Detecting and Mapping Thematic Changes in Transient Networks

Chaomei Chen
College of Information Science and Technology
Drexel University
3141 Chestnut Street, Philadelphia, PA 19104 USA
chaomei.chen@cis.drexel.edu

Abstract

Research fronts are the forefront of a scientific field. Timely identifying emerging trends and abrupt changes in scientific literature is not only beneficial for users of digital libraries, but also instrumental for users to trace the movement of a research front. A research front can be seen as a dynamic and transient mapping from the state of the art to its immediate prior art - the intellectual base. Citation and co-citation analysis primarily focuses on the structure and dynamics of the intellectual base, information retrieval and related research mainly focuses on intrinsic properties of text documents. In this article, we describe a novel approach that enables us to detect and visualize the transient relationship over time. Temporal associations between content words connected to a surge of interest in scientific papers and high-impact articles in the intellectual base are identified and visualized as a time-variant network. We apply the approach to the detection of thematic changes over time in information visualization and research in terrorism.

1. Introduction

A *research front* in a scientific field represents the state of the art of the collective knowledge of a scientific community [1]. A fundamental way for a research front to link to its immediate prior art – *the intellectual base* - is via citations in scientific publications, i.e. via references made by scientists in their papers. Timely identifying emerging trends and abrupt changes associated with a research front is not only beneficial for users of digital libraries, but also instrumental for users to trace the movement of a research front and better utilize resources and services provided in digital libraries.

Relevant research falls into two broad categories: the study of the intellectual base, and the study of text documents. The most representative work of the former is citation and co-citation analysis, which typically focuses on the extent to which scientific publications have influenced subsequent publications in terms of citation impact. The most representative work in the latter is research in information retrieval, especially including a variety of theories and techniques developed for detecting thematic trends in text documents. However, few studies have quantitatively investigated the interconnection between the contents of the most recent scientific publications and the intellectual base. To our knowledge, citing documents and cited documents have not been studied and represented in a coherent and unifying framework.

Studying the relationship has several practical implications. One of the limitations of citation-based methods is the delay caused by the very nature of citation data; citing articles must go through peer review and publication processes before citations made in these articles become countable. Citation data alone cannot tell us anything more recent than when they are produced. The latest trends are more likely to be detected by studying recent text documents. On the other hand, citations provide valuable information on how the scientific community as a whole reacts to particular publications, and such acceptance takes time. Such information is beyond the reach of statistical models of intrinsic properties of text documents. In other words, a high frequency term may be associated with a surge of interest, but it is not necessarily a good predictor of a subsequent popularity by peer scientists. In this article, we describe a novel approach that enables us to detect and visualize the transient relationship between newly published scientific papers and their intellectual base.

The rest of the article is organized as follows. We first introduce related work, including co-citation analysis, content word analysis, detecting abrupt changes and emerging trends, and visualizing changes over time. Then we explain our approach in detail and demonstrate the use of this approach in detecting and visualizing changes in research on information visualization and on terrorism.

IEEE COMPUTER SOCIETY

2. Related work

Quantitative approaches to the study of science have been mainly developed by the scientometrics community, including pioneer works such as [1-7]. The major problem now is the lack of information and knowledge management techniques that enable scientists and a variety of digital library users to detect and track significant changes over time as part of the big picture of science.

2.1. Co-citation analysis and word analysis

Derek Price [1] suggested that scientific literature contains two distinct parts: a classic part and a transient part. The make-up of these two parts varies from field to field; mathematics, for example, is strongly predominated by the classic part, whereas the transient literature rules in physics. Furthermore, he introduced the notion of *research fronts* – the transient part of scientific literature - as a collection of papers that have draw much of the attention of scientists. Based on an examination of citation patterns of scientific papers, he conjectured that it is possible to identify objectively defined subjects in citation networks. He particularly emphasized the significance of understanding the nature of such moving frontiers in the development of a quantitative method for delineating the topography of current scientific literature. In this article, we consider research fronts as a transient mapping between the most salient themes in the latest publications and highly cited publications corresponding to such themes.

Research has shown that clusters of highly co-cited articles can reveal the extent a field of study shifts its focus [5, 8, 9]. Small and Griffith examined issues concerned with identifying specialties by mapping the structure of scientific literatures, especially through analyses of co-citation networks [10]. Small subsequently found rapid changes of focus in collagen research [5].

Besides co-citation analysis, co-word analysis is a method initially developed in 1960s [11]. Co-word analysis frequently uses the so-called inclusion index, which measures the extent to which the appearance of one word in a text document predicts the appearance of another. However, few techniques are available for conducting co-citation analysis and content word analysis in a coherent and integrative framework.

The most notably exception of the lack of combined approaches to co-citation analysis and content word analysis is the study conducted more than a decade ago by Braam, Moed, and van Raan [3]. They studied the literature of atomic and molecular physics over a period of 10 years by combining co-citation analysis and content word analysis. However, more precisely, they combined the results of co-citation analysis and content word analysis. Content words were extracted from articles citing papers in co-citation clusters.

The focus of their study was on co-citation clusters rather than co-citation networks of documents per se. Indeed, their assumption is that content words reflect specialties more directly than co-citation clusters; therefore, they evaluated the comprehensiveness of specialties as represented by co-citation clusters by examining the congruence between the results of the two methods, i.e. the extent to which they converge.

In summary, there is the lack of techniques that enable users to handle the transient and volatile nature of scientific frontiers. Word analysis can reveal significant trends and abrupt changes in the latest publications, but it must make a connection between such properties and articles identified by co-citation analysis.

2.2. Detecting abrupt thematic changes

Topic Detection and Tracking (TDT) projects aim to develop and evaluate technologies required to segment, detect, and track topical information in a stream consisting of news stories. TDT addresses five types of tasks: 1) story segmentation, 2) topic detection, 3) topic tracking, 4) first story detection, and 5) story link detection. Topic detection focuses on discovering previously unseen topics, whereas topic tracking focuses on monitoring stories known to a TDT system. Roy, Gevry, and Pottenger [12] presented methodologies for trend detection. Kontostathis, Galitsky, Pottenger, Roy, and Phelps [13] gave a comprehensive survey of emerging trend detection in textual data mining.

