

Trailblazing through a Knowledge Space of Science: Forward Citation Expansion in CiteSeer

Chaomei Chen, Xia Lin, Weizhong Zhu
College of Information Science and Technology
Drexel University
3141 Chestnut Street, Philadelphia PA 19104-2875, USA
chaomei.chen@ischool.drexel.edu, xia.lin@ischool.drexel.edu, wz32@drexel.edu

Understanding emerging trends and patterns in science and technology is essential not only to scientists and engineers in their own fast-advancing fields but also to a wide variety of individuals and organizations who are also interested in tracking the development of thematic topics. This is a challenging task because many existing tools are not particularly designed to deal with the dynamics of intellectual structures that transcend the boundaries of individual documents or isolated topics. In this article, we introduce a conceptual and operational platform that extends the traditional notion of traveling along individual citation pathways and defines operators for recursive and holistic theme expansion based on citation connectivity. We describe the implementation of a forward expansion operator and illustrate its potential with the CiteSeer metadata. In addition, we integrate the forward expansion operators with information visualization techniques.

1. Introduction

Understanding emerging trends and patterns in science and technology is essential and challenging. It is necessary for scientists and engineers to keep abreast of their own fast-advancing fields; it also has profound implications to a wide variety of individuals and organizations who are interested in tracking the development of thematic topics. However, many existing tools do not readily deal with the dynamics of intellectual structures that go beyond the boundaries of individual documents or isolated topics.

Consider this scenario: a researcher is interested in the latest developments in the area of *information visualization*. She may search on Google or other Internet search engines. She may also search bibliographic and citation databases such as CiteSeer¹ and the Web of Science². It is likely that she will find a long list of relevant documents or published articles. If she finds a particularly relevant article on information visualization, she can trace through citation links and see if there are additional potentially relevant articles because of the existence of such links. So far this scenario is a simplified notion of trailblazing in a conceptual space (Bush, 1945). If she could improvise traces left by others in such a space, then she would be able to navigate through the space much more effectively and find the most valuable information to meet her needs.

While citation links in scholarly publications, among other techniques such as text mining and domain analysis, have provided researchers a valuable type of traces, the traditional notion of citation-based navigation mostly operates on the basis of individual links as opposed to a holistic view of emergent patterns from individual links. Our researcher needs answers to a number of questions which are yet to be readily answered with the traditional notion of citation-based navigation. She needs to know not only who has cited a particular article on her list, but also any emergent patterns or similarities associated with citations to this group of articles as a whole. How many schools of opinions are there regarding this group of intellectual works? What

¹ <http://citeseer.ist.psu.edu/>

² <http://scientific.thomson.com/products/wos/>

are the collective impacts of a group of relevant articles? From a longitudinal perspective, what roles have a collection of articles been playing in a development of a subject?

2. Forward Expansion in a Knowledge Space of Science

As the great asset of science and technology, the stability of scientific knowledge varies widely. The core of scientific knowledge is relatively stable, whereas the forefront of scientific knowledge is often fast-paced, volatile, and transient (Chaomei Chen, 2003; Cole, 1992; Price, 1965; Ziman, 1968). Fundamental changes in the conceptual structures of the core knowledge of a scientific field are known as scientific revolutions (Kuhn, 1962), or conceptual revolutions. More frequent changes in the forefront of scientific knowledge can be seen as the movement of a research front. In this context, the quality of scientific knowledge accessible through a network of digital libraries depends on the extent to which these resources reflect the research fronts as well as the structure of the core knowledge. It is essential for users, including learners and teachers, to maintain an awareness of the value of learning and teaching materials in the dynamic structure of scientific knowledge.

Understanding the conceptual structure associated with a scientific knowledge domain is essential to education as well as research. The ability to analyze, synthesize, and evaluate the value of scientific knowledge is not only a critical skill that learners should aim to achieve (Bloom, 1956), but also an important survival skill for researchers to keep up with the rapid advances of their own fields. We use the term knowledge space to emphasize the dynamic nature of such concept spaces in terms of how scientists perceive the intellectual value of knowledge structures. As Vannevar Bush envisaged in his Memex (Bush, 1945), the value of a knowledge structure is how we make various intellectual connections, or trails, and how we may be inspired by such intellectual connections made by others. Although hypertext, notably via the revolution of the World-Wide Web, has made it possible to accomplish a great deal of what Bush envisaged, problems such as lost in cyberspace and cognitive overload have been identified (Conklin, 1987). Users of digital libraries are facing similar challenges (Bollen, Luce, Vemulapalli, & Xu, 2003). Furthermore, users need tools that enable them to keep track the evolution and impact of scientific knowledge over time.

