The Centrality of Pivotal Points in the Evolution of Scientific Networks

Chaomei Chen
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104-2875 USA
+1 215 895 6627

chaomei.chen@cis.drexel.edu

ABSTRACT

In this paper, we describe the development of CiteSpace as an integrated environment for identifying and tracking thematic trends in scientific literature. The goal is to simplify the process of finding not only highly cited clusters of scientific articles, but also pivotal points and trails that are likely to characterize fundamental transitions of a knowledge domain as a whole. The trails of an advancing research field are captured through a sequence of snapshots of its intellectual structure over time in the form of Pathfinder networks. These networks are subsequently merged with a localized pruning algorithm. Pivotal points in the merged network are algorithmically identified and visualized using the betweenness centrality metric. An example of finding clinical evidence associated with reducing risks of heart diseases is included to illustrate how CiteSpace could be used. The contribution of the work is its integration of various change detection algorithms and interactive visualization capabilities to simply users' tasks.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; H.3.3 [**Information Search and Retrieval**]: Information filtering; H.4.3 [**Communications Applications**]: Information browsers

General Terms

Algorithms, Measurement, Design

Keywords

Information visualization, intellectual turning points, knowledge domain visualization, research fronts, betweenness centrality

1. INTRODUCTION

Scientific literature is made of three types of components in terms of how often and how long they are cited after their publication: classics, transient ones, and forgotten ones [20]. Classic work is constantly cited even years after its original publication. Transient ones peak during a relatively short period of time, and then vanish. Forgotten ones are virtually unnoticed by fellow scientists. It is possible for an article in one category to move to another. For example, something being unnoticed for a decade may suddenly draw a surge of attention. It is important for scientists to know the origin and major intellectual milestones of a field. It is also important to be able to assess the value of a particular contribution as perceived by fellow researchers in the field. For sociologists and philosophers of science, it is valuable to be able to study how a domain's knowledge structure changes over time efficiently. It is crucial for biomedical researchers and clinicians to be fully aware the latest outcomes of clinical trials given their clinical implications.

A research front refers to the state of the art of a domain's knowledge. It is desirable to be able to track the dynamics of a research front as it changes over time. In fact, detecting emerging trends and abrupt changes is a common challenge found in several disciplines, including identifying business cycles in economy, following the impact of scientific breakthroughs, and making sense of the wide-ranging impact of profound events as in topic detection and tracking. Key techniques have been devised to support individual tasks such as first story detection, query expansion for efficient information retrieval, and visually exploring a body of document collections.

In this paper, we describe the design of CiteSpace and its new features for identifying pivotal points in evolving scientific networks. CiteSpace was first reported in [7] as a progressive knowledge domain visualization tool for identifying intellectual turning points in a paradigm shift. The original prototype streamlines tasks such as time slicing, network modeling, pruning, and merging. Turing points corresponding to two superstring revolutions in physics were identified based on visually salient properties of visualized networks of co-cited articles. An enhanced version of CiteSpace was subsequently reported in [4], in which Kleinberg's burst detection algorithm [16] was adapted and extended to visualize bipartite scientific networks of surged terms and highly cited articles. Earlier versions of CiteSpace were also used to visualize the domain structure of information visualization [6]. The latest addition to CiteSpace is the new feature for identifying pivotal points based on Freeman's betweenness centrality (BC) metric [11]. In contrast to the use of BC in eliminating inter-cluster links in recent community-finding studies, we argue that high-BC nodes are good candidates for pivotal points in scientific networks. By combining the computational BC metric and visually salient attributes of nodes

in a time-coded network, the double-reinforced network visualization is expected to improve its clarity and interpretability.

The rest of the paper is organized as follows. We first introduce key concepts and discuss related work. We will then formulate the research question and our conceptual model for unifying intelligent and user interface components of trend detection and interactive visualization. An illustrative example of identifying pivotal points in a network of articles reporting clinical trials concerning heart diseases is included. Finally, the significance of the work and the potential of its future development

2. RELATED WORK

The following review focuses on intelligent and user interface aspects of identifying thematic trends in scientific literature.

2.1 Intelligent Analysis

Topic detection and tracking (TDT) originated in the information retrieval community is a particularly relevant source of inspiration. The goal of TDT is to identify a topic from a stream of text data and track the topic over time. Research in TDT has several more specific strands; for example, the goal of First Story Detection (FSD) is to identify the first appearance of a topic in a text stream. Swan and Jensen [26] constructed timelines with statistical models of word usage in TimeMines.

Research in knowledge discovery and data mining has also led to some promising techniques. A notable example is the burst detection algorithm [16], which can be used to measure how fast a given type of events taking place. Example applications of this algorithm include the detection of surges of words in the union speech and scientific papers. In this paper, we describe an extension of the use of burst detection not only in identifying a research front, but also in identifying the intellectual base that is particularly associated with such surge.

