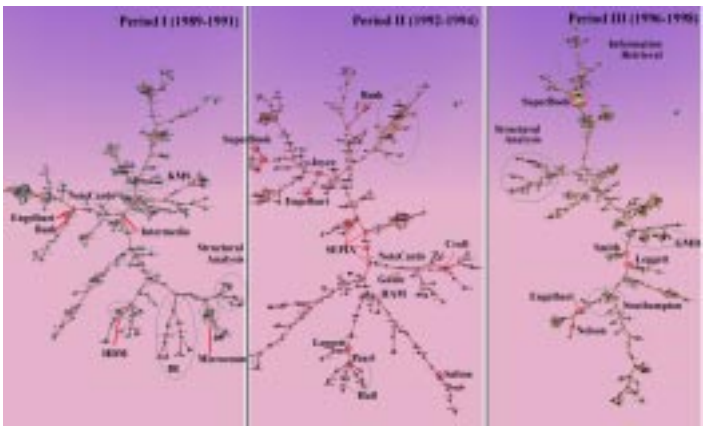

Figure 14. Three snapshots of the evolution of the hypertext literature (1989-1991, 1992-1994, 1996-1998). (© 1999 Springer-Verlag London)

Figure 14 includes three author co-citation maps: the left-hand side map (L) for (1989—1991), the middle map (M) for (1992-1994), and the right-hand side map (R) for (1996—1998). The names of specialties are annotated by hand in these maps. The map L includes 196 authors who have five or more citations during the first period. We followed the links from the map to detailed citation information and found that authors were clearly grouped with reference to papers describing hypertext systems, which are all well-known today, including NoteCards, Intermedia, KMS, and Microcosm. Information retrieval has a much longer history as a field of research. A sub-field in hypertext is rooted in information retrieval. And this specialty of information retrieval was already in place in the first period.

The second period, 1992—1994, was predominated by SEPIA, a famous collaborative hypermedia system developed at GMD, Germany. Six members of GMD occupied the central area of the map M. In this map, Pearl became the branching node for the Microcosm group, and Leggett and the Pearl's branch appeared on the same major branch. We have identified an open hypermedia specialty in the overall author co-citation map. This movement indicated the emergent open hypermedia specialty. Remarkably, Salton and Croft are not in the same major branch in this map. Although we need to check more detailed citation information regarding both Salton's and Croft's work in this period, this example shows that author co-citation maps can be used to highlight the impact of particular aspects of one's work.

The latest period ranges from 1996 through 1998. We expected to identify a specialty of the World-Wide Web in map L, given the apparent influence of the rapid advances of the Web and the growth of the WWW as a research field. However, this is not clear from the map. We need to examine the citation details more thoroughly. The information retrieval specialty is located towards the North of the map. The structural analysis specialty is located to the West of the map.

3.6.2. Towards the Visualization of a Paradigm

Kuhn's work on the structure of scientific revolution has influenced a wide range of scientific disciplines. However, the existence of a paradigm and paradigm shifts in the past have never been shown in an empirical approach. If we can transform the visualization of intellectual structures into a tool that represents a paradigm visually, then such tools will enable scientists to identify more detailed information they need within this context. Furthermore, it will

then be possible to visualize the co-existence of two competing paradigms and a paradigm shift. We are currently investigating this approach. Here we provide a few examples to illustrate our ideas.

In 1988, Halasz addressed seven issues for new generations of hypermedia systems [15]. It was this article that had been the most cited article throughout the ACM Hypertext conferences. In 1997, a spatial hypertext reached the top of the most cited article list. For the first time, the top cited paper was written by researchers other than Halasz, although in 1996 and 1998, another article written by Halasz in 1990 regained the first position.