Empirical evidence has consistently shown that graphical representations of abstract information structures have significant effects on users interacting with such structures (C. Chen & Rada, 1996). The educational value of graphical maps has also been noted in learning practice; for example, graphical maps can be used as part of learning methods based on Vygotsky's intellectual scaffolding theory of learning to facilitate learners to master cognitive skills defined in Bloom's learning taxonomy (Donelan, 2005). Earlier work in the digital library community has demonstrated the value of concept spaces automatically generated based on scientific literature (H. Chen, Houston, Sewell, & Schatz, 1998; H. Chen, Ng, Martinez, & Schatz, 1997) and the role of visual representations of categories in such concept spaces, such as the use of multi-layered self-organizing maps (SOM) (H. Chen et al., 1998).

In this article, we focus on how citation links in scientific publications can be utilized for navigation tasks in a knowledge space of a scientific discipline. Existing studies in citation analysis (Garfield, 1955), co-citation analysis (Chaomei Chen, 2006; Chaomei Chen & Paul, 2001; Small, 1973), and author co-citation analysis (Chaomei Chen, 1999; White & McCain, 1998) have made use of citation and co-citation links in identifying missing intellectual links and visualizing intellectual structures. CiteSeer is a widely used repository, or a digital library of scientific publications (Lawrence, Giles, & Bollacker, 1999). In fact, the CiteSeer metadata has

been made available³, including citation links found in more than half million scientific articles. In this article, we use the most comprehensive version of the CiteSeer metadata, namely the Dublin core metadata standard with additional metadata fields, such as citation links (References and IsReferencedBy), author affiliations, and author addresses. We will also use the term citation for reference, and the term citedby for IsReferenceBy. We will use the term *forward* to denote the direction from an article A to an article B if A is cited by B, i.e. $A \leftarrow B$. Implicitly, B is more recent than A because in order for B to cite A it must be published after A. In contrast, if A cites B, then *backward* citation is the direction from A to B, i.e. $A \rightarrow B$. In this article we investigate how one can make use of forward citations in identifying new trends in CiteSeer.

In order to conceptualize the forward expansion as an operator that can be applied recursively to a set of articles, we first introduce the notion of a citation viewpoint (σ), an aggregative mapping from one set of articles \mathbf{S} to another set of articles \mathbf{S}' . $\sigma : \mathbf{S} \rightarrow \mathbf{S}'$. The σ mapping formalizes the interrelationship between a citing article (source) and a cited article (target). Thus, given an article s in \mathbf{S} , $\sigma(s)$ is an article in \mathbf{S}' and $\sigma(s)$ is cited by s . Note that $\sigma(s)$ can be cited by multiple articles in \mathbf{S} , i.e. the image of a forward mapping is a subset of the source set: $\sigma^{-1}(\sigma(s)) \subseteq \mathbf{S}$. By the same token, the forward expansion operator is defined as a set mapping σ^{-1} on the source set: $\sigma^{-1} : \mathbf{S} \rightarrow \mathbf{S}^*$, where \mathbf{S}^* is called a forward expanded set.

There can be a number of ways to construct the set \mathbf{S} , for example, by a topical search, by finding all articles published in a conference series, or by collecting all articles that satisfy a set of user defined criteria. Traditional citation analysis, especially author co-citation analysis and document co-citation analysis in effect focuses on the image, or co-domain of σ , i.e. $\sigma(\mathbf{S})$.

Focusing on $\sigma(\mathbf{S})$ has a number of inherited drawbacks in terms of the timeliness of any patterns and trends one may detect from $\sigma(\mathbf{S})$. It is widely acknowledged that although the expected delay from the time a scientific discovery is made to the time it is published varies from one discipline to another, such delays tend to undermine the overall timeliness of citation analysis if we only focus on what is cited within a relatively narrow scope.

Another practical issue of concern with the conventional σ is that the initial set \mathbf{S} may not be constructed in a way such that it will adequately cover a topic. If one constructs the initial set by a topical search in a database, the breadth and depth of the search results are likely to be affected by the specific design of the database. The user may use a suboptimal search term or set search criteria inadequate. More importantly, if the user is interested in how this initial set of articles have influenced others, then an obvious way for the user is to trace along cited-by links from an article to its citers. Currently, the support of such needs is not readily available. For example, the *Web of Science* (WoS) allows the user to navigate from an article to articles that cite the current article through a link called Times Cited, but, to our knowledge, the user has no way to perform the task multiple articles with a single batch operation. CiteSeer allows a user to follow a link called Context from the main page of an article and view a series of excerpts from articles that cite the current article. These excerpts are called citation contexts, showing how the current article is referenced specifically. Similar to WoS, performing the task on a set of articles instead of one by one individually is also not available from CiteSeer's current user interface.