Recently there is a growing interest in identifying communities as characterized by associations found in the community of scientists, the World Wide Web (WWW), and scientific literature [12, 21]. The interest is in part rooted in the revival of social network analysis triggered by the rapid advances recently in complex network analysis. Granovetter's work on the strengths of weak ties underlines some fundamental characteristics of weak ties in social networks [13]. Unlike strong ties with clusters, which tend to carry redundant information, weak ties tend to carry more noteworthy information. Exclusive weak ties between two different clusters are called bridges. Most community-discovering algorithms rely on betweenness centrality, a metric proposed by Freeman [2, 11]. The betweenness centrality of a node in a given graph measures the probability that the node is right on an arbitrary shortest path in the graph.

Nodes of strong betweenness centrality tend to be part of bridges connecting tightly connected sub-graphs. Such tightly connected sub-graphs are regarded as discovered communities. Many algorithms are based on a similar concept of edge betweenness centrality to locate inter-community edges, and identify communities by subsequently remove such edges.

In our approach, we use the betweenness centrality of a node to identify pathways between different thematic clusters. Rather than removing high centrality edges, we draw the user's attention to high-centrality nodes in interactive visualizations because we expect studying these pathways can lead to profound insights into how two themes differ.

The innovative use of betweenness centrality is also a significant improvement to enable users find intellectual turning points more accurately and easily. In the initial implementation of CiteSpace [7], we conceptualized a turning point as a node that exclusively joins two or more sub-graphs. However, users had to resort to visual search to spot nodes with such topological properties as no computational or other intelligent data modeling mechanism was in place to support the task of identifying a pivotal point in an integrated intelligent user interface. The use of betweenness centrality in this work allows us to compare the computational and visual results and see to what extent they are consistent to each other.

2.2 Visualizing Thematic Changes

The earlier work by Henry Small on collagen identified changes of research focus in terms of membership changes in clusters of co-cited scientific papers [25]. Interrelationships between papers can be quantified in a number of ways, including co-citations and bibliographic coupling [15] as well as content similarities modeled by Vector Space Models (VSM) [23] and Latent Semantic Indexing (LSI) [9]. Unifying frameworks have been proposed in the past, for example, Generalized Similarity Analysis (GSA) [5]. Co-citations are increasingly used as an effective grouping mechanism. However, traditional co-citation analysis does not incorporate much intelligent analysis components beyond clustering. When a body of scientific literature is visualized, the support for interactive exploration tends to be limited and often isolated from real-world repositories of scientific papers, with exceptions of few systems [27].

The lack of integration between co-citation analysis and seamless accessibility to scientific literature is the result of a complex set of factors, notably the proprietary nature of the Science Citation Index and the proprietary nature of digital archives of scientific papers. CiteSeer has done an exemplar job in establishing a credible resource in computer science literature, including access to full-text papers, citations, and citation contexts. Identifying communities is still at an early stage in this context, whereas visualizing thematic changes has an even longer way to go.

ThemeRiver [14] is a good example of how thematic changes can be captured by visualizing how term frequencies change over time. Timeline [18] arranges clusters of scientific papers based on bibliographic coupling along several parallel timelines. One can see how papers on a topic are distributed at a glance.

Brandes et al. [3] visualize the evolution of a discourse as a spiral extending through several layers of time planes. Erten et al. [10] visualize categories of papers in the ACM Digital Library across three time slices to show the changes.

Although a wide variety of intelligent user interfaces have been developed, to our knowledge, there is still the lack of integrated systems that facilitate the identification of emerging trends in scientific literature, especially for discovering and understanding fundamental turning points across a time interval that spans years or decades. The lack of seamless integration between data visualization and data access hinders the understanding and interpretation of the visualized models because information available from a single source is often inadequate; it also hinders users to retrieve the most important scientific work effectively and timely if they do not have the tools to help them sift through the vast amount of information.

2.3 Intelligent User Interfaces

Currently, without utilizing artificial intelligence techniques, such as machine learning or user profiling, CiteSpace is not a typical intelligent user interface in its strict sense. In a broader sense, it can fit in the framework of intelligent user interfaces because it provides intelligent modeling and discovery functions of an underlying domain so as to improve the efficiency of human-machine interaction [17].

