InfoVis 2004 Contest: WilmaScope Graph Visualisation

Adel Ahmed* Tim Dwyer† Colin Murray‡ Le Song§ Ying Xin Wu¶

School of Information Technologies
The University of Sydney, Australia

National ICT Australia‖

Abstract

Our visualisation of the IEEE InfoVis citation network is based on 3D graph visualisation techniques. To make effective use of the third dimension we use a layered approach, constraining nodes to lie on parallel planes depending on parameters such as year of publication or link degree. Within the parallel planes nodes are arranged using a fast force-directed layout method. A number of clusters representing different research areas were identified using a self organising map approach.

Keywords: infovis contest, graph visualization, self organizing map

1 Outline

Our general approach is a fairly standard visualisation pipeline with the following steps: (1) parse the data file and generate the internal data structures; (2) classify papers and authors; generate graphs, filtering edges and colouring nodes according to classifications; (3) find clear layout for the graph; and (4) generate still images and animations of the graph.

Steps 1–3 are achieved using custom Java programs. The result of step 3 is a set of files which can be loaded into the WilmaScope (`http://wilma.sourceforge.net`) graph visualisation system. A custom layout engine was created for WilmaScope which provided improved layout for the scale-free networks that were created by step 2. The final images and animations were created by loading the WilmaScope graph files with the node-positions determined in step 4, into a 3D modelling package.

2 Classification

The paper classification method is based on a spherical self-organizing-map (SOM) [5]. The vector data using to train the SOM is calculated based on a term histogram. Each document is transformed to a vector based on the frequency of words in its title, abstract and keywords. We use a stemmer to change the words into their morphological root and we use a stop-word list to eliminate the meaningless words such as "the", "some", etc. After that, the most popular words ("information", "visualization", "data" etc.) and the words that appear less than 5 times in total are eliminated. 827 words remain so each vector is 827 dimensional. Some papers were removed because no words remained after the above preprocessing. In order to increase the accuracy, we weighted each word according to the term frequency / inverse document frequency. The clustering is then done by a 42 neuron spherical self-organizing map.

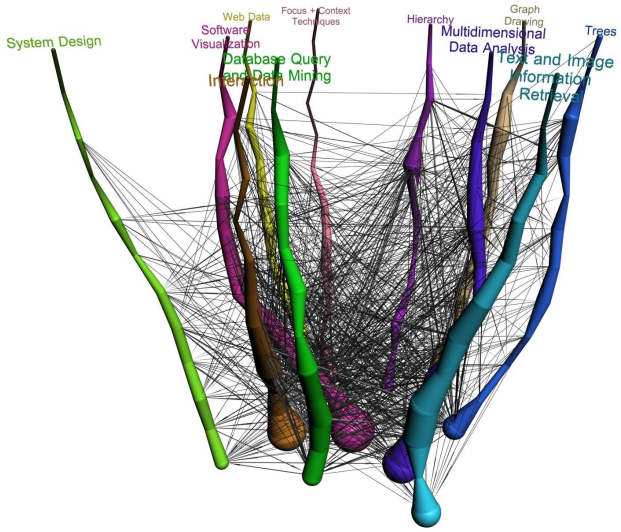

Figure 1: Research Area Evolution

The following research areas were identified based on the first method: "Database Query and Data Mining", "System Design", "Web Data", "Interaction", "Graph Drawing", "Focus + Context Techniques", "Software Visualization", "Hierarchy", "Multidimensional Data Analysis", "Trees", "Text and Image Information Retrieval"

3 Graph Definitions

In order to investigate the data-set from different perspectives, several different graph definitions were used. A graph $G = (V, E)$ consists of a set of nodes V and edges E where each edge $e = u, v$ represents a relationship between two nodes $u, v \in V$.

Citation Network — The citation network is the most direct mapping of the source data to a graph structure G. Each node $v \in V$ is a paper. We considered two definitions for the set V: (1) all papers including papers from outside the IEEE InfoVis proceedings such that $|V| = 1800$; (2) only those papers from the IEEE InfoVis proceedings: $|V| = 600$.

The graph G is directed in that each edge is an ordered pair (u, v) such that u cites v.

Co-Citation Network — Nodes are again papers. Edges are unordered pairs $e = u, v$ such that $u, v \in V$ are papers that appear to-

*e-mail: aahmed@it.usyd.edu.au
†e-mail:dwyer@it.usyd.edu.au
‡e-mail:cmurray@it.usyd.edu.au
§e-mail:lesong@it.usyd.edu.au
¶e-mail:Christine.Wu@nicta.com.au

‖National ICT Australia is funded by the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council

Figure 2: Citation network

gether (co-cited) in at least one other paper's list of citations. Each edge e is assigned a weight w_e according to the number of papers in which u and v are co-cited. $|V| = 523$ and $|E| \leq 7000$. $|E|$ is easily reduced to a more manageable number by thresholding on w_e.