Forward citation expansion is expected to improve the timeliness and the coverage of σ . The diagram in Figure 1 illustrates the strategy of the citation expansion method. The initial set \mathbf{S}

³ <http://citeseer.ist.psu.edu/oai.html>

consists of three tables, namely citers, authors, and references. The S' set consists of two tables that make up the references with respect to the set S , namely citers and authors. In this framework, σ is no longer an isolated one off mapping from one set to another. Instead, each instance of σ is an element of a citation mapping chain. The fundamental question is not whether it is possible to trace the citers of an individual article; instead, the question is how one can device the mapping mechanism so that the operation can be applied recursively.

The diagram shows a dark gray area (1), a gray area (2), and a light-gray area (3). Three sets of tables have different shade of green: A, B, and C. Traditional citation and co-citation analysis focuses on a mapping from $B \rightarrow A$ from area 2 to area 1. The objective of a forward expansion is to move from $B \rightarrow C$ from area 2 to area 3. Articles in C will form the new citation viewpoint. The new σ, σ' , often has a more recent and timely set of articles as the domain of σ' . The image, or co-domain of σ' can be expected to be more recent and timely than the earlier generation of σ . A slightly formal notation is as follows.

Let $\sigma: S \rightarrow S'$ and $\sigma^{-1}: S' \leftarrow S$. 1-step expansion is defined as: $\sigma^* \sigma^{-1}: (S^* \leftarrow S) \rightarrow S'$, where S^* is the set of articles expanded by cited-by links. In other words, 1-step expansion reverses the citation mapping from the initial set S as: $\sigma: (S^* \leftarrow S) \rightarrow S'$. A more generic k-step expansion can be similarly defined as: $\sigma^* \sigma^{-k}$.

We expect that such forward expansion operators can extend the coverage of a topic and the timeliness of the base articles. The following sections will describe the design of such operations at the user interface level and illustrate how a forward expansion can become a new way of trailblazing the knowledge space of scientific literature.

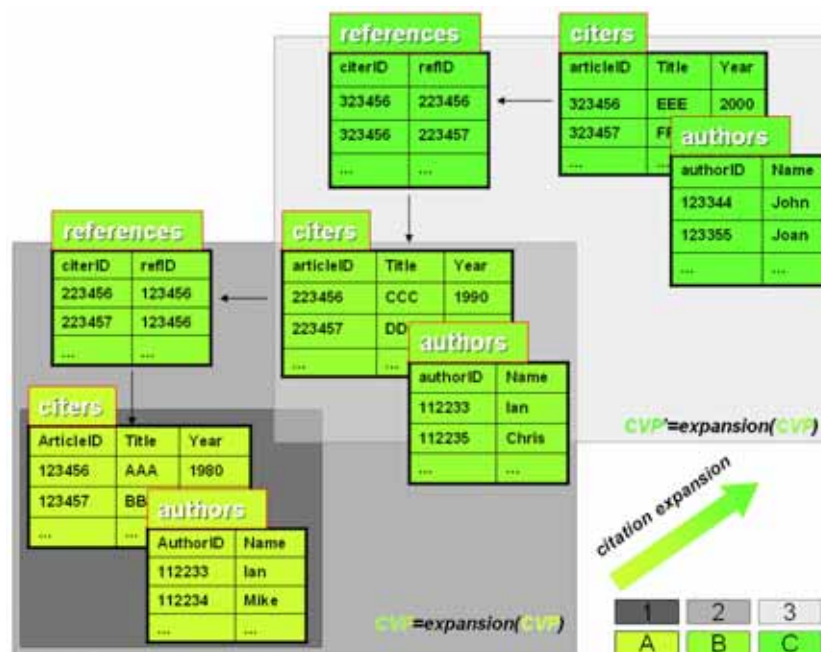


Figure 1. Multiple *Citation Viewpoints* (CVPs) form a chain for citation expansion. The unit of analysis is the role of a group of articles in the development of a topic rather than that of a single article.

3. Forward Expansion in CiteSeer

The CiteSeer metadata consists of more than half million scientific articles written by 1.3 million authors together with 1.2 million references and 4.1 million forward citation reference links, known as cited-by or is-referenced-by in the metadata. See Table 1 below for detail. At the time of writing, CiteSeer hosted at the Pennsylvania State University has 739,135 articles. In other words, there are approximately 165,000 more articles in the current CiteSeer database than in the metadata.

Table 1. Statistics of the CiteSeer metadata used for this study.

Tables	# Records
Articles	574,178
Authors	1,327,197
References	1,271,875
Referenced By	4,113,194

We implement the forward expansion in Java in a new version of CiteSpace. Detailed descriptions of CiteSpace are available elsewhere (Chaomei Chen, 2006). The CiteSeer metadata is parsed and populated into a mysql database. The mysql server is hosted on a Linux cluster. Our Java application access the database over the Internet. Given the volume of the data, the current architecture meets our needs adequately. The choice of Java and mysql is part of our longer-term plan to make the software and the data openly available as a service to serve the scientific community. We are considering alternative architectures for further refinements of the work in terms of extensibility, interoperability, and response speed.