Two studies in the intelligence user interfaces community are particularly relevant: Rosetta [1] and Conversation Map [22]. Rosetta indexes computer science articles based on words and phrases used by their citers in order to improve the match between a search query and indexed documents. It is a widely recognized problem that users often have little idea of how to make use of query language features in order to formulate their information needs effectively. One of the advantages of citation-based approaches, as used in CiteSpace, over traditional query-based approaches, is that search by citation links capitalizes on the knowledge of scientists in the domain, which is similar to ask a domain expert to tell us which two articles are related. The Conversation Map is built on the basis of text analysis and social networks of newsgroups. It combines social and semantic networks of newsgroups for users to explore.

The following section introduces the conceptual model of CiteSpace, which outlines an integration of intelligent trend detection and visualization techniques towards a user friendly environment in which one can learn about a subject, its major milestones in the past, and key changes involving its major components.

3. CONCEPTUAL MODEL

According to Norman [19], the hallmark of a good user interface design is that it is self-explanatory and requires the minimal amount of user manuals to explain how it is supposed to be used. A good conceptual model should give users a clear idea of how a system will behave and let users feel that they are in control.

3.1 A Stream of Snapshots

Our conceptual model is simple. Suppose we want to study the evolution of a subject over a given period of time. Imagine we have a special camera and you take a sequence of snapshots of the subject in the backdrop of scientific literature. Since we are interested in trends and changes over time, we can join the sequence of snapshots as if we are creating a panorama from several photographs. Given two adjacent snapshots, the most likely scenario is that something on both snapshots overlaps. The most natural way to join the two snapshots is to align up the overlap parts. If each snapshot is a network, the joint network is a panorama of the entire time span.

How does this panorama help us to detect trends and turning points? Let's use citation networks as an example, but the principle can be generalized to other phenomena. Suppose we have two snapshots of a network. We call one *before*, and the other *after*. Suppose the *before* network is dominated by a paper *x*, but there is a change and no one cites the paper any more in the *after* network; instead, the after network is dominated by another paper *y*. Assume there is yet another paper *z*; and *z* remains citable in both periods. Here is what we would expect to see - the before network is a cluster centered around *x*, the after network centered around *y*, and the two networks are connected via *z* single handed. If there is no such change, there will be a single cluster in the joint network, i.e. showing a trend. If there is a change like this, z would have a high betweenness centrality and it would be a pivotal point, or a turning point for that matter.

What types of user tasks does our conceptual model invite? We would want to know who is the *z*, and more about it if possible. We might also want to know the nature of either the *before* cluster or the *after* cluster. We might even want to compare where they differ. Click through tens or hundreds of nodes would be far too time consuming and burdensome to remember what we just read. A more desirable way to do it would be casting a net to select everything in a cluster and obtain an intelligent summary. If the data at hand is incomplete, we may like to see if the missing data can be filled by searching elsewhere.

Figure 1 shows schematic diagrams of key components explained in the conceptual model and how underlying operations are matched to the design of visualization. More details are given in the following section.
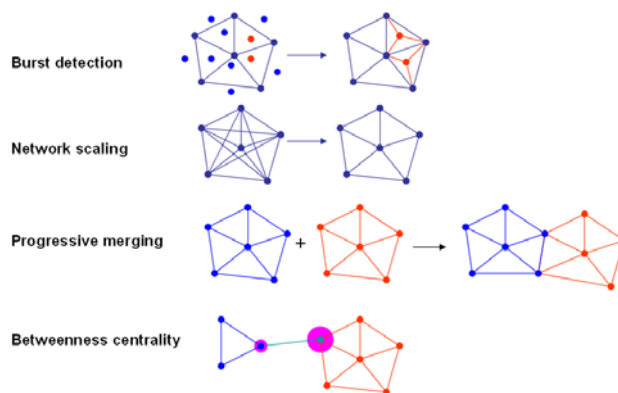

Figure 1. Key intelligent and visualization components in CiteSpace.

3.2 CiteSpace

Time slicing, pruning, and progressive merging options are implemented in the initial version of CiteSpace – a Java application reported in [7]. The use of burst detection on single words is included in [4]. In this paper, burst detection is used to detect terms, including words and phrases of up to 4 words. The use of betweenness centrality is a significant new addition. The group select function and instant search on third-party resources are also new improvements.

CiteSpace uses an algorithm called Pathfinder network scaling [24] to prune excessive links in a network in order to improve the clarity of the subsequent visualization. The Pathfinder algorithm has been used to extract the most salient structures from proximity networks. Let $G = (V, E)$ denote a network, i.e. a graph with weighted edges, where V is the set of vertices and E is the set of weighted edges, or links. Pathfinder leaves V untouched, but reduces the number of links in E based on a *triangle inequality*.