Allan, Papka, and Lavrenko [14] presented a single pass clustering algorithm, in which the entire content of each story is represented as a query. A newly arrived story is compared with all the existing stories. If the new story is similar enough to one of the existing ones, then to the algorithm, it is really nothing new. In contrast, if no match was found between the new story and existing stories, then the newly arrived story is marked as something completely new. It was reported that the single pass algorithm can distinguish the Oklahoma City Bombing stories from the earlier stories on the Word Trade Center bombing. However, their algorithm could not distinguish OJ Simpson trial stories from other court cases.

Swan and Allan [15] constructed a statistical model that can detect the arrival of significant information. The key to their approach is to use a χ^2 distribution to identify the probability of a single random word passing the χ^2 test and thereby to determine whether the appearance of a feature is purely by chance or not. Their approach was tested with a subset of the Topic Detection and Tracking (TDT) pilot study's corpus, which contains CNN broadcast news and Reuters newswires.

2.3. Burst Detection

An interesting observation of news stories is that events are often reported in burst. Yang, Pierce, and Carbonell [16] depicted a daily histogram of story counts over time. News stories about the same event tend to appear within a very narrow timeframe. The gap between two bursts can be used to discriminate distinct events. More recently, Jon Kleinberg developed a burst detection algorithm and applied to the arrivals of email and words used in titles of articles [17].

Kleinberg's method is based on the assumption that whenever an important event occurs or is about to occur, there should be a sharp increase of certain words that characterize the event. He called such sharp increases *bursts*. Essentially, Kleinberg's burst detection algorithm analyzes the rate of increase of word frequencies and identifies the most rapidly growing words. He tested his algorithm on the full text of all the State of the Union addresses since 1790. The burst detection algorithm identified important events occurring at the time of some of the speeches. For example, *depression* and *recovery* were bursty words in 1930-1937, *fighting* and *Japanese* were bursty in 1942-1945, and *atomic* was the buzz word in 1947 and 1959.

The major advantage of Kleinberg's approach is that it can handle multiple levels of bursts. There are alternative approaches, such as wavelet analysis, Hidden Markov Models (HMMs) [18], and machine-learning and other data mining and knowledge discovery algorithms. Data mining and knowledge algorithms provide a rich source of resources that can be used to enhance information visualization. Algorithms that can detect abrupt changes with the least number of false alarms are certainly useful to filter out noises from the original data so that subsequent modeling and visualization can produce much improved and sharp focused information visualization.

2.4. Visualizing Changes

ThemeRiver [19] is a visualization system that uses the metaphor of a river to depict thematic flows over time in a collection of documents. The thematic changes are shown along a time line of corresponding external events. A thematic river consists of frequency streams of terms; the changing width of a stream over time indicates the changes of term occurrences. The occurrence of an external event may be followed by sudden changes of thematic strengths. On the one hand, searching for an abruptly widened thematic stream is a much more intuitive task to detect a new story than text-based TDT systems that can only report changes in terms of statistics. On the other hand, ThemeRiver does not incorporate additional data filters that would make changes in thematic patterns easy to detect.

Powerful data modeling techniques can make salient patterns clearer and in sharper contrast. Some of the major technical problems are how to make the changes over time easy to understand and how to preserve the overall context in which changes take place.

Web ecology is concerned with relationships among users and their information environment, and its evaluation over time in terms of the content, the topological properties, and the usage [20]. Disk Tree represents the hyperlink structure of a website. It uses breadth-first search (BFS) algorithm to extract a tree structure from the original network, and then visualizes the tree structure. Time Tube shows an array of disk trees to show the evolution of a website. They are designed to support tasks such as comparison and trend detection.

Brandes and Corman [21] used semi-transparent layers one on another to display the progress of a discourse. Erten, Harding, Kobourov, Wampler, and Yee [22] reported an interesting example of temporal graph visualization of the categorization of papers in ACM conference proceedings. The category graph for a given time period is an undirected graph. Vertices represent categories and edges represent categories that co-occurred in these conference papers. The weight of a node represents the concentration of research on a category. The weight of an edge represents the strength of the relation between two categories.

Chen and Morris [23] demonstrated an animated visualization technique that can display the process of network evolution. A Pathfinder network version and a minimum spanning tree (MST) version were compared in a study of co-citation networks over a time span of nearly 60 years. Although the MST version contained many hubs, the animation revealed that the Pathfinder version maintained the local accuracy of network evolution in a much better way. Animated visualization techniques were also used to depict how citation landscapes change over time [24].

More recently, we reported the CITESPACE system for co-citation analysis and visualization [25]. It was designed to detect intellectual turning points in scientific literature, especially to help users to identify articles that triggered conceptual revolutions. CITESPACE supports a number of analyzing and visualizing methods, namely time slicing, thresholding, modeling, pruning, merging, and mapping.

However, existing visualization methods are still inadequate to help users to identify significant changes. For example, how many citations to a scientific publication in a year should count as a significant increase? Straightforward counting cannot take us very far. Visualizations based on ambiguous information cannot give sharp and clear-cut images. Furthermore, to what extent can we compensate the citation latency by indicators that could be detected earlier, for example, a surge of interest detected based on the abstracts of freshly published articles?

In this article, we describe a novel method that combines co-citation analysis and text analysis of content words. We build upon the co-citation network visualization techniques developed in CiteSpace and establish a generic framework so that heterogeneous aspects of scientific publications can be accommodated in harmony (See Figure 1). The method is also extensible to incorporate additional components from scientific publications such as authors and authors' academic affiliations as well as their geographic locations.