Table 2 shows some examples of forward expansion. Some have led to a significant increase in terms of the number of articles. The overall expansion effect is between 3 and 10 times of the size of the initial set. The difference between the initial set and the expanded set may be useful in identifying fast-growing or high-impact topics, but further studies are needed. We illustrate the methodological details through the *information visualization* example.

Table 2. Some examples of the forward expansion.

Search in title	Initial search	Forward expansion	Expansion effect
Concept drift	20	202	10 times
Digital library or digital libraries	562	3,225	6 times
Information visualization	122	787	6 times
Semantic web	160	494	3 times

The user interface of the prototype is shown in Figure 2. The user interface is designed to support the search and expansion procedure as follows. First, the user types a title phrase or an author's name to search in the CiteSeer metadata. The next section of the user interface will display how many articles that meet the search criteria. The unique ID assigned by CiteSeer, the title, and the year of publication of an article is also displayed to the user. The user now has three options: 1) do a new search, 2) generate a document co-citation visualization based on the initial search, and 3) perform a forward citation expansion and then visualize co-citation networks based on the expanded set of articles.

Since options 1 and 2 are straightforward, we will explain option 3 in more detail. Options and controls for citation-based expansion are grouped in the user interface. The default expansion is forward expansion, which moves the standing point of a citation view closer to the present time than that of a citation view without any expansion. Backward expansion can be also useful, for example, when one needs to trace the origin of a subject matter. In this article, we will focus on the forward expansion rather than a backward expansion.

