Visualizing Knowledge Domain Citation and Semantic Structure

Richard H. Fowler
Department of Computer Science
University of Texas – Pan American
Edinburg, TX, USA

Kyle A. Picou
Department of Computer Science
University of Texas – Pan American
Edinburg, TX, USA

Wendy A. L. Fowler
Department of Computer Science
University of Texas – Pan American
Edinburg, TX, USA

Yavuz Tor
Computing and Information Technology Center
University of Texas – Pan American
Edinburg, TX, USA

Abstract - *Researchers are faced with a wide range of tasks when interacting with the literature of a scientific field. These tasks range from determining the field's seminal documents, for the individual beginning investigation in an area, to keeping abreast of current literature and emerging trends, for a scientist working in the field. Visualization can provide one mechanism for ordering documents and revealing structure and relations within a knowledge domain. The system described in this paper provides visual representations of document collections, using both citation information for individual documents and the semantic structure of the document collection, to form interactive visualizations that the user can explore. The system is currently in use with the Citeseer index and provides tools that display the citation structure of a user-defined domain. This bibliometric network is augmented by semantic information derived using a cosine term vector analysis of documents to provide a similarity metric among documents. Supplementary network information from this semantic analysis is used to augment the citation network and provide domain information that reflects documents' content relations.*

Keywords: information visualization, knowledge domain, citation network, bibliometric network, semantic information, graph visualization

1 Introduction

The investigation of knowledge domains is a difficult task. Standard search tools do not provide adequate assistance when researchers are investigating a topic. Though search tools are able to access and return hundreds of documents based on a query, the investigation of these individual returned documents is time consuming. To reduce the time required to gain an adequate overview of a research domain or stay abreast of emerging trends, researchers require tools that can help locate seminal and research frontier papers in a domain. The combination of bibliometric and semantic information can help researchers find these papers. For example, visually evident clusters can direct researchers more quickly to possible seminal papers in the knowledge domain they are investigating.

In this paper we investigate the use of bibliometric information and semantic content extraction in the creation of node link visualizations. The bibliometric information is the citation network formed by documents resulting from a query for "information visualization" in the Citeseer [1] index. The semantic information was provided by calculating the cosine angle between document term vectors. This information is then visualized using the Prefuse [2] visualization toolkit as a node link structure laid out using a force directed algorithm.

2 Related Work

Knowledge domain visualization has been carried out by a variety of different methods. The general steps for any knowledge domain visualization are to obtain a data set, map the data set to a visual structure, extract relational semantic information if so desired [3], and finally display the resulting structure. How these steps are carried out depends on the data set to be visualized and the goals of the visualization.

The methods for gathering a data set can be either programmatic or non programmatic in nature. Generally, the data set is gathered from an available archive of scientific documents in the domain that the researcher wishes to investigate. Archives like the ACM and the IEEE digital libraries provide bibliography information and short abstracts of the articles they index. The problem with these archives is one of both cost and programmatic access to the information. The

use of search tools like Citeseer and Google Scholar has grown to address these problems. These tools provide both programmatic and free access for researchers. Alternative methods have also been explored such as requesting personal bibliography files from known experts [4]. These types of methods usually require an additional data cleaning step before the data set can be used.

Mapping the data set to a visual structure typically entails the application of dimensionality reduction techniques. The general process makes use of the vector space model or some variant to create a set of *n* variables in *m* dimensional vector space. These variables are then projected down to the required number of dimensions. Though extracting the *m* dimensional vectors is relatively easily done, the projection of these variables into two or three variables is a challenging problem at the core of information visualization. Multidimensional scaling is one of many techniques used to perform this dimensionality reduction. Usually, the extraction and use of any semantic information is performed in parallel with this process [5], [6], [7], [8].

Finally, visualization of the resulting structure is performed. Some mapping methods like multidimensional scaling provide a structure that can be directly displayed. Others require the use of techniques such as force directed placement to visualize the resulting structure. Because of the node link structure that is often formed from the use of bibliographic data, force directed placement is widely used because it is relatively easily implemented and produces aesthetically pleasing results [9], [10], [11], [12].