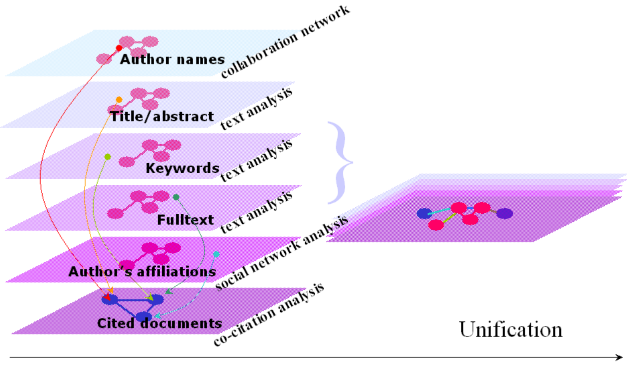

Figure 1 Our approach enables a unification of various quantitative methods for studying scientific literature.

3. Method

The procedure of our approach consists of the following steps. First, choose a topic of interest. In this article, we demonstrate two examples: one on information visualization and the other on terrorism or terrorist. Second, compile datasets of bibliographic records of scientific papers that match the chosen topic. In particular, each record contains the title, the abstract, information on authors, keywords provided by authors as well as keywords provided by indexers, and a list of references. The two illustrative examples are based on science citation data compiled by ISI's Science Citation Index (SCI).

The next step is to select only those content words that have shown a surge of usage during the time interval of the retrieved data. Both our examples in this article range from 1990 through January 2004. Content words can be selected from titles, abstracts, and keyword fields such as descriptors and identifiers. Our experience suggests title words and abstract words together tend to provide more meaningful results than any other options. For example, title words alone or descriptors alone do not always lead to additional insights in resultant maps because of the relatively lower prevalence of such words.

Content words of selected types are passed through abrupt change detection algorithms. The goal is to identify a subset of words that have demonstrated a sufficient surge of popularity. After the September 11 terrorist attacks, one of the abruptly increased words in scientific papers on terrorism is the abbreviation of Post-Traumatic Stress Disorder (PTSD). In this study, we used Kleinber's burst detection algorithm to detect the surge of a word in citing articles. The algorithm also identifies the span of a burst. Headline news may last days or weeks, whereas a scientific topic may last years and even decades.

The set of burst words and a subset of sufficiently cited articles from the original dateset are used to form a series of heterogeneous matrices. These matrices are subsequently visualized as networks and synthesized as an overall visualization. The selection of cited articles essentially follows the same procedure as described in [25]. The entire time interval can be divided into several equal-length time slices. Within each sliced sub-interval, one can specify a citation threshold that an article must meet in order to be included in subsequent analysis; similarly, one can specify a co-citation threshold that an article must be co-cited with other articles frequently enough to be included in the picture. If the frequency of a word bursts during the interval of a time slice, i.e. if there is an overlap between the burst span and the time slice interval, then the word will become part of the matrix corresponding to the time slice; otherwise, the word will be excluded. Thus, each time slice is represented by words that have abrupt increases in terms of their frequencies within the same period of time.

Given a time slice, the corresponding matrix consists of citing words and cited articles. The word-word associations in such matrices are determined by co-occurrences of a pair of words in a citing document, for example, in the title and the abstract. The article-article associations are defined by co-citations. The word-article connections indicate that the word is found in articles citing the target article and that there is a surge of the use of this particular word within the time span of the analysis.

Each matrix defines a network with citing words and cited articles as nodes. Excessive and redundant links in such networks are pruned by Pathfinder network scaling algorithm, also explained in [25]. Briefly speaking, Pathfinder network scaling imposes a triangle inequality condition over a network. It verifies if a link satisfies this condition and removes links that violate the condition.

4. Results

Two examples are included in this article. The first example is on information visualization (1990-2003) and the second example is on terrorism/terrorist (1990-2004).

4.2. Example A: Information Visualization

The Information Visualization dataset contains 193 citing articles published between 1990 and 2004 on the topic. We only removed a dozen of the most common stopwords such as *the*, *an*, *how*, and *where*. We expected that abrupt change detection would further eliminate most stopwords if not all. As a result, among 3,777 content words from titles, abstracts, descriptors, and identifiers of the 193 articles, 654 words are retained because of detected bursts. Abstract words turn out to be significant. Removing other types of words hardly made any difference in terms of the burst words. The merged network contains 62 nodes. Thresholds for co-citation analysis are shown in Table 1.

Table 1 Information Visualization (1991-2004): Time slices and thresholds (c for citation, cc for co-citation, ccv for cosine coefficients).

2-year slices	c | cc | ccv	space	nodes	links
[1991, 1992]	1 | 1 | 0.01	16	0	0
[1993, 1994]	1 | 1 | 0.01	67	7	21
[1995, 1996]	2 | 1 | 0.01	303	9	34
[1997, 1998]	3 | 2 | 0.01	600	12	54
[1999, 2000]	3 | 2 | 0.01	1148	21	117
[2001, 2002]	3 | 2 | 0.01	1443	19	27
[2003, 2004]	3 | 2 | 0.01	952	8	12

Figure 2 shows a synthesized network over seven time slices between 1991 and 2004. Citing words are displayed in red. The size of a word is proportional to its frequency in the first year of its burst span. The color of a link indicates the first year that the strength of the link exceeds corresponding thresholds. The thickness of a link is proportional to the strength of the link. The thicker a link is shown, the stronger the co-citation association is. Cited articles are labeled in dark, showing the author's name, the year of publication, volume numbers and page numbers. If it is a book, it is shown with a VBOOK in its label. The font size of an article label is also proportional to its citations. Larger sized labels mean more frequently cited articles. The number at the lower left corner of a label is the citation count when the article passed the thresholds for the first time.