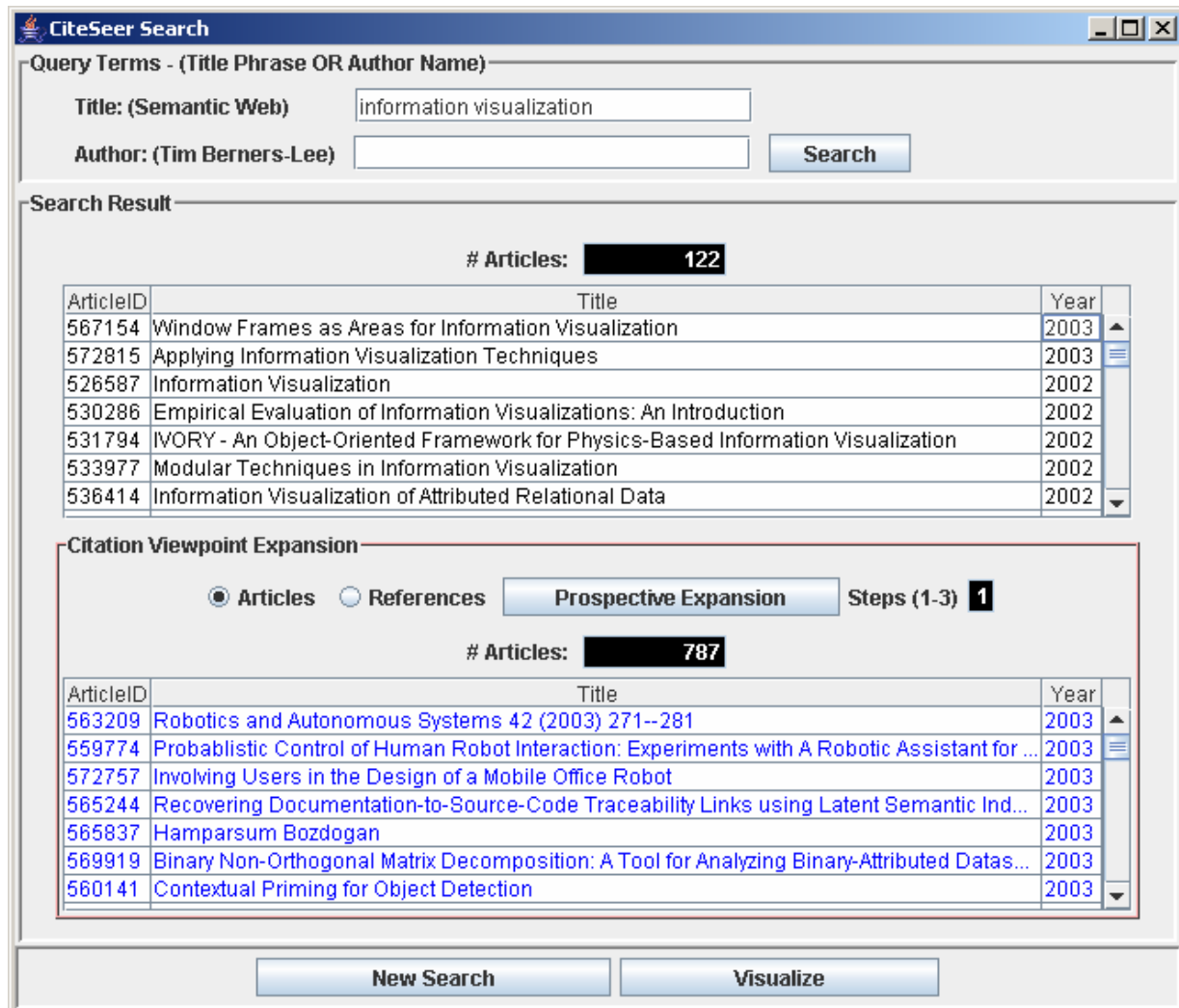


Figure 2. The CiteSeer metadata can be searched and expanded along citation links through this graphical user interface. The initial search for “information visualization” in the title field resulted in 122 articles. A direct forward expansion found 787 articles that cited at least one of the initial 122 articles. Both of the initial 122-article set and the expanded 787-article set are analyzed and compared.

The number of articles returned from the Prospective Expansion, i.e. forward expansion, their titles and the year of publication are also displayed to the user. The user at this stage can choose to do a new search or visualize the expanded topic. If the user chooses to visualize the expanded topic, full records will be retrieved from the normalized database. A separate visualization window will be launched for interactive visualization.

4. Results

The initial search for “information visualization” led to a 122-article set. First, this set of articles was visualized in the form of a document co-citation network. Figure 3 shows a screenshot of the visualized network, containing 121 nodes and 170 links found between 1990 and 2005. The network consists of two major components, one to the left and the other to the right. The left sub-network contains two sections from different years. Shneiderman’s 1998 article *the Eyes Have It* is the most popular in this group. The right sub-network includes a number of noun phrases extracted from the abstracts, such as *user interface*, *minimization algorithm*, *layout procedure*, *visualization process*, and *visualization strategy*. Burst detection in this case found about a few burst terms.

Betweenness centrality has been used to measure the extent to which a node has a central position in terms of connecting various paths in a given network. In the context of co-citation analysis, if an article has a high centrality score, it means the majority of articles in the subject commonly cite the article along with other articles. High-centrality articles in the visualized network underline the most prominent preferences or research issues concerning the initial search results. For example, the 1998 article by Shneiderman entitled “The Eyes Have it” has the highest centrality of 0.18. This is an article that describes a data and task taxonomy for developing tools and understanding users’ needs.

It is known that citation data may not present a comprehensive picture of a subject. On the other hand, knowing the constraints, it is also useful to conceive the CVP image based on our initial search can be used as a reasonable reference framework to infer the impact of these visualized articles on their citers. More importantly, is there anything missing from the image? The current image is our citers’ view. What about the views of more recent articles that cite our citers? The next step is for us to expand our citer set and produce a new image based on the expanded set of articles. What new topics would emerge?

Figure 4 shows a network of 499 articles and terms that are cited by the citers of our citers. In other words, these articles are commonly cited by articles that cite articles in our initial set of articles. First of all, the network appears to be considerably different from the visualization of the initial results. For example, the most frequently found terms include *singular value decomposition* and *expectation maximization*. The term *singular value decomposition* is a key technology in Latent Semantic indexing (LSI). Landauer’s 1995 article shown in the map is related to LSI. The term *expectation maximization* is commonly found in data mining and machine learning literature. Other terms not shown on the map include *search engine*, *information retrieval*, and *dimensionality reduction*. These concepts are known to be related to the topic of information visualization, but they are usually not regarded as the core of information visualization. In other words, it would be unlikely to capture such themes by searching for information visualization. The fact that forward expansion identifies these topics may provide a researcher who is not aware of such connections a useful list of areas to investigate further.