3 System

3.1 Data Set

We chose the Citeseer document library as a source, based on its open availability and support for citation identification. Citeseer has metadata for some 700,000 documents available for download. For each document within the library, this metadata includes a unique document identifier, a title, topic, information about authors, year of publication, list of references, and a few paragraphs from the document, as well as other information. The document's list of references is constructed using a unique identifier for documents in the Citeseer document library. This means that references to outside documents are not considered.

We used Apache's Lucene to build an index of all Citeseer document metadata, creating term vectors for each document using the description of the document. In the examples shown in this paper we then sent the query "information visualization" to the index and constructed our example dataset based on the list of 1,138 documents resulting from this query. For the citation graph constructed from this subset, a mapping from the document set to a node link structure is defined, such that:

- Each document within the set is represented as a node, or vertex, in the graph.
- Each citation is represented as a link, or edge, in the graph.
- As detailed below, for every link we assign a weight based on the similarity of the two documents that are represented by the two nodes joined by this link.

Term vectors obtained from the document description metadata are used to form between document distances used in later semantic network formation. Using the term vectors for document pairs, the distance of two documents is derived from these term vectors. Similarity of two term vectors is calculated based on the vector-space model. Given two documents, d_1 and d_2, the distance between d_1 and d_2 is calculated as,

$$\text{Similarity } (d_1,d_2) = cos\theta = \frac{\vec{d_1}\vec{d_2}}{|\vec{d_1}| \bullet |\vec{d_2}|}$$

where $\vec{d_1}$ and $\vec{d_2}$ are the term vectors for d_1 and d_2 respectively. This similarity measure has the following properties [3]:

- $\vec{d_1}\vec{d_2} \in [0,1]$. Similarity measure is always between 0 and 1
- $\vec{d_1}\vec{d_2} = \vec{d_2}\vec{d_1}$.
- $\vec{d_1}\vec{d_1} = \vec{d_2}\vec{d_2} = 1$

The resulting node-link structure is stored in a relational database which is read by the online visualization algorithm. Two relational tables are used to store this information: *Document (id, title, year, topic, authors, description)* that represents documents within the dataset and *Similarity (source, target, weight)* that represents the links. The visualization algorithm reads the relational tables for the node link structure and creates the node list and edge list which are similar in structure to relational tables. With these lists of nodes and edges, the graph is constructed for visualization.

3.2 Citation Graph Visualization

Once the citation and semantic information have been extracted from the data set, the Prefuse visualization toolkit is used to render the resulting node-link structures. Documents are displayed as nodes in a

graph labeled by their unique Citeseer document id. Document title and complete document information are accessible by the user. Nodes are also mapped to a gray scale to reflect age of the document. The more recently the document has been published the darker its node is displayed. Links between documents are displayed as straight line edges between nodes. A link exists between documents if there is a citation between the two documents, and the links are considered undirected. These links form the citation network of the graph.

Semantic information is used to augment to graph display based on document citations alone. Document similarity information was added to the layout algorithm as a modifier for spring length and spring coefficient. Because we wished to visually support the intuitive notion that documents are closer in semantic space if their content is similar, the inverse of document distance was used to modify the layout parameter that determined edge length. This resulted in short edges for documents that were highly similar by the term vector cosine angle measure. To increase the display affect of the relevance calculation on the resulting graph, the document distance was also used to modify the spring coefficient. This resulted in higher spring coefficients for edges linking documents with more similar content. This helped create tighter document clusters in the final graph display.