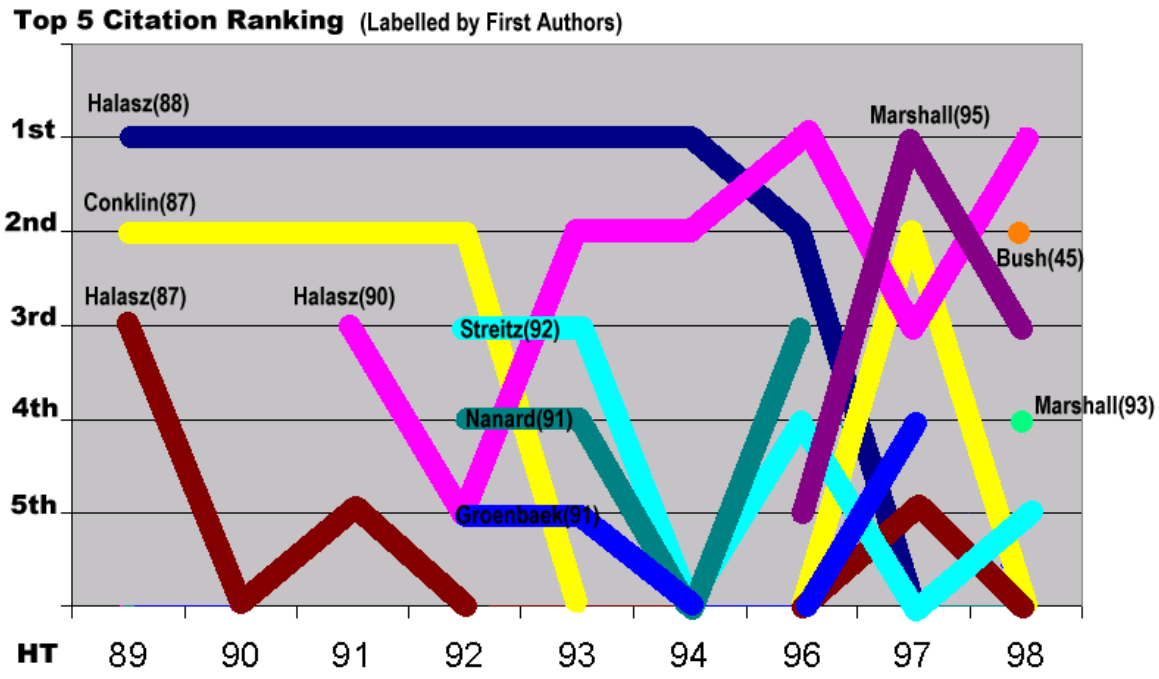

Figure 15. Do we see a paradigm shift?

Figure 16 is a citation and co-citation landscape of the field of hypertext. The landscape was derived from an author co-citation analysis of the ACM Hypertext conference proceedings between 1989 and 1998. The highest peak at the center of the landscape represents the impact of Halasz's work on this community. According to Kuhn's paradigm theory, once a paradigm is established, scientists working within the paradigm do not challenge the assumptions and principles. Does the single peak in the landscape indicate the existence of a paradigm in the field of hypertext? If so, is the paradigm defined by Halasz's articles that have influenced so many researchers in this field? How can we

visualize a paradigm in a discipline in general? Is it possible to visualize a paradigm shift with this visualization framework?

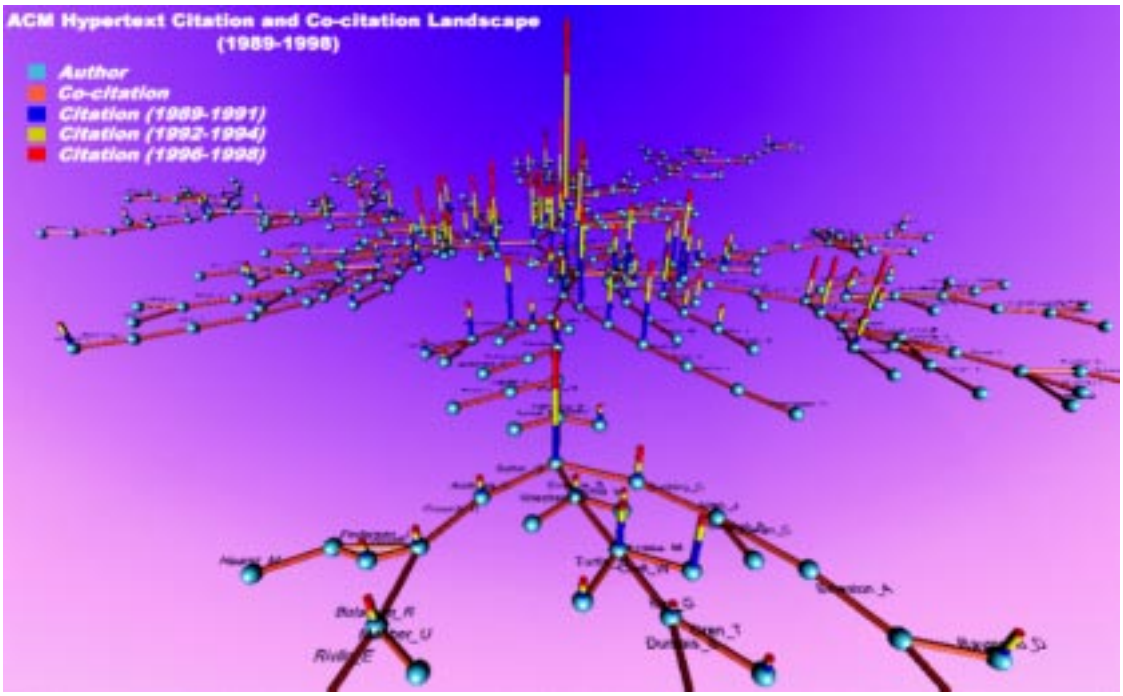

Figure 16. A citation and co-citation landscape of the field of hypertext. Does the single peak in the center indicate the existence of the paradigm since the beginning of this field? Can we track its movement so as to detect a paradigm shift in the future?