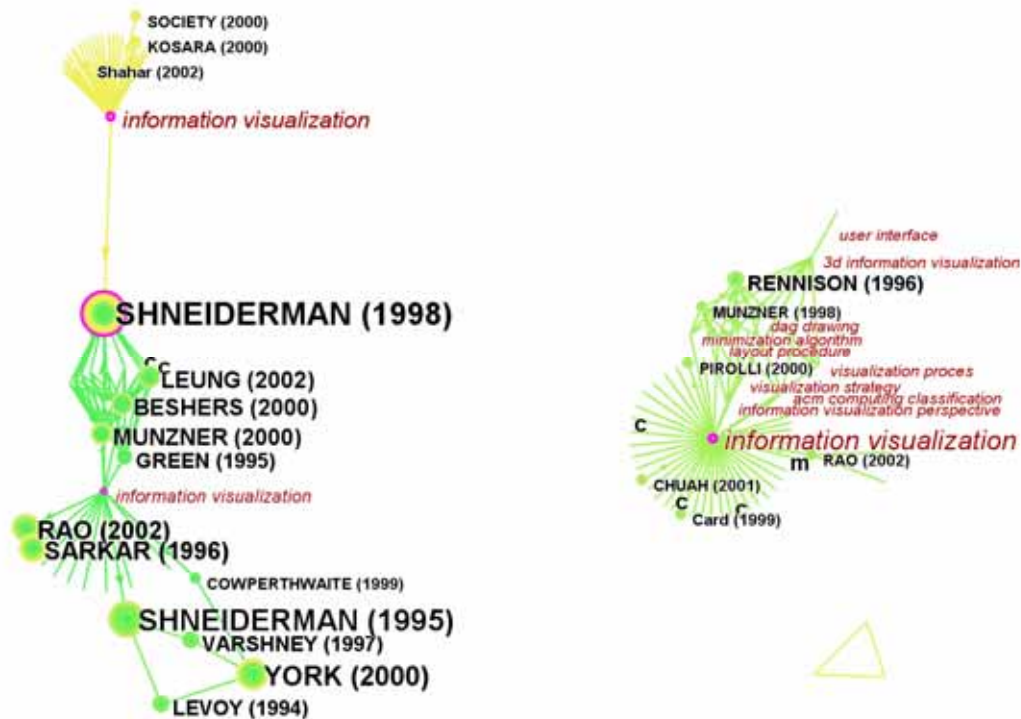


Figure 3. A hybrid network of terms and articles based on the initial 122 articles resulted from the title-field search of "information visualization" in the Citeseer metadata. The network is a Pathfinder network, meaning links meet the triangle inequality condition. (Network: Nodes=121, Links=170)

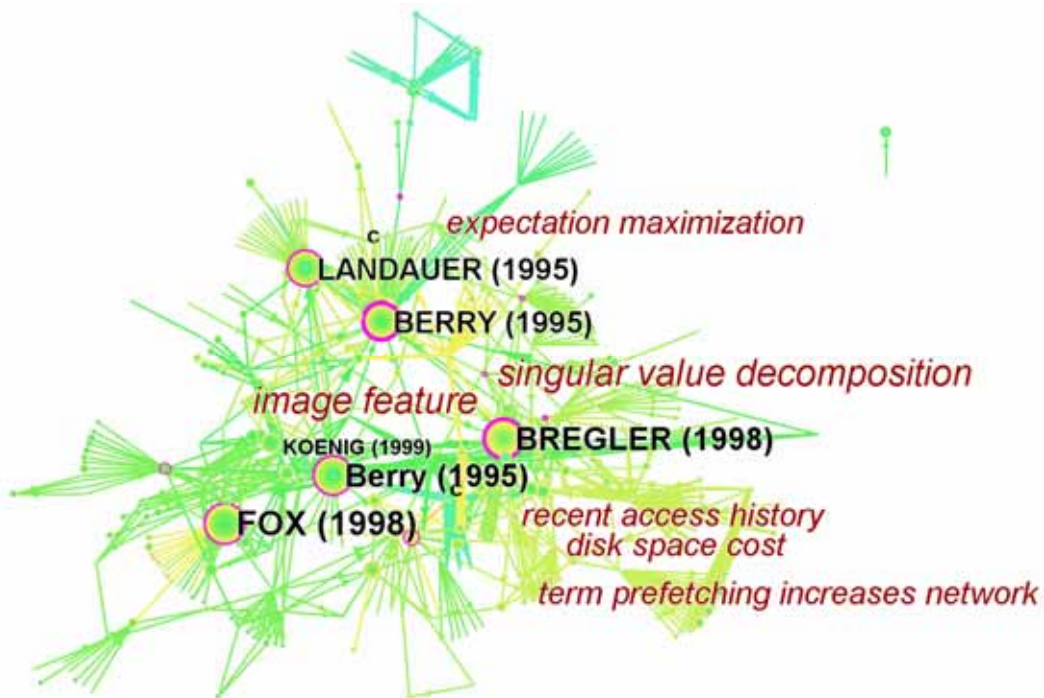


Figure 4. A visualized network of a one-step forward citation expansion of the initial 122-article set on "information visualization." (Network: N=499, E=988).