Given a metric space, a triangle inequality can be defined as follows,

$$w_{ij} \le (\textstyle\sum_k w^r n_k n_{k+1})^{1/r}$$

where w_{ij} is the weight of a direct path between *i* and *j*, $wn_k n_{k+1}$ is the weight of a path between n_k and n_{k+1}, for $k = 1, 2, \ldots, m$. In particular, $i = n_1$ and $j = n_k$. In other words, the alternative path between $\underline{i}$ and *j* may go all the way round through nodes $n_1, n_2, \ldots, n_k$ as long as each intermediate links belong to the network.

The pruning power of Pathfinder is determined by two parameters, q and r. Parameter q specifies the range within which a path of length q can be used to compare with a given link e in a triangle inequality check. This is similar to a search for alternative flight routes between two cities comparing with the non-stop flight. The link weight could be the cost, which is why such paths are also called minimum cost paths. If q = 2, we only consider alterative routes of two links, i.e. 1-stop flights; if q = 3, consider both 1- and 2-stop routes against the non-stop one. The rule is the tightest when q = N – 1, where N is the size of V, i.e. en route all possible cities. The r parameter defines a metric space over a given network based on the Minkowski distance so that one can measure the length of a path connecting two nodes in the network. The Minkowski distance becomes the familiar Euclidean distance when $r = 2$. When $r = \infty$, the weight of a path is defined as the maximum weight of its component links, and the distance is known as the maximum value distance.

The strongest pruning configuration is to set q = N – 1, and r = ∞. It has been shown that Pathfinder networks tend to capture more meaningful patterns in network evolution than minimum spanning trees [8].

Until the work of CiteSpace [7], Pathfinder had been only used to prune single networks. The merge of two Pathfinder networks is not defined. In CiteSpace, we extend the Pathfinder network scaling from a single network to a time series of networks. Merging two Pathfinder networks $Gi = (V_i, E_i)$ and $Gj = (V_j, E_j)$ joins the two vertex sets, $V_i \cup V_j$, and divides links into three sets $E_i - E_j$, $E_j - E_i$, and $E_i \cap E_j$ and processes them as follows. If links in $E_i \cap E_j$ violate additional triangle inequality checks, they will be removed from the merged network; otherwise, they will become part of the merged network. The merge operation also makes it possible to apply a divide-and-conquer approach to process large networks.

As mentioned earlier, the initial version of CiteSpace relies on users' visual search to identify pivotal points. Users could experience a considerable cognitive load if the visual search involves a large number of nodes and links. The integration of betweenness centrality and visualization is expected to simplify the task. New features in CiteSpace are explained in next section.

4. SUPPORTING SENSE MAKING TASKS

The new additions to CiteSpace include the use of betweenness centrality to identify pivotal points and marquee selection to search for additional information on the Internet.

4.1 Pivotal Points

The bewteenness centrality is computed for each node in the merged network. The implementation is based on a fast algorithm by Brandes [2].

The betweenness centrality of a node v is defined as follows as $g(v)$

$$g(v) = \sum_{s \neq v \neq t} \sigma_{st}(v)/\sigma_{st}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of shortest paths from s to t going through v.

Nodes with high centrality are marked with purple rings in the visualization for easy recognition. The user can double click on each node and bring up additional information. Currently, a double click invokes a search on PubMed, a service of the National Library of Medicine, includes over 15 million citations

for biomedical articles back to the 1950's. These citations are from MEDLINE and additional life science journals. PubMed includes links to many sites providing full text articles and other related resources.

4.2 Hybrid Networks

In traditional studies of research fronts, a research front and its intellectual base may not necessarily present in the same representation. More often they don't. Algorithms that can detect abrupt changes, such as burst detection, can be used to identify a set of fast-rising terms used by scientists in their latest publications. The current research front can be characterized by such sets of terms. In addition, because these new publications make references to earlier work that has influenced their work, the collection of references cited by the research front papers define the intellectual base. In such a way, we can extend homogeneous networks of co-cited papers to heterogeneous networks of multiple types of entities and multiple types of relations. For example, legitimate entity types may include author, paper, index term, grant, funding program, inventor, patent, and many more.

In this paper, we limit to two types of entities: fast-rising terms identified by burst detection and highly cited papers selected based on user specified thresholds from within CiteSpace.

4.3 Graphical User Interface

The main user interface for CiteSpace is shown in Figure 2. The user can perform various data import and export functions as well as configure the entire streamlined modeling and visualization process.
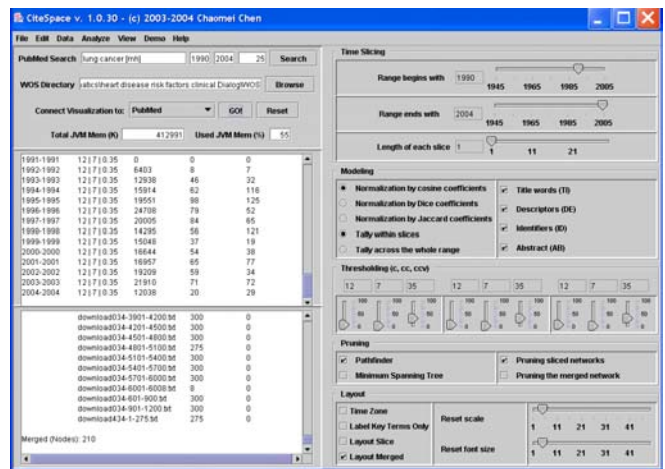
Figure 2. Graphical user interface of CiteSpace.