Blue links are links appearing in the earliest time slice, whereas red links are links that didn't appear until the last time slice. The earliest links in the Information Visualization map are in purple in the lower right corner, corresponding to the slice of 1993 and 1994. Links added in the next slice (1995-1996) are further down in green. Links emerged between 1997 and 1998 form a dense pink cluster to the left, including the 1993 article by Robertson and his colleagues on 3D interactive visualization. Several citing words surged during this period, including *graph* (found in 30 articles that cited Robertson 1993 paper), *multimedia* (30), *drawing* (30), and *search* (30). Additions in the following slice (1999-2000) form the brown cluster in the center of the image, containing graph drawing articles by Tamassia (1987), Dibattista (1994), and Hendley (1995). Note that these are the first authors only as they are shown in SCI's format. The high-profile citing words in this cluster include *patterns* in 39 articles, *project* (36), and *drawings* (36). The bright yellow cluster centered by *knowledge* corresponds to the period of 2001 and 2002. The word *knowledge* appeared in 61 articles within this short period. Other words such as multivariate (33), life (29) and cycle (29) are also prominent in this cluster. The few red links were added during the latest time slice between 2003 and 2004.

Figure 3 shows a pruned version of the network shown in Figure 2. Pathfinder network scaling was used to remove redundant links and preserve the most salient links only. Thus, networks in the two figures have the identical set of articles and terms, but the pruned network shown in Figure 3 has fewer links than the one in Figure 2. Because CITESPACE redraws the network completely after pruning, the position of a node is likely to change in the new layout. One can locate a cluster in the pruned version based on the colors of links. For example, the yellow cluster centered by word *knowledge* in the original network shown in Figure 2 is transformed to a yellow star-like structure in the pruned version.

In CITESPACE, users can interactively scale the font sizes of labels. The branch at the lower left corner shows that articles to do with visualizing *multivariate* information have particularly cited Ware's 2000 book on information visualization and Inselberg's 1990 article, which in turn is often co-cited with Chernoff's 1973 article. These two articles are indeed about multivariate information visualization. Furthermore, the presence of the burst word *multivariate* told us that this pattern in fact emerged in the latest time slice (2003-2004).

4.3. Example B: Terrorism/Terrorists

The Terrorism/Terrorist dataset (1990-2004) contains 15,887 content words, including title words, abstract words, descriptors, and identifiers. Among them, 235 words were detected to have significant bursts. A merged network is shown, containing 91-nodes. Time slices and thresholds used are detailed in Table 2.

In addition to burst detection for words, we experimented burst detection for citations and expected that articles with sudden citation jumps should be representative in an image of a research front. The most meaningful ranking was given by the reciprocal of burst weight in Kleinberg's algorithm, shown as r in Table 3.

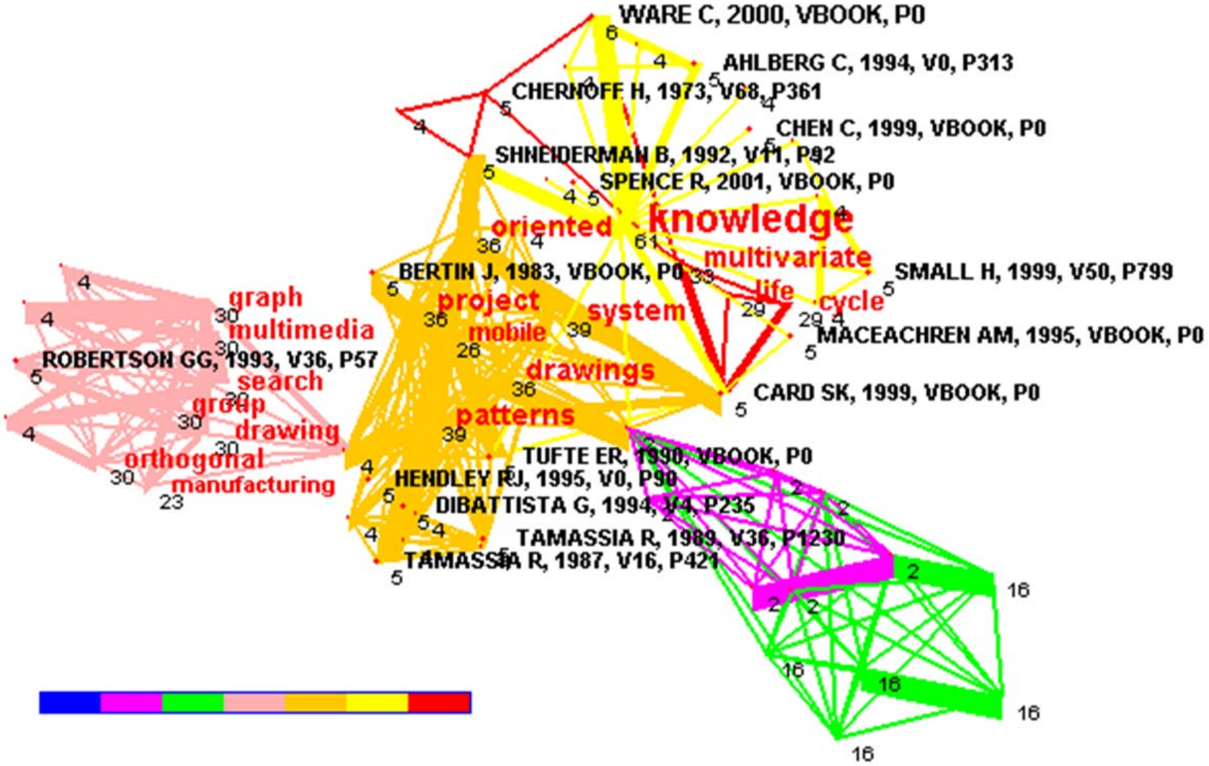
WARE C, 2000, VBOOK, P0
AHLBERG C, 1994, V0, P313
CHERNOFF H, 1973, V68, P361
CHEN C, 1999, VBOOK, P0
SHNEIDERMAN B, 1992, V11, P92
SPENCE R, 2001, VBOOK, P0
oriented
knowledge
multivariate
BERTIN J, 1983, VBOOK, P0
SMALL H, 1999, V50, P799
life
cycle
graph
project
system
multimedia
mobile
MACEACHREN AM, 1995, VBOOK, P0
ROBERTSON GG, 1993, V36, P57
drawings
search
group
CARD SK, 1999, VBOOK, P0
patterns
drawing
orthogonal
TUFTE ER, 1990, VBOOK, P0
manufacturing
HENDLEY RJ, 1995, V0, P90
DIBATTISTA G, 1994, V4, P235
TAMASSIA R, 1989, V36, P1230
TAMASSIA R, 1987, V16, P421

Figure 2 Frequently cited articles and trendy citing words in Information Visualization (1991-2004).