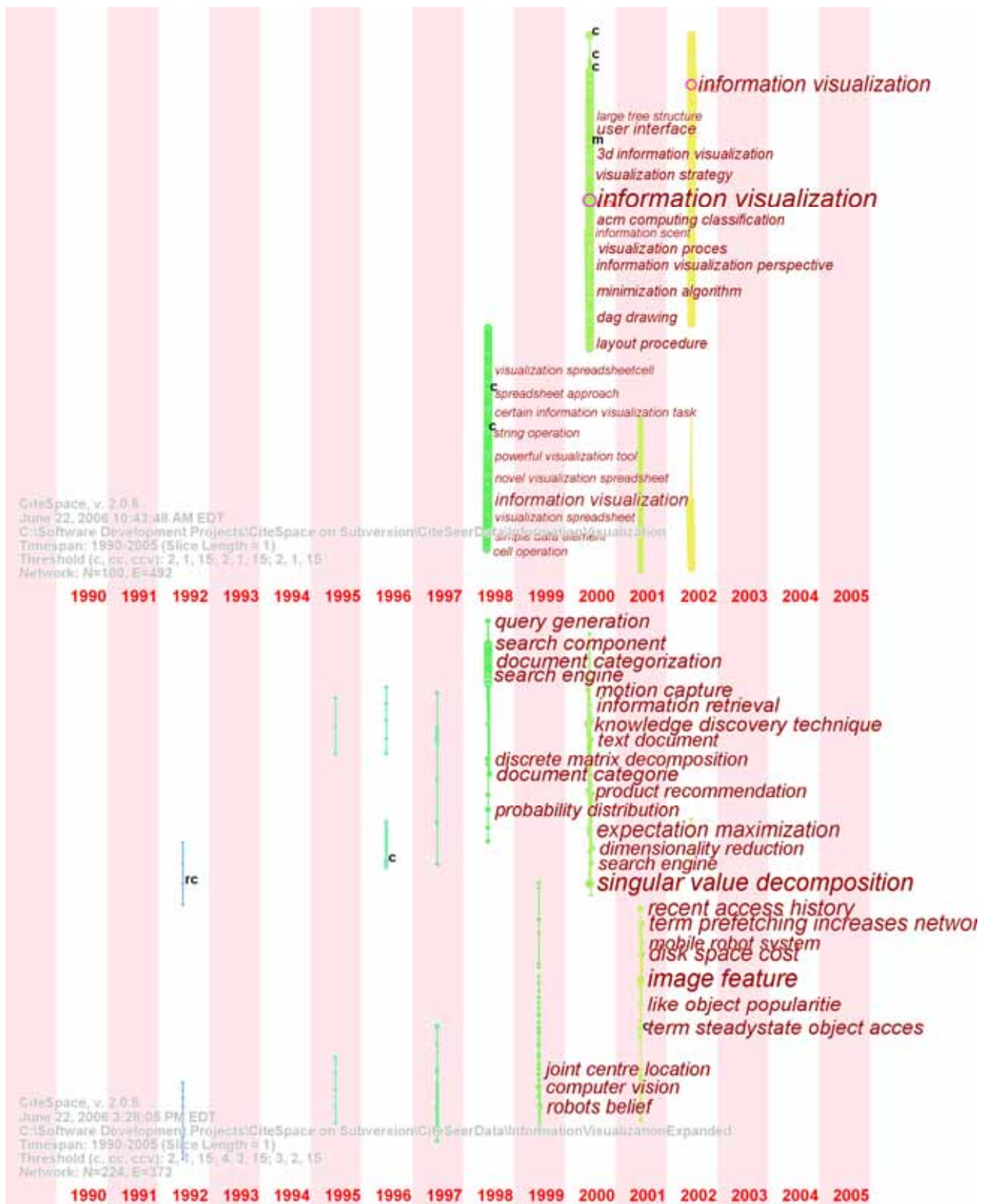


Figure 5. Timelines of key terms extracted from abstracts in both the initial set (top image) and the expanded set (bottom image).

In order to compare the initial and the expanded datasets, we visualize the most frequent terms found in the two sets of articles. These terms are extracted from the abstracts based on part-of-speech tags of noun phrases. Figure 5 shows timezone visualizations generated by CiteSpace. The top one represents the timeline of most frequent terms in the initial dataset. The bottom one represents the timeline of the expanded dataset. As shown in the top image, key terms such as *visualization spreadsheet* and *novel visualization spreadsheet* are found in 1998 and *large tree structure*, *visualization strategy*, and *DAG drawing* are found in 1999. In contrast, the bottom image shows terms such as *query generation*, *search engine*, and *document categorization* in 1998, *computer vision* in 1999, *information retrieval*, *knowledge discovery technique*, *expectation maximization*, and *singular value decomposition* in 2000, and *image feature* and *mobile robot system* in 2001. These terms inform us peripheral aspects of information visualization that have been identified in terms of citations, including Internet search, document categorization, reducing the high dimensionality of document collections, computer vision and robot systems.

5. Discussions

We have gained some valuable information concerning the effectiveness of forward citation expansion in CiteSeer. The most obvious advantage of working with the CiteSeer metadata versus the Web of Science is its open access nature. Because it is openly accessible in its entirety, one can move back and forth along citation links in CiteSeer with an implementation of the forward expansion operator. Such tasks are impossible in the Web of Science, where programmatically access data on-demand is currently not permitted for user applications.