The primary use scenario is that the user searches a bibliographic database such as the Web of Science, DialogWeb, or other resources. The user can search for a topic with a few search terms, or simple specify all articles published in one or several journals or conference proceedings. The user can save the search results in files stored in a user directory. Start CiteSpace and identify the data directory so that CiteSpace can import the data.

To illustrate the process, let's image the user's goal is to explore the literature of clinical trials of risk factors related to heart diseases. Our user would like to find out about the following information:

- *What are the most influential clinical trials over the last decade?*

- *What are the most intensively studied risk factors of heart diseases?*

- *What are the most remarkable changes concerning clinical evidence related to heart diseases over the last 5~10 years?*

- *What is the latest trend in this area?*

The dataset of citations was compiled based on a search for `((heart disease) and (risk factor*) and clinic*)` in SciSearch® 1974-1989 and 1990-2004 in DialogWeb. 275 and 6008 records were found in 1974-1989 and 1990-2004, respectively. In this paper, we limit the analysis to 1992-2004 because there are not many papers before this period.

In the modeling panel of the interface, select the widely used cosine coefficients and tick all subject terms to build a hybrid

network over each time slices. The threshold selection is done with three points – the earliest slice, the middle, and the latest slice. Three thresholds can be selected for each point, or simply leave them to default values. The three thresholds are citation threshold *c*, co-citation threshold *cc*, and co-citation coefficients threshold *ccv*. The higher a threshold, the fewer entities may enter the subsequent analysis. To allow more entities into the analysis, lower the thresholds. Note that many papers in the dataset may not be featured because they do not have enough citations.

The default pruning algorithm in CiteSpace is Pathfinder. The user can decide whether to prune networks in individual slices or prune the merged network as well.

Figure 3. The cluster view in CiteSpace, showing a hybrid network of surged citing terms and highly cited references. Labels of cited references are turned off in this image. Annual citations to each article are rendered as citation tree rings with the earliest citations corresponding to the inner most rings. Nodes with outmost rings in purple are identified as pivotal points. The color of a link denotes the earliest time slice in which the connection was first made. Texts in black are the most representative MeSH headings assigned by National Library of Medicine to articles within each rectangle.

4.4 Interacting with CiteSpace Visualization

CiteSpace support two types of visualizations. One is the cluster view, in which a modified spring embedder algorithm is used to generate the spatial layout of the visualization (see Figure 3). The other is the timezone view, in which a constraint spring embedder is used so that entities are locked in vertical time zones corresponding to their publication time (see Figure 4). Entities can move freely vertically within their time zone.