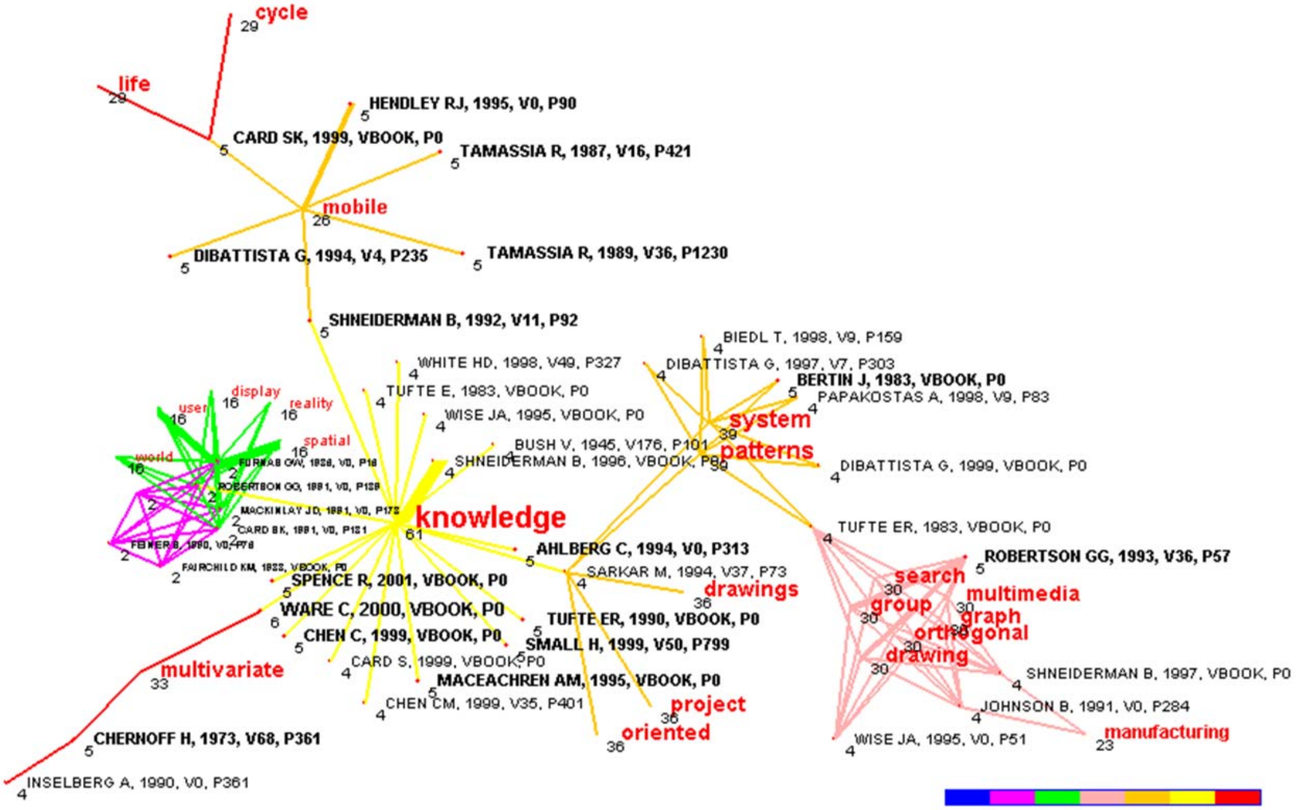
cycle
life
HENDLEY RJ, 1995, V0, P90
CARD SK, 1999, VBOOK, P0
TAMASSIA R, 1987, V16, P421
mobile
DIBATTISTA G, 1994, V4, P235
TAMASSIA R, 1989, V36, P1230
SHNEIDERMAN B, 1992, V11, P92
BIEDL T, 1998, V9, P159
WHITE HD, 1998, V49, P327
DIBATTISTA G, 1997, V7, P303
BERTIN J, 1983, VBOOK, P0
TUFTE E, 1983, VBOOK, P0
PAPAKOSTAS A, 1998, V9, P83
display
user
reality
WISE JA, 1995, VBOOK, P0
system
spatial
BUSH V, 1945, V176, P101
patterns
world
FURNAS GW, 1986, V0, P16
SHNEIDERMAN B, 1996, VBOOK, P0
DIBATTISTA G, 1999, VBOOK, P0
ROBERTSON GG, 1991, V0, P189
MACKINLAY JD, 1991, V0, P173
knowledge
CARD SK, 1991, V0, P181
TUFTE ER, 1983, VBOOK, P0
FEINER S, 1990, V0, P76
AHLBERG C, 1994, V0, P313
search
ROBERTSON GG, 1993, V36, P57
FAIRCHILD KM, 1988, VBOOK, P0
SARKAR M, 1994, V37, P73
multimedia
SPENCE R, 2001, VBOOK, P0
drawings
group
graph
WARE C, 2000, VBOOK, P0
TUFTE ER, 1990, VBOOK, P0
orthogonal
CHEN C, 1999, VBOOK, P0
SMALL H, 1999, V50, P799
drawing
CARD S, 1999, VBOOK, P0
multivariate
SHNEIDERMAN B, 1997, VBOOK, P0
MACEACHREN AM, 1995, VBOOK, P0
JOHNSON B, 1991, V0, P284
CHEN CM, 1999, V35, P401
project
oriented
manufacturing
CHERNOFF H, 1973, V68, P361
WISE JA, 1995, V0, P51
INSELBERG A, 1990, V0, P361

Figure 3 A Pathfinder-pruned version of the network shown in Figure 2.