We encountered a number of practical issues in dealing with the CiteSeer metadata. For example, the order of authors is not preserved in the CiteSeer metadata, and it is impossible to restore the order from the metadata. Since CiteSeer relies on automated citation indexing, some references in articles have not been correctly extracted. More importantly, missing or incorrectly parsed references do have adverse impacts on the quality of resultant expansions. The most serious type of problems we have experienced with the CiteSeer metadata are data integrity problems. Based on a small sample, we found that the data quality on the live CiteSeer website appears to be much superior to the quality of the CiteSeer OAI metadata. CiteSeer OAI metadata is automatically generated and it is expected to have errors. The quality and, to a great extent, the validity of citation expansion directly depends on the quality of citation links and `IsReferencedBy` links.

In connection to forward expansion, we assume that if article A cites article B, i.e. $A \rightarrow B$, then $B \leftarrow A$, i.e B is cited by A. This is a fundamental assumption for forward expansion. However, in the CiteSeer metadata, we found that this assumption could be invalid from time to time. The integrity between referenced and is-referenced-by is not reinforced in the metadata. This is a potentially serious problem because the quality of forward expansion may suffer. Therefore we would like to encourage information scientists to consider making batch forward expansion a standard feature for citation analysis and digital libraries. We would like to draw attention to the metadata research community that open access and comprehensive data such as CiteSeer can be very instrumental in the development of new techniques and knowledge.

6. Related Work

The most influential work, as we mentioned in the introduction, is due to Vannervar Bush's visionary Memex and his notion of trailblazing in an abstract and interconnected knowledge space (Bush, 1945). Co-citation analysis, including author co-citation analysis (White & McCain,

1998) and document co-citation analysis (Small, 1973), pioneers the idea of moving back and forth along citation paths.

The Web of Science supports stepwise forward expansion but not a batch mode. CiteSeer's Context is an excellent feature for understanding the value of a cited article in original contexts of citations. Other relevant but different efforts include open archive projects such as OpCite⁴. What we proposed and implemented here is the recursively applicable batch operators for forward expansion and backward expansion. In addition, it is potentially valuable to integrate expansion operators with a wide variety of information visualization environments.

7. Conclusions

We have introduced the notion of citation viewpoints to support the development of novel operators for recursive and holistic trailblazing through scientific literature. We have shown that an initial search topic can be expanded by forward expansion to a diverse range of topics based on citation links as traces in a knowledge space. In conclusion, we recommend that forward expansion should be considered as a standard citation analysis methodology and a practice for topic-oriented information retrieval. We also recommend the use of visualization tools as a mean to explore and understand complex information.

Note

CiteSpace is available as a Java application at <http://cluster.cis.drexel.edu/~cchen/citespace>.

References

- Bloom, B. (1956). *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain*. New York: David McKay.
- Bollen, J., Luce, R., Vemulapalli, S. S., & Xu, W. (2003). Usage analysis for the identification of research trends in digital libraries. *D-Lib Magazine*, 9(5).
- Bush, V. (1945). As we may think. *The Atlantic Monthly*, 176(1), 101-108.
- Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management*, 35(2), 401-420.
- Chen, C. (2003). *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. London: Springer.
- Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359-377.
- Chen, C., & Paul, R. J. (2001). Visualizing a knowledge domain's intellectual structure. *Computer*, 34(3), 65-71.
- Chen, C., & Rada, R. (1996). Interacting with hypertext: A meta-analysis of experimental studies. *Human-Computer Interaction*, 11(2), 125-156.
- Chen, H., Houston, A. L., Sewell, R. R., & Schatz, B. R. (1998). 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.
- Chen, H., Ng, T. D., Martinez, J., & Schatz, B. R. (1997). A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment on the Worm Community System. *Journal of the American Society for Information Science*, 48(1), 17-31.
- Cole, S. (1992). *Making Science: Between Nature and Society*. Cambridge, MA: Harvard University Press.

⁴ <http://opcit.eprints.org/>

- Conklin, J. (1987). Hypertext: An introduction and survey. *Computer*, 20(9), 17-41.
- Donelan, C. I. (2005). *From spider maps to double-cell diagrams: Graphic organizers support student learning*. Retrieved April 3, 2005, 2005, from <http://www.enc.org/features/focus/archive/graphic/document.shtm?input=FOC-003559-index>
- Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(108-111).
- Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Lawrence, S., Giles, C. L., & Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6), 67-71.
- Price, D. D. (1965). Networks of scientific papers. *Science*, 149, 510-515.
- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24, 265-269.
- White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, 49(4), 327-356.
- Ziman, J. M. (1968). *Public Knowledge: An Essay Concerning the Social Dimension of Science*. Cambridge, England: Cambridge University Press.