Author Network — Nodes are authors of IEEE InfoVis papers. An edge $e = u,v$ exists if there is at least one paper for which u and v are both authors. The weighting w_e is assigned based on the total number of papers coauthored by u and v. $|V| = 981$ and $|E| > 10000$. Again, in the examples E is filtered based on a threshold on w_e.

Research Area / Author Network — A slight modification to the Author network involved adding a node for each of the 11 research areas identified above. Edges were then created between each author node and the node for each research area in which that author had published a paper.

Research Area Evolution Network — We created a graph (network) of citations between research areas in different years. That is, we created a graph $G = (V,E)$ where each node $u \in V$ represents all the papers published in a particular year in one of the above research areas and each directed edge $(u,v) \in E$ indicates a citation in a paper u of a paper in v.

Before visualising this graph we created additional edges between pairs of nodes representing sequential years in the same research area. Thus all nodes in each research area formed a chain ("worm"), ordered by year of publication. See Figure 1.

4 Graph Layout

Force-directed layout was chosen for its innate ability to show clustering and outliers. In order to better utilise the 3D visualisation space we constrained nodes to lie on a series of parallel planes, see [4]. Various assignments of nodes to planes were tried, for example: in-degree (eg. number of citations) or year of publication.

In order to layout the large graphs in reasonable time a version of the FADE algorithm [7] was implemented, reducing the complexity of the force-directed algorithm with a spatial partitioning.

We found that the citation and author networks were scale-free, with the degree of connectedness of nodes following a power law such that a small number of nodes participated in the majority of edges. We therefore customised the FADE algorithm in several different ways to better show the network topology. The first variation uses edges that repel instead of attracting. No repulsive force is used, instead a force uniformly attracting all nodes to the origin is added. The result is that high degree nodes are pushed to the outside where they can be seen more clearly. The second variation is a more typical FADE method but proceeds in several stages. Firstly, positions are found for the highest degree nodes. These positions are then fixed while lower degree nodes are positioned. This approach is based on an idea from [1] but the third dimension is used to better show the degree-based partitioning of nodes by assigning the different classes of nodes to different layers. One of our renderings, with layers constrained to concentric spherical shells, is shown in Figure 2.

We arranged the research area evolution graph by constraining research area nodes to layers defined by year. See Figure 1.

5 Generating still images and animations

In order to render the final images and animations the graphs arranged by WilmaScope were imported into 3DS Max using a number of custom scripts. The animations were carefully constructed to demonstrate effective navigation through the graph structure — focusing on one node then navigating to another following a camera path that attempts to preserve a viewer's mental map of the graph.

6 Results

Detailed results are discussed in our competition report. However, our most notable findings were (briefly): most highly cited papers were clearly visible as "peaks" in the degree based partitioning of the citation network; erroneous future citations were clearly evident from the research area evolution "worms" visualisation (see Figure 1); and research organisations — headed by a small number of famous, senior researchers, with lesser, typically, postdoctoral researchers in the middle tier and students and affiliates in the bottom tier — were visible as "mountains" in the Co-Citation network (see Figure 2).

References

[1] Visualisation of power-law network topologies. In *Proceedings of the Eleventh IEEE International Conference on Networks (ICON 2003)*, pages 69–74. IEEE, 2003.

[2] Jan William Buzydlowski. *A comparison of self-organizing maps and pathfinder networks for the mapping of co-cited authors*. PhD thesis, Drexel University, 2002.

[3] Chaomei Chen and Steven Morris. Visualizing evolving networks: Minimum spanning trees versus pathfinder neworks. In *Proceedings of the IEEE Symposium on Information Visualisation*, pages 67–74, 2003.

[4] David Dodson. Comaide: Information visualization using cooperative 3d diagram layout. In *Proceedings of the 3rd International Symposium on Graph Drawing (GD'95)*, volume 1027 of *Lecture Notes in Computer Science*, pages 190–201. Springer, 1995.

[5] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Information Sciences*. Springer, third edition, 2001.

[6] A. Quigley. *Large Scale Relational Information Visualization, Clustering, and Abstraction*. PhD thesis, University of Newcastle, Australia, 2001.

[7] A. Quigley and P. Eades. Fade: Graph drawing, clustering, and visual abstraction. In *Proceedings of the 8th International Symposium on Graph Drawing (GD2000)*, volume 1984, pages 197–210. Springer, 2000.