The visualization interface includes a visualization panel, a browser panel containing selected items as hyperlinks to PubMed, and three tables listing the details of selected nodes, major MeSH headings and other MeSH headings found in PubMed. The color bar on the top of the visualization is the color mapping legend for links in the map. A link in blue would be made in the first time slice, whereas a link in red would be made in the most recent time slice. Links in a cluster tend to be similar in color – this is the footprint of a research front.
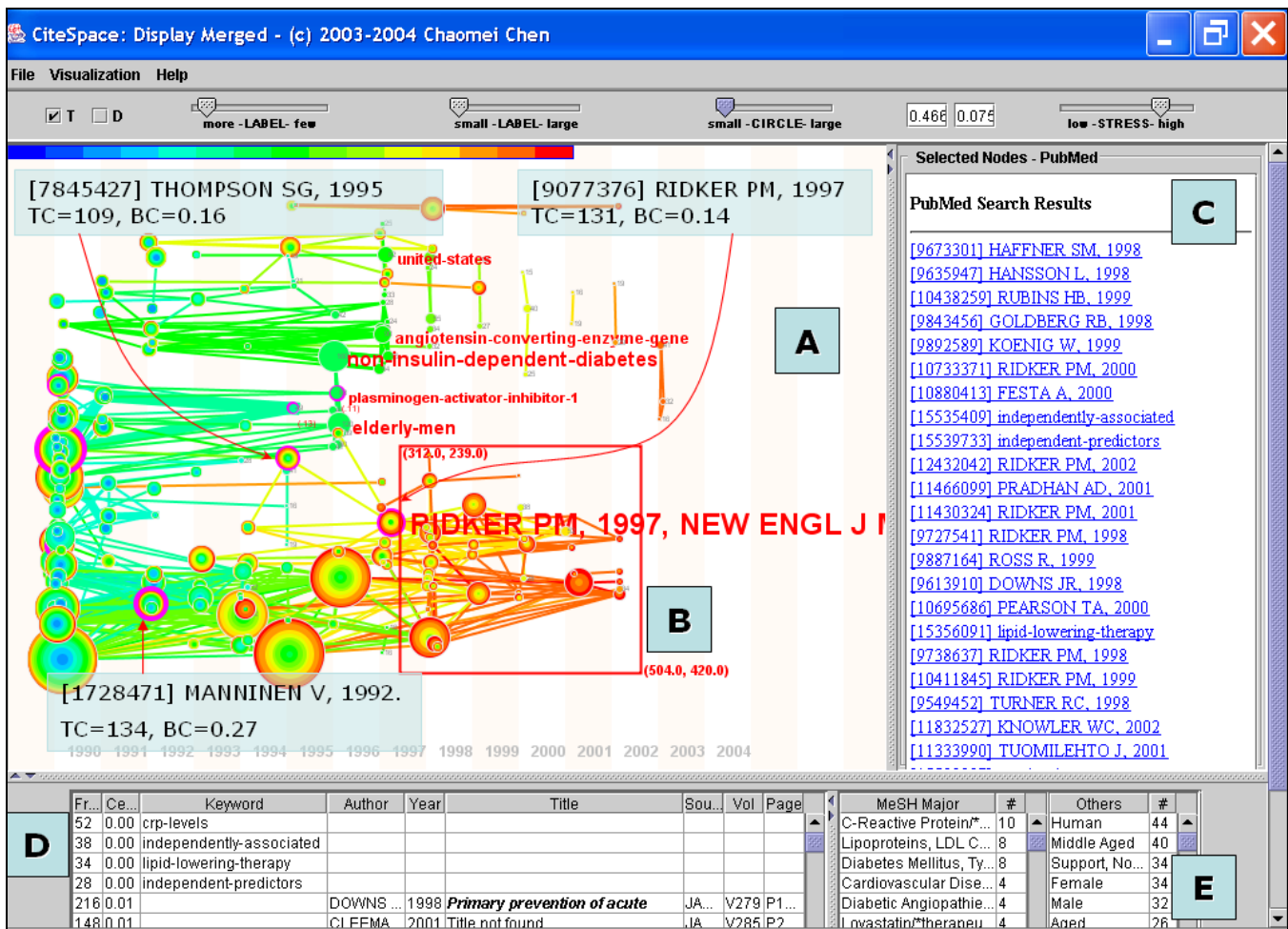
Figure 4. The timezone view. Surged terms and highly cited articles are arranged by their publication date. Items prior to 1990 are collapsed into a single timezone.

Users can interact with the visualized network in several ways. Double click on a node will invoke a search on PubMed. If there is a match and the node is a paper, a hyperlink to the abstract of the selected node will appear in the browser panel (region C in Figure 4). Click on the hyperlink and the abstract will appear in the second half of the browser. Science Citation Index (SCI), the product of the Institute for Scientific Information (ISI), does not include the title of a cited work in the cited reference (CR) field. The standard format contains five fields: author (AU), year of publication (PY), source (SO), volume (VO), and page (PG). In contrast, a PubMed record normally contains several MeSH terms. These MeSH terms are assigned to a record by human indexers at the National Library of Medicine to categorize the topic of the paper. The MeSH terms are selected from a controlled vocabulary. They are valuable for the user to identify the major topic of a paper. MeSH terms are not available from ISI's citation data. CiteSpace extracts such MeSH terms and displays them in the tables ranked by their frequencies when multiple nodes are selected in the visualized network.

Region D in Figure 4 shows a table of selected nodes, sorted by the first column, the citation frequency of a paper, or the frequency of a term. If the list contains both terms and articles, then terms will be shown first, followed by articles. In such lists, terms represent a *research front*, whereas articles form the *intellectual base*. The second column shows the node

betweenness centrality. For most rows, this column is 0.00. A non-zero centrality means the node may be a pivotal point. The title column displays the title of a paper if a match is found on PubMed; otherwise, "Title not found" is shown.

The solid rectangle B in Figure 4 marks the scope of a so-called *marquee selection*. It allows the user to select a group of nodes and send for PubMed search. To make a marquee selection in CiteSpace, hold the ALT key and drag the mouse so that the frame covers the nodes you want to include in the PubMed search.