For the "information visualization" example the overall graph structure showed several densely connected components with less connected structures surrounding them. Figure 1 is an image of the most densely connected portion of the graph structure. These densely connected components tended to either be surveys of current methods in information visualization or early works in a subdomain of information visualization. Figure 2 is illustrative of the first category

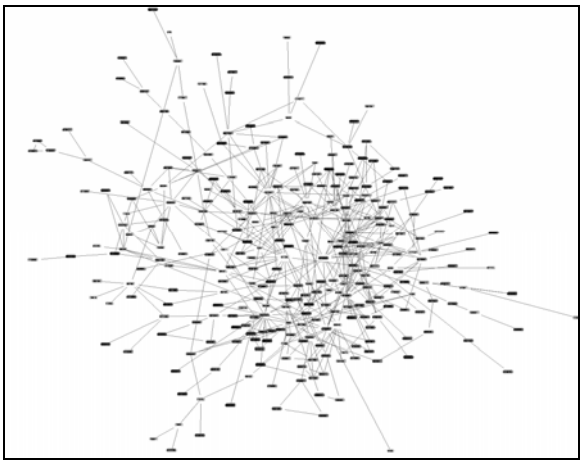

Figure 1. Overview image of the most connected portion of the citation graph structure. Dense clusters of nodes often indicate sub domains within the citation graph structure.

of documents. The highlighted document is a survey of graph visualization and navigation techniques from March, 2000. Neighboring documents are also highlighted and are indicative of the types of reference material that researchers would need to gather to successfully complete such a survey document. Also included in the cluster are documents that cite the survey.

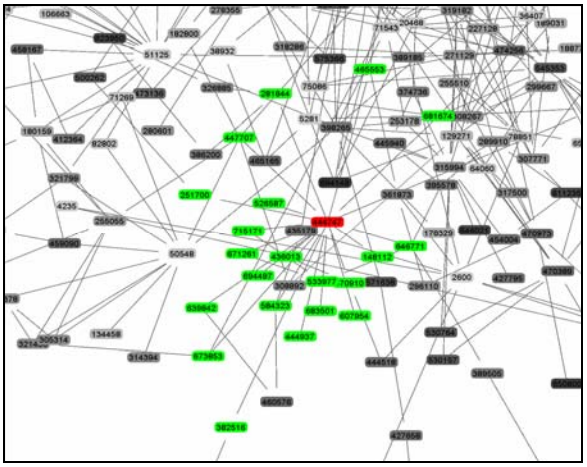

Figure 2. Detailed image of a highly connected potion of the graph. The highlighted node is a survey of graph visualization and navigation techniques. The classification of this node as a hub is easy to make. Darker nodes represent more recent documents.

3.3 Citation Minimum Spanning Tree Visualization

To further investigate graph clustering, a minimum spanning tree was derived using the citation data. This reduced link number and helped to disambiguate many of the graph structures. For the purposes of the spanning tree algorithm, the links are considered undirected. The reason for this was that considering links as directed pruned far too many edges to consider the graph structure meaningful. Figure 3 shows the minimum spanning tree that was obtained from the same data set used for the display of Figure 1. Several document clusters centered on a node of high degree are clearly visible. These high degree nodes are the same as those nodes from the previous citation graph; however, the citation tree allows for much more separation and disambiguation of the nodes. The effect of rendering the semantic information from the document set is also more visible. The large distance separating nodes clearly indicates that those documents possess content that is less relevant to each other by the document vector cosine angle measure.