The sharpest citation jump was detected with Schuster's 2001 article. Again, these are the first authors only. This article started off with 14 citations in 2002 and moved to 24 in 2003. Compared with other articles in the table, this was indeed an abrupt raise. The Franz 1997 article has the longest burst span since 1999. It appears other articles in the table tend to take off around the third year of their publication. All articles in the table are featured in the visualization (Figure 4 and Figure 5).

We include two versions of the network: first, the one without pruning and then the one with pruning. The map shows seven clusters in distinct colors, corresponding to seven time slices. Unlike the Information Visualization network, these clusters are isolated from one another at this level. Terrorism/terrorist research may not like information visualization as a coherent scientific subject matter; rather, research in terrorism and terrorist activities tends to be strongly driven by external events.

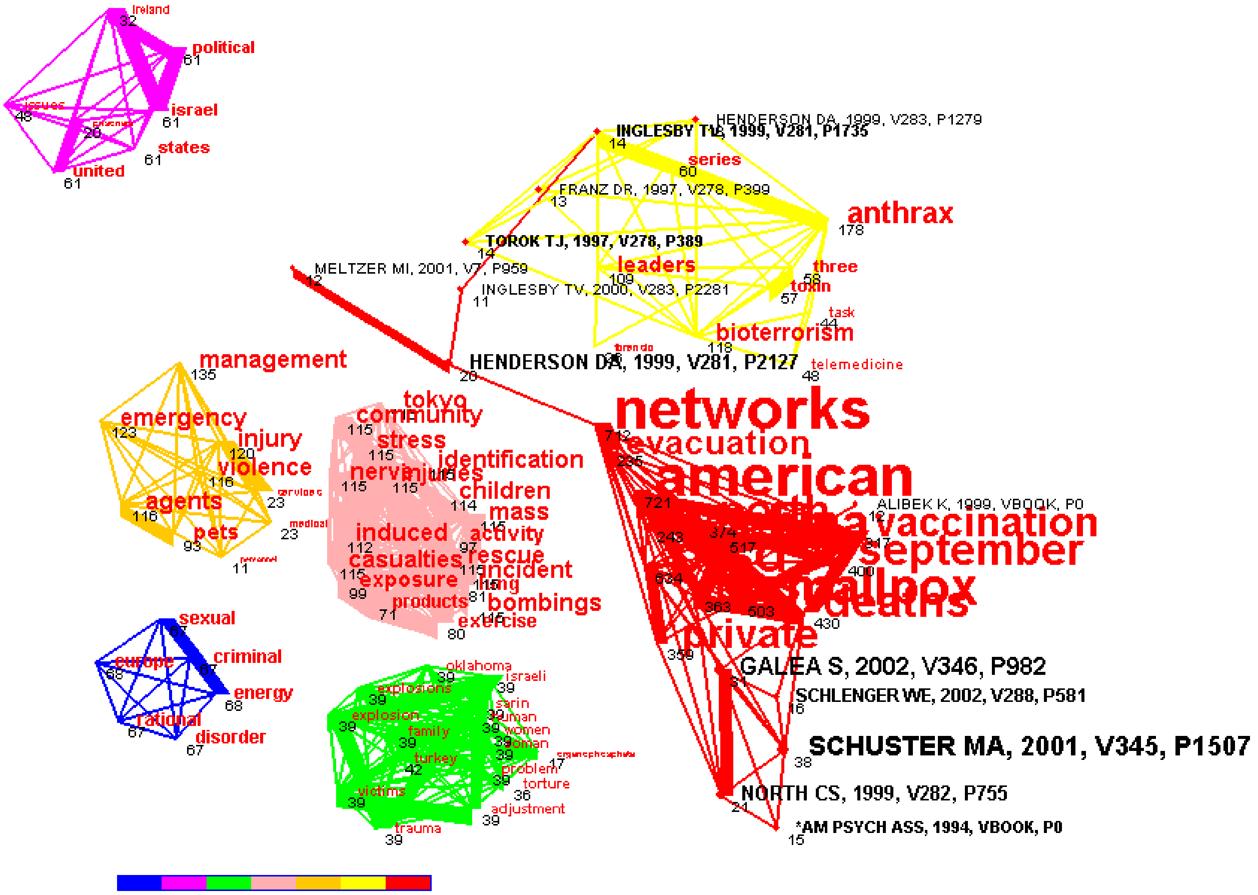

Figure 4 Thematic changes in the strongly event-driven terrorism/terrorist literature over time (1990-2004), including the 1995 Oklahoma City bombing (green: 1994-1995), the 1995 Tokyo subway sarin attack (pink: 1996-1997), anthrax and bioterrorism (yellow: 2000-2001), and the September 11 terrorist attacks (red: 2002-2003).

This dataset generated a much larger pool of content words (15,887). The resultant map contains many more words than the Information Visualization map. The green cluster (1994-1995), for example, contains *Oklahoma*, *Turkey*, *Israeli*, and *sarin*. The Oklahoma City Bombing was in April 1995. There was a major increase in non-lethal terrorist attacks against property in Turkey in 1995 by the Kurdistan Workers' Party (PKK). The PKK also committed lethal acts of terrorism. Israeli Prime Minister Yitzhak Rabin was assassinated by a Jewish Israeli extremist in November. In March 1995, members of the Japanese cult Aum Shinrikyo placed containers of the deadly chemical nerve agent sarin on five trains of the Tokyo subway system during the morning rush hour, killing 12 persons. The attack was the first major use of chemical weapons by terrorists.

Table 2. Terrorism/terrorist data: Time slices and corresponding articles selected.