If there are several areas that contain authors with very high citation rates, will the paradigm view still be applicable to this domain? If there appear to be several smaller paradigms, does the paradigm view still make sense? In a long run, our ultimate goal is to provide scientists with visualizations that can clearly show the existence of a paradigm or the co-existence of two or more competing paradigms.

4. Challenging Issues

Visualizing knowledge structures is a very promising field of research and its commercial potential is great. Habn and Mani recently identified approaches to automatic text summarization into knowledge-poor and knowledge-rich approaches [14]. Knowledge-poor approaches refer to those essentially relying on statistical and probabilistically methods to solve problems in text summarization, whereas knowledge-rich approaches make use of external sources

of rules and heuristics in the summarization process. In their opinion, at least for the short term, knowledge-poor approaches are likely to dominate applications, particularly when augmented with extraction learning mechanisms. Knowledge-rich approaches will begin to catch up and eventually replace extraction when we have reasonably sized grammars and domain knowledge sources.

Visualizing knowledge structures is in a similar situation – there are a similar division of approaches in terms of whether the underlying knowledge modeling is based on statistically or probabilistically detectable patterns or heuristics and additional information that exists outside the scientific literature. In fact, the majority of knowledge visualization approaches essentially rely on statistical patterns.

Researchers need to address a range of challenging issues to advance theories and techniques for visualizing knowledge structures. In particular, we would like to highlight the following five challenging issues:

1. What is the best way to detect new and significant developments in knowledge?
2. What measures a new paradigm?
3. Can we detect a paradigm shift of science solely based on the reflection of scientific literature?
4. What is the best way to track the growth of knowledge: a landscape of documents or an intellectual network of people?
5. How do we match the visualized intellectual structure to what scientists have in their mind?

Author co-citation analysis (ACA) provides a promising basis for modeling and visualizing knowledge structures reflected in scientific literatures. The greatest advantage of ACA-based approaches is that the strengths associated with particular intellectual bonds between scientists are measured by an integral part of scientific practice. On the other hand, this practice is likely to hinder the process for analysts to obtain the necessary data timely. A more fundamental challenge is rooted in the fact that we are extracting patterns from the scientific literature. A new and potentially significant piece of work might not be captured and recognized in this way, at least very unlikely in knowledge-poor approaches.

Is it possible to track the rise and fall of a scientific paradigm as defined by Thomas Kuhn in his philosophy of science [18]? White and McCain [36] have done a splendid job in their award-winning article to demonstrate how this could be achieved. They have identified information retrieval and citation analysis as two intellectual camps in information science. In order to track the development of science, the ultimate challenge is to detect and predict the shift of a paradigm as it is taking place.

Judging the extent to which a visualized knowledge structure matches or mismatches to what leading scientists have in their mind is yet another fundamental challenge. To meet this challenge, we must equip ourselves with a deeper understanding of the nature of knowledge and cognition. This is likely to be an interdisciplinary endeavor, involving disciplines such as epistemology, sociology of knowledge, and philosophy of science.

5. Conclusions

In this chapter, we have highlighted the history and the state of the art of visualizations of knowledge structures. We have emphasizes the central idea of painting the big picture of science and its profound connections to philosophy of science, especially Kuhn's structure of scientific revolution. Some of the most influential theories and methods of analysis, modeling, and visualization have been included. Challenging issues to be resolved in the future are discussed.

In conclusion, visualizing knowledge structures is a challenging but ultimately rewarding route to capture the essence of a scientific paradigm. Potential applications will be valuable to multiple domains. These are tough challenges, but the potential benefits are tremendous and profound.

References

1. Bookstein, A., Klein, S.T. and Raita, T., Detecting content-bearing words by serial clustering. in *SIGIR '95*, (Seattle, WA., 1995), ACM Press, 319-327.
2. Card, S., Mackinlay, J. and Shneiderman, B. (eds.). *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, 1999.
3. Chen, C. *Information Visualisation and Virtual Environments*. Springer-Verlag, London, 1999.
4. Chen, C. Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management, 35* (2). 401-420.
5. Chen, C. and Carr, L., Visualizing the evolution of a subject domain: A case study. in *IEEE Visualization '99*, (San Francisco, CA, 1999).
6. Chen, C. and Paul, R.J. Visualizing a knowledge domain's intellectual structure. *Computer*, *34* (3). 65-71.
7. Chen, H., Houston, A.L., Sewell, R.R. and Schatz, B.R. Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science, 49* (7). 582-608.