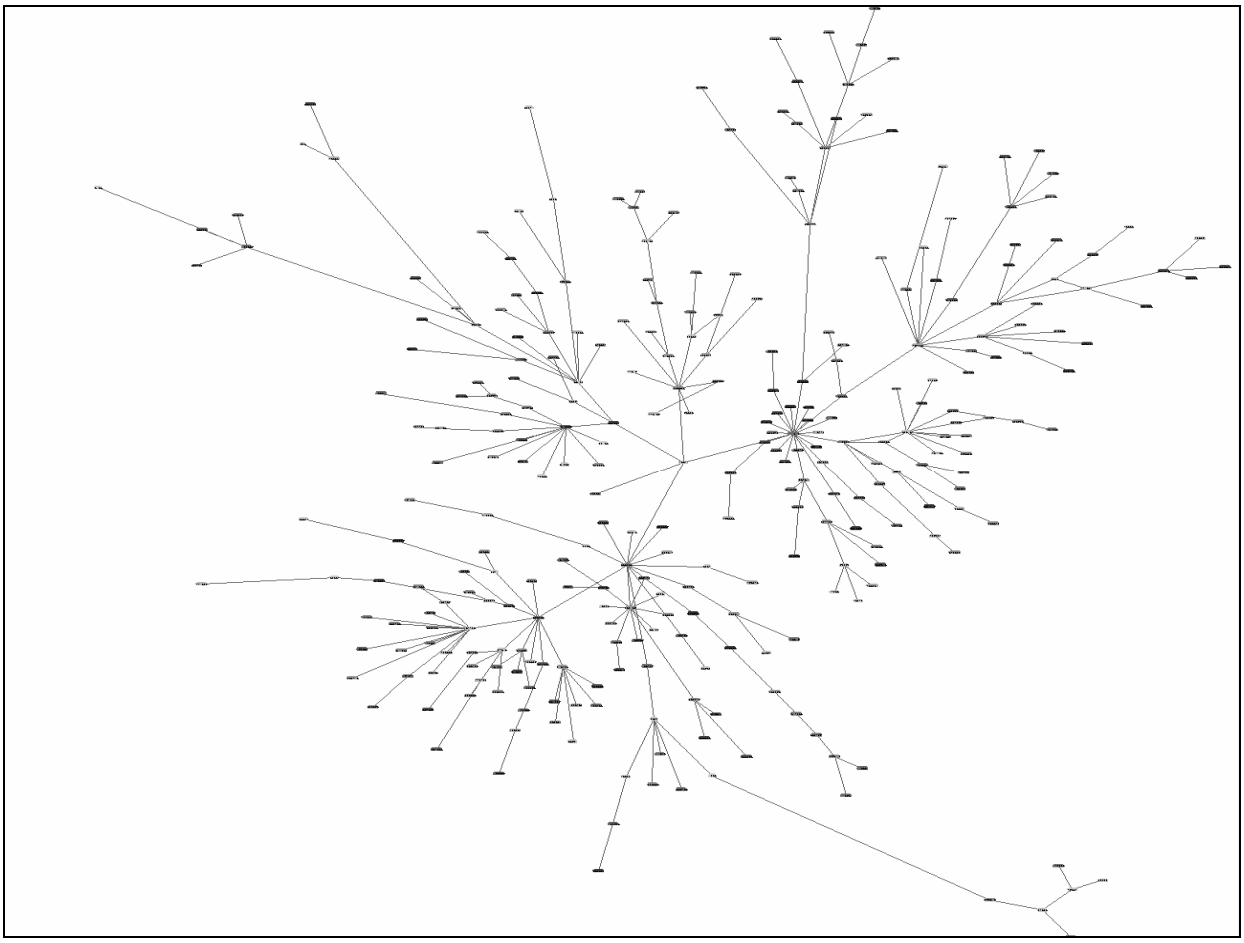

Figure 3. Minimum spanning tree of the citation data. Clusters of documents are much more readily visible in the resulting node link structure, as compared to the graph of Figure 1.

Figure 4 shows in detail the document cluster from Figure 2. Here the semantic information from the document set is much more easily discernable in contrast to the display of the previous figure. The closely surrounding nodes that are more recently published documents all contain subject information ranging from graphsplatting to visual web mining and fit well within the category graph visualization techniques. Nodes farther away are from subject matter that would be considered less directly relevant to graph visualization technique. For example, the farthest node away represents a document on non linear magnification techniques.

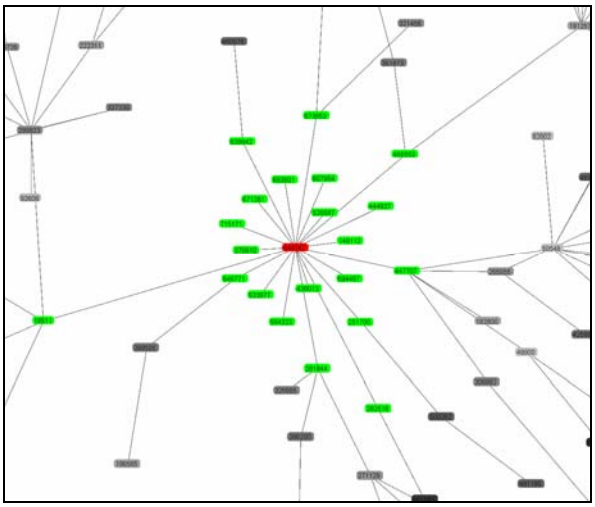

Fig. 4. Detailed image of graph cluster. The highlighted node is the same survey of graph visualization and navigation techniques from Figure 2.