2-year slices	c | cc | ccv	space	nodes	links
[1990, 1991]	20 | 1 | 0.15	1440	6	15
[1992, 1993]	16 | 2 | 0.15	1355	7	21
[1994, 1995]	13 | 3 | 0.15	1522	16	120
[1996, 1997]	10 | 5 | 0.15	2710	18	153
[1998, 1999]	10 | 5 | 0.15	2993	9	33
[2000, 2001]	10 | 5 | 0.15	5436	13	37
[2002, 2003]	10 | 5 | 0.15	17835	27	91

The bright yellow cluster in the upper half of the map corresponds to the time slice of 2000 and 2001. The most prominent words are *anthrax* in 178 articles and *bioterrorism* in 118 articles. The highly cited articles in this period include Inglesby's 1999 article (14), Torok's 1997 article (14 citations within the period), and Franz's 1997 article (13). From the bibliography of these papers listed below, one should be able to understand instantly why an anthrax paper in terrorism research would cite Inglesby's 1997 paper.

Table 3. Citation bursts (shaded years).

	Citation Detail Range: 1990-2003	Citation Burst Span				
r		99	00	01	02	03
25.4	SCHUSTER MA, 2001				14	24
27.6	FRANZ DR, 1997.	7	4	9	13	2
32.2	INGLESBY TV, 1999.	2	3	11	11	3
33.3	TOROK TJ, 1997.	3	4	10	9	3
35.8	HENDERSON DA, 1999.	2	3	10	7	5
42.0	NORTH CS, 1999.		1	1	14	7
69.0	INGLESBY TV, 2000.			3	9	2

- *Inglesby TV, et al. (1999) Anthrax as a biological weapon: medical and public health management. Working Group on Civilian Biodefense. JAMA, 281:1735-45.*
- *Torok TJ, et al. (1997) A large community outbreak of salmonellosis caused by intentional contamination of restaurant salad bars. JAMA, 278:389-395.*
- *Franz, DR, et al. (1997) Chemical recognition and management of patients exposed to biological warfare agents. JAMA, 278:399-411.*

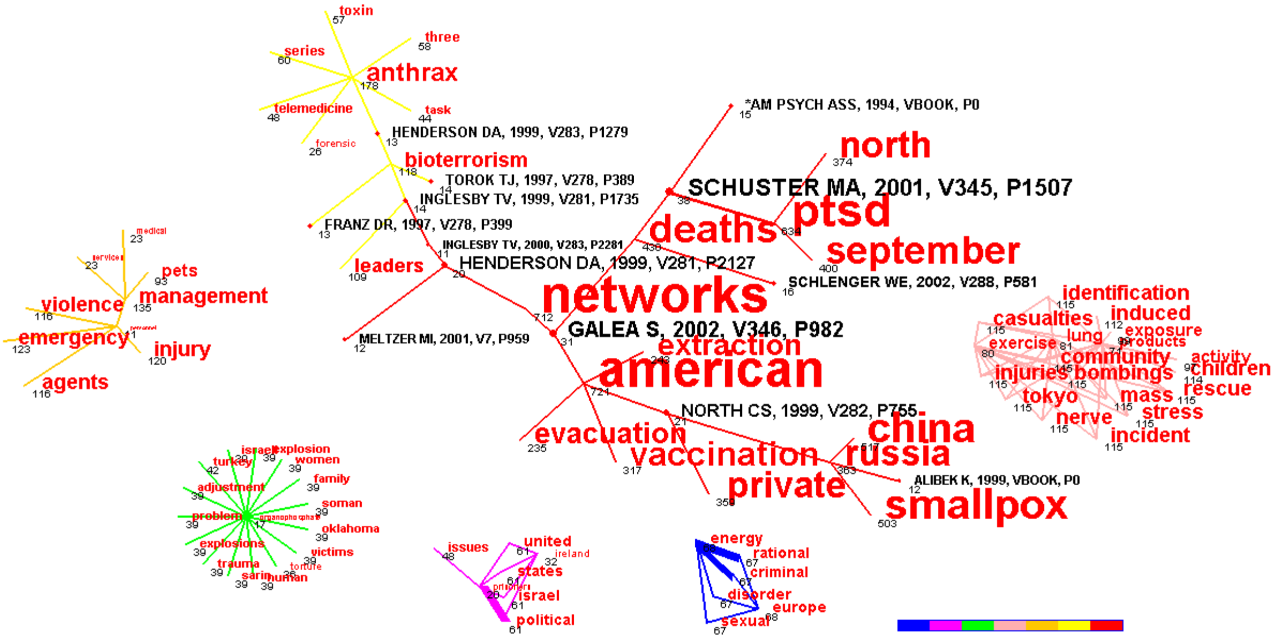

Figure 5 A Pathfinder-pruned version of the network shown in Figure 4.

The red cluster corresponds to the latest time slice (2002-2003). Because words are also shown in red, the words are barely readable. The second map was pruned by Pathfinder algorithm. The significant messages from the red cluster are clear. The cluster is dominated by a large number of burst words, such as *American* (721), *networks* (712), *PTSD* (634), *September* (400), *China* (517), and *smallpox* (503). These words reflect the impact of events such as the September 11 terrorist attacks on issues that scientists were most concerned with.

PTSD stands for Post-Traumatic Stress Disorder. The National Center of PTSD was created within the Department of Veterans Affairs in 1989, to address the needs of veterans with military-related PTSD. The National Center's website highlights several events that may cause PTSD, including the War in Iraq and the September 11 anniversary.

Articles mostly cited in this context include Schuster 2001 (38), Galea 2002 (31), North 1999 (21), and Schlenger 2002 (16). Their titles are listed below.

- *Schuster MA, et al. (2001) A national survey of stress reactions after the September 11, 2001, terrorist attacks. N Engl J Med 2001, 345:1507–1512.*
- *Galea S, et al. (2002) Psychological sequelae of the September 11 terrorist attacks in New York City. N Engl J Med, 346:982-987.*
- *North CS, et al. (1999) Psychiatric disorders among survivors of the Oklahoma City Bombing. JAMA, 282(8):755-762.*
- *Schlenger WE, et al. (2002) Psychological reactions to terrorist attacks. JAMA, 288:581-588.*

On the one hand, the titles of these articles suggest that these articles were cited in the context of how to handle PTSD. The title words *stress reactions*, *psychological sequelae*, *psychiatric disorders*, and *psychological reactions* all reinforce the connections to PTSD after the September 11 terrorist attacks. On the other hand, one question would be why the words such as *psychological* were absent from the visualization.