5. RESULTS

CiteSpace is designed to help users who seek answers to the following questions. Given a knowledge domain, what are the leading research questions? How many distinct research fronts and their intellectual bases can be identified in clusters? What is the nature of each such cluster? What are the most popular MeSH terms assigned to papers in a given cluster? How do two clusters differ, in terms of their subject terms, and/or MeSH terms?

5.1 Pivotal Points

The significance of a pivotal point can be understood by examining the MeSH terms assigned to the papers in corresponding clusters. Figure 4 shows the timezone view of the visualization of the merged network. Each vertical strip represents

103

a year. Papers published in a particular year are located in the corresponding strip – a time zone. The interrelationship between a research front and its intellectual base is easier to recognize in the timezone view than in the cluster view. A research front tends to appear as a cluster of terms in red with links to its intellectual base on its left. The width of the stretch between a research front and its intellectual base indicates how far back the research front sets its foundation on.

The following discussion focuses on the three pivotal points annotated in Figure 4. The region B would capture the latest trend in clinical trials of heart disease risk factors because the majority of links are in yellow, brown, or red, suggesting they were made in the most recent three time slices. MeSH Major headings shown in region E indicate that the nature of this cluster can be best captured by the MeSH heading "*C-Reactive Protein/*analysis.*" The pivotal point immediately before this cluster is an article published in 1997 by P. M. Ridker. It has been cited 131 times with a normalized betweenness centrality of 0.14. This pivotal-point article provides clinical evidence that C-reactive protein predicts the risk of future myocardial infarction and stroke, and that the use of aspirin appears to reduce the risk. Although a thorough evaluation is in order, given the theme of the subsequent cluster of papers citing this pivotal-point article, the identification of this clinical evidence is promising because we would probably miss this one if we rank the importance of nodes by citation counts, i.e. the degree centrality. The conclusion section of the abstract states:

The base-line plasma concentration of C-reactive protein predicts the risk of future myocardial infarction and stroke. Moreover, the reduction associated with the use of aspirin in the risk of a first myocardial infarction appears to be directly related to the level of C-reactive protein, raising the possibility that antiinflammatory agents may have clinical benefits in preventing cardiovascular disease.

An even earlier pivotal point is a 1995 article by S. G. Thompson, with a total of 109 citations and a betweenness centrality of 0.16. This article studied the risk of myocardial infarction, which was also studied by the 1997 pivotal point. The conclusion of the Thompson study is as follows:

In patients with angina pectoris, the levels of fibrinogen, von Willebrand factor antigen, and t-PA antigen are independent predictors of subsequent acute coronary syndromes. In addition, low fibrinogen concentrations characterize patients at low risk for coronary events despite increased serum cholesterol levels.

If we keep tracing back, the next pivotal point is a 1992 article by V. Manninen. It was cited 134 times. Its betweenness centrality is 0.27. This article reported the prognostic value of serum triglyceride concentration. Its conclusion is as follows:

Serum triglyceride concentration has prognostic value, both for assessing coronary heart disease risk and in predicting the effect of gemfibrozil treatment, especially when used in combination with HDL-C and LDL-C.

A closer examination of the visualization also reveals some potentially significant changes. MeSH heading Lovastatin/*therapeutic was found in earlier papers, whereas Pravastatin/*therapeutic was found in more recent papers. According to MedlinePlus, both *lovastatin* and *pravastatin* are drugs used to reduce the amount of cholesterol in blood. High cholesterol is one of the risk factors for heart diseases. The following evidence was extracted from two of the highly cited but not pivotal points themselves:

Shepherd (1995) reported on *pravastatin*:

Treatment with pravastatin significantly reduced the incidence of myocardial infarction and death from cardiovascular causes without adversely affecting the risk of death from noncardiovascular causes in men with moderate hypercholesterolemia and no history of myocardial infarction.

Brown (1990) concluded on *lovastatin*:

In men with coronary artery disease who were at high risk for cardiovascular events, intensive lipid-lowering therapy reduced the frequency of progression of coronary lesions, increased the frequency of regression, and reduced the incidence of cardiovascular events.

This type of clinical evidence should be very useful both to clinicians and to patients. We are planning the next step to have clinicians involved in evaluative studies of medical decision support using CiteSpace.

6. FUTURE CHALLENGES

The major contribution of this work is the integration of a number of intelligent modeling and search components at the back end and the interactive user interface showing various aspects of a visualized network at the front. This is also an extensible architecture that can incorporate new intelligent data analysis components as well as other visualization and views. Searching PubMed is a real-world test. Incorporating search in other types of resources is a promising route to pursue. The work also provides a potentially useful environment for seamless data fusion.

Connecting to full-text archives and comprehensive digital libraries is necessary to undertake deeper text analysis and incorporate data mining and natural language processing to a greater extent. It will lead to the development of deeper intelligent agents.