4 Conclusions

The system shows that visualizations of bibliographic data from freely available data sources based on semantic content can be a useful adjunct to other visual representations in discerning structure within a document corpus. Using well understood document

indexing technologies coupled with object oriented visualization libraries allows for the rapid investigation of knowledge structures not easily grasped with text based search tools. Further investigation of the bibliographic data using additional data sources, metrics, and visualization techniques will undoubtedly produce a better understanding. Tuning the current semantic term vector content extraction metric to incorporate additional information from the document excerpts and incorporating additional document element can enhance the investigation process by providing additional similarity data to visualize. Although the current system's use of minimum spanning trees allows for online calculation, other link reduction techniques, such as Pathfinder network scaling, can provide additional representations of underlying structure. A number of additional visualizations also suggest themselves from the current system, for example flow visualizations of bibliographic data based on publication date. We are currently in the process of investigating these types of visualizations more fully.

5 Acknowledgements

This work was supported by the University of Texas – Pan American Computing and Information Technology Center and Department of Education grant P116Z0202159 to the first author.

6 References

[1] Li, H., Lee, W. C., Councill, I., Giles, C. L. "CiteSeer^X: an Architecture and Web Service Design for an Academic Document Search Engine", *Proc. of WWW 2006*, *The International World Wide Web Conference, 2006*, pp. 187.

[2] Heer, J., Card, S. K., Landay, J. A. "prefuse: a toolkit for interactive information visualization", *CHI 2005, Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, 2005, pp. 421-430.

[3] Becker, J., Kuropka, D. "Topic-based Vector Space Model", *Proc. of the 6th International Conference on Business Information Systems*, 2003, pp. 7-12.

[4] Murray, C., Ke, W., Borner, K. "Mapping Scientific Disciplines and Author Expertise Based on Personal Bibliography Files", *Proc. of the Conference on Information Visualization (IV'06)*, 2006, pp. 258-263.

[5] Paulovich, F. V., Nonato, L. G., Minghim, R. "Visual Mapping of Text Collections through a Fast High Precision Projection Technique", *Proc. of the Conference on Information Visualization (IV'06)*, 2006, pp. 282-290.

[6] Chen, C., Lobo, L. "Semantically Modified Diffusion Limited Aggregation for Visualizing Large – Scale Networks," *Proc. of IV 2003, The 7th Annual International Conference on Information Visualization, 2003,* pp. 576.

[7] Chen, C., Morris, S. "Visualizing Evolving Networks: Minimum Spanning Trees versus Pathfinder Networks", *IEEE Symposium on Information Visualization 2003*, 2003, pp. 67-74.

[8] van Eck, N. J., Frasincar, F, van den Berg, J. "Visualizing Concept Associations Using Concept Density Maps", *Proc. of IV 2006, Conference on The Information Visualization*, 2006, pp. 270-275.

[9] Ichise, R., Takeda, H., Muraki, T. "Research Community Mining with Topic Identification", *Proc. of IV 2006, Conference on The Information Visualization*, 2006, pp. 276-281.

[10] Chen, T. T., Hsieh, L. C. "Uncovering the Latent Underlying Domains of a Research Field: Knowledge Visualization Revealed", *Proc. of IV 2006, Conference on The Information Visualization*, 2006, pp. 252-256.

[11] Paulovich, F. V., Minghim, R. "Text Map Explorer: a Tool to Create and Explore Document Maps", *Proc. of IV 2006, Conference on The Information Visualization*, 2006, pp. 245-251.

[12] van Ham, F., Wijk, J. J. "Interactive Visualization of Small World Graphs", *IEEE Symposium on Information Visualization 2004*, 2004, pp. 199-206.