Overall the detection and visualization of citing words and cited articles in a coherent framework have produced clearer and more understandable representations of an underlying research front. A link between a word and an article in such maps clearly tells us something about why the article was cited. The new method enables us to present more information to the user than what traditional co-citation analysis could do. In particular, the new method strengthens co-citation analysis in the following ways. First, the citing words provide a meaningful labeling mechanism for identifying the nature of co-citation clusters. Traditionally labeling is a challenge in co-citation analysis because of the lack of words available in references. Citing words are more appropriate to be used as labels than the references of articles. Second, the provision of citing words in the process eases the latency problem inherited by co-citation networks. Analysts no longer have to wait until an article gathers sufficient citations before one can take such articles into account. Our approach establishes an important link between the transient part of a research front and the intellectual base. Third, the connection between word analysis and co-citation analysis provides valuable insights into the potential impact of an emerging trend or a sudden change of research focus. Fourth, the framework is extensible. Additional heterogeneous interrelationships from scientific publications can be integrated, such as authors, authors' academic disciplines, and their geographic locations, with content words and referenced articles.

5. Conclusions

In conclusion, the combination of content word analysis and co-citation analysis within the same analyzing and visualizing framework contributes to the quantitative approach to the study of scientific literature. We expect the combined representations of the transient mapping from the state of the art of a field to the intellectual base will enable scientists and other users to track significant developments of a field in a holistic way. Our method has expanded the scope of the coverage of co-citation networks and content word networks if they are used separately. The method is also extensible to incorporate additional phenomena such as scientific collaboration, social networks, and knowledge diffusion across disciplinary as well as geographic boundaries.

References

[1] D. D. Price, "Networks of scientific papers," *Science*, vol. 149, pp. 510-515, 1965.

[2] E. Garfield, "Citation indexes for science: A new dimension in documentation through association of ideas," *Science*, vol. 122, 1955.

[3] R. R. Braam, H. F. Moed, and A. F. J. van Raan, "Mapping of science by combined co-citation and word analysis, I: Structural aspects and II: Dynamical aspects," *Journal of the American Society for Information Science*, vol. 42, pp. 233-266, 1991.

[4] R. J. W. Tijssen and A. F. J. v. Raan, "Mapping changes in science and technology: Bibliometric co-occurrence analysis of the R&D literature," *Evaluation Review*, vol. 18, pp. 98-115, 1994.

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

[6] H. Small, "Visualizing science by citation mapping," *Journal of the American Society for Information Science*, vol. 50, pp. 799-813, 1999.

[7] L. Leydesdorff, *The Challenges of Scientometrics: The Development, Measurement, and Self-Organization of Scientific Communications*, 2nd ed: Universal Publishers, 2001.

[8] H. G. Small, "A co-citation model of a scientific specialty: A longitudinal study of collagen research," *Scoial Studies of Science*, vol. 7, pp. 139-166, 1977.

[9] C. Chen, *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. London: Springer, 2003.

[10] H. G. Small and B. C. Griffith, "The structure of scientific literatures I: Identifying and graphing specialties," *Science Studies*, vol. 4, pp. 17-40, 1974.

[11] M. Callon, J. Law, and A. Rip, "Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World." London: Macmillan Press, 1986.

[12] S. Roy, D. Gevry, and W. M. Pottenger, "Methodologies for trend detection in textual data mining," presented at Proceedings of the Textmine '02 Workshop at the Second SIAM International Conference on Data Mining, Washington, DC., 2002.

[13] A. Kontostathis, L. Galitsky, W. M. Pottenger, S. Roy, and D. J. Phelps, "A Survey of Emerging Trend Detection in Textual Data Mining," in *A Comprehensive Survey of Text Mining*, M. Berry, Ed.: Springer, 2003.

[14] J. Allan, R. Papka, and V. Lavrenko, "Online new event detection and tracking," presented at ACM SIGIR, 1998.

[15] R. Swan and J. Allan, "Extracting significant time varying features from text," presented at Eighth International

Conference on Information Knowledge Management (CIKM'99), Kansas City, Missouri, 1999.

[16] Y. Yang, T. Pierce, and J. Carbonell, "A study on retrospective and online event detection," presented at the 21st ACM International Conference on Research and Development in Information Retrieval (SIGIR'98), 1998.

[17] J. Kleinberg, "Bursty and hierarchical structure in streams," presented at Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

[18] J. Aizen, D. Huttenlocher, J. Kleinberg, and A. Novak, "Traffic-based feedback on the web," *PNAS*, 2004.

[19] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "ThemeRiver: Visualizing thematic changes in large document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, pp. 9-20, 2002.

[20] E. H. Chi, J. Pitkow, and e. al., "Visualizing the evolution of web ecologies," presented at CHI '98, Los Angeles, 1998.

[21] U. Brandes and S. R. Corman, "Visual unrolling of network evolution and the analysis of dynamic discourse," *Information Visualization*, vol. 2, pp. 40-50, 2003.

[22] C. Erten, P. J. Harding, S. G. Kobourov, K. Wampler, and G. Yee, "Exploring the computing literature using temporal graph visualization," presented at Conference on Visualization and Data Analysis (VDA), 2004.

[23] S. Lerman, G. Xu, and A. Tsatsaroni, "A sociological description of changes in the intellectual field of mathematics education research: Implications for the identities of academics," *Proceedings of the British Society for Research in Learning Mathematics*, vol. 23, pp. 43-48, 2003.

[24] C. Chen, T. Cribbin, R. Macredie, and S. Morar, "Visualizing and tracking the growth of competing paradigms: Two case studies," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 678-689, 2002.

[25] C. Chen, "Searching for intellectual turning points: Progressive knowledge domain visualization," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 101, pp. 5303-5310, 2004. http://www.pnas.org/cgi/reprint/0307513100v1.pdf