8. Davidson, G.S., HENDRICKSON, B., K.JOHNSON, D., MEYERS, C.E. and WYLIE, B.N. Knowledge mining with VxInsight: Discovery through interaction. *Journal of Intelligent Information Systems*, *11* (3). 259-285.
9. Fairchild, K., Poltrock, S. and Furnas, G. SemNet: Three-dimensional graphic representations of large knowledge bases. in Guidon, R. ed. *Cognitive Science and its Applications for Human-Computer Interaction*, Lawrence Erlbaum Associates, 1988, 201-233.
10. Garfield, E. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122* (108-111).
11. Garfield, E. Scientography: Mapping the tracks of science. *Current Contents: Social & Behavioural Sciences*, *7* (45). 5-10.
12. Garfield, E., H., S.I. and Torpie, R.J. *The use of citation data in writing the history of science*. Institute for Scientific Information, Philadelphia, 1964.
13. Garfield, E. and Small, H. Identifying the changing frontiers of science, The S. Neaman Press, 1989.
14. Habn, U. and Mani, I. The challenges of automatic summarization. *IEEE Computer*, *33* (11). 29-36.
15. Halasz, F. Reflections on NoteCards: Seven issues for the next generation of hypermedia systems. *Communications of the ACM*, *31* (7). 836-852.
16. Hjorland, B. *Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science*. Greenwood Press, Westport, 1997.
17. Kruskal, J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*. 1-27.
18. Kuhn, T.S. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 1962.
19. Lin, X., Soergel, D. and Marchionini, G., A self-organizing semantic map for information retrieval. in *SIGIR '91*, (Chicago, IL, 1991), ACM Press, 262-269.
20. McCain, K.W. Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, *41* (6). 433-443.
21. Munzner, T., H3: Laying out large directed graphs in 3D hyperbolic space. in *the 1997 IEEE Symposium on Information Visualization*, (Phoenix, AZ, 1997), IEEE, 2-10.
22. Neal, R. *Bayesian Learning for Neural Networks*. Springer-Verlag, New York, 1996.
23. Price, D. Networks of scientific papers. *Science*, *149*. 510-515.
24. Price, D.D. *Little Science, Big Science and Beyond*. Columbia University Press, New York, 1986.
25. Price, D.D. Networks of scientific papers. *Science*, *149*. 510-515.
26. Robertson, G.G., Mackinlay, J.D. and Card, S.K., Cone trees: Animated 3D visualizations of hierarchical information. in *CHI '91*, (New Orleans, LA, 1991), 189-194.
27. Salton, G. Developments in automatic text retrieval. *Science*, *253*. 974-980.
28. Schvaneveldt, R.W. (ed.), *Pathfinder Associative Networks: Studies in Knowledge Organization*. Ablex Publishing Corporations, Norwood, New Jersey, 1990.

29. Schvaneveldt, R.W., Durso, F.T. and Dearholt, D.W. Network structures in proximity data. in Bower, G. ed. *The Psychology of Learning and Motivation, 24*, Academic Press, 1989, 249-284.
30. Sher, I. and Garfield, E., New tools for improving and evaluating the effectiveness of research. in *Research Program Effectiveness*, (Washington, D.C., 1966), Gordon and Breach, 135-146.
31. Small, H. Co-citation in scientific literature: A new measure of the relationship between publications. *Journal of the American Society for Information Science, 24*. 265-269.
32. Small, H. A co-citation model of a scientific specialty: A longitudinal study of collagen research. *Scoial Studies of Science, 7*. 139-166.
33. Small, H. A passage through science: Crossing disciplinary boundaries. *Library Trends, 48* (1). 72-108.
34. Small, H. Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50* (9). 799-813.
35. White, H.D. and Griffith, B.C. Author co-citation: A literature measure of intellectual structure. *Journal of the American Society for Information Science, 32*. 163-172.
36. White, H.D. and McCain, K.W. Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science, 49* (4). 327-356.
37. Wise, J.A. The ecological approach to text visualization. *Journal of the American Society for Information Science, 50* (13). 1224-1233.
38. Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A. and Crow, V., Visualizing the non-visual: Spatial analysis and interaction with information from text documents. in *IEEE Symposium on Information Visualization '95*, (Atlanta, Georgia, USA, 1995), IEEE Computer Society Press.

List of Symbols

$tf \times idf$

$\mathbf{T_k}$

$\mathbf{D_i}$

tf_{ik}

n_k

w_{ij}

$p\,(x \mid \theta)$

$p\,(x \mid \theta')$

$p\,(\theta)$

$p\,(\theta')$

$p\,(\theta \mid x)$

p_I

$p(n, T)$

$El = \Sigma_n\, n \times p(n, T)$

$\xi(s)$

$ccv(t)$

$\lambda(\{S_\alpha\})$