Validating the results with domain experts is an integral part of the development process. We intend to involve domain experts in future studies. Additional use scenarios will be developed, especially including medical decision making and healthcare informatics. Evaluative studies in comparison with complex queries used to retrieve clinical evidence in evidence-based medicine will be an important step towards an intelligent user interface that can transform our ability to detect and track the latest evidence in our knowledge.

7. NOTES

The current version of CiteSpace and several interactive examples are available at: http://cluster.cis.drexel.edu/~cchen

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] Bradshaw, S., Scheinkman, A. and Hammond, K. Guiding people to information: Providing an interface to a digital library using reference as a basis for indexing. In

Proceedings of Intelligenet User Interfaces (IUI 2000) (New Orleans, LA, 2000), ACM Press, 2000, 37-43.

[2] Brandes, U. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, *25*, 2 (2001), 163-177.

[3] Brandes, U. and Corman, S. R. Visual unrolling of network evolution and the analysis of dynamic discourse. *Information Visualization*, *2*, 1 (2003), 40-50.

[4] Chen, C. Detecting and mapping thematic changes in transient networks. In *Proceedings of the 8th International Conference on Information Visualization (IV'04)* (London, 2004), IEEE Computer Society Press, 2004, 1023-1032.

[5] Chen, C. Generalised Similarity Analysis and Pathfinder Network Scaling. *Interacting with Computers*, *10*, 2 (1998), 107-128.

[6] Chen, C. *Information Visualization: Beyond the Horizon*. Springer, London, 2004.

[7] Chen, C. Searching for intellectual turning points: Progressive Knowledge Domain Visualization. *Proc. Natl. Acad. Sci. USA*, *101* (2004), 5303-5310.

[8] Chen, C. and Morris, S. Visualizing evolving networks: Minimum spanning trees versus Pathfinder networks. In *IEEE Symposium on Information Visualization* (Seattle, Washington, 2003), IEEE Computer Society Press, 2003, 67-74.

[9] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41*, 6 (1990), 391-407.

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

[11] Freeman, L. C. A set of measuring centrality based on betweenness. *Sociometry*, *40* (1977), 35-41.

[12] Girvan, M. and Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, *99* (2002), 7821-7826.

[13] Granovetter, M. Strength of weak ties. *American Journal of Sociology*, *8* (1973), 1360-1380.

[14] Havre, S., Hetzler, E., Whitney, P. and Nowell, L. ThemeRiver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, *8*, 1 (2002), 9-20.

[15] Kessler, M. M. Bibliographic coupling between scientific papers. *American Documentation*, *14* (1963), 10-25.

[16] Kleinberg, J. Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Edmonton, Alberta, Canada, 2002), ACM Press, 2002, 91-101.

[17] Maybury, M. T. and Wahlster, W. *Readings in Intelligent User Interfaces*. Morgan Kaufmann Publishers, San Francisco, CA, 1998.

[18] Morris, S. A., Yen, G., Wu, Z. and Asnake, B. Timeline visualization of research fronts. *J. Am. Soc. Inf. Sci. Technol.*, *55*, 5 (2003), 413-422.

[19] Norman, D. A. *The Invisible Computer: Why Good Products Can Fail, The Personal Computer is so Complex, and Information Appliances are the Solution*. MIT Press, Cambridge, MA., 1998.

[20] Price, D. D. Networks of scientific papers. *Science*, *149* (1965), 510-515.

[21] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. and Parisi, D. Defining and identifying communities in networks *arXiv: cond- mat/ 0309488 v1*, 2003.

[22] Sack, W. Conversation Map: A content-based usenet newsgroup browser. In *Proceedings of Intelligent User Interfaces (IUI 2000)* (New Orleans, LA, 2000), ACM Press, 2000, 233-240.

[23] Salton, G., Allan, J. and Buckley, C. Automatic structuring and retrieval of large text files. *Communications of the ACM*, *37*, 2 (1994), 97-108.

[24] Schvaneveldt, R. W. (ed.), *Pathfinder Associative Networks: Studies in Knowledge Organization*. Ablex Publishing Corporations, Norwood, New Jersey, 1990.

[25] Small, H. G. A co-citation model of a scientific specialty: A longitudinal study of collagen research. *Soc. Stud. Sci.*, *7* (1977), 139-166.

[26] Swan, R. and Jensen, D. TimeMines: Constructing timelines with statistical models of word usage. In *the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2000), 2000.

[27] White, H. D., Lin, X., Buzydlowski, J. W. and Chen, C. User-controlled mapping of significant literatures. *Proc. Natl. Acad. Sci. USA*, *101* (2004), 5297-5302.