Exploring Categorical Relationships in History of InfoVis publications

CS533 Project Proposal by Alex Gukov (agukov@gmail.com)

1 Domain, Tasks and Dataset

2.1 Domain

When working with a new area of research it is often difficult to grasp its influences and primary directions by only examining individual papers. A high level overview can accelerate this process significantly.

My project focuses on creating such an overview for the InfoVis publication history over an eight year period (1995-2002). Moreover, my visualization will also show the relationship between the co-citation network and the category information extracted from paper metadata. The visualization can then help answer questions such as which papers/authors have influenced a particular subfield and how different subfields are related to each other.

2.2 Tasks

The following are questions which need to be answered by a visualization in order to provide a useful overview of a research field:

2.2.1 Most influential papers/authors

a) What authors have written the most papers?
b) What papers or authors have been cited the most?
c) What is the most influential paper within the last year?

2.2.2 Relationships between authors

a) Identify authors who frequently cite each other or another author.
b) Identify authors who frequently write papers together.

2.2.3 Categories within the field

a) Which categories exist within the field?
b) What is the relationship between authors which publish in the same category?
c) What are the most influential papers/authors within a category?
d) How was a particular category formed (author, paper) and how did it evolve?
e) Are there disconnected groups of researchers working within the same category?

2.2.4 Quality of the data

A visualization which answers above questions can be very useful when exploring a new field. However, when applying the visualization to a new dataset we may also have questions about the quality of data itself. This is particularly important for category information because it is not given to us directly, but rather extracted from the text of the articles themselves(not in our case), keywords, titles or abstracts.

An important application of this project will be to visually identify the most appropriate attributes to use for category grouping in a given set of publications.

2.3 Dataset

My project will use the InfoVis 2004 contest data set. The set presents a subset of the publication history of Information Visualization field over an eight year period (1995-2002). Due to inconsistencies and duplicates in the original data some participants filtered the dataset first. My project will use the result of preprocessing work done by Ke, Borner and Viswanath. The data set includes the following information distributed over a number of relational tables.

2.3.1 InfoVis articles

There are 614 articles from the InfoVis field. Organization of these articles is the primary goal of this project. The metadata for each article includes the publication date, author, references, abstract and keywords. Note that only 429 papers have an abstract, 424 papers have keywords and 340 papers have both.

2.3.2 References

There are total of 3780 references to ACM papers. Among these 1970 are to InfoVis papers within the primary article set, while 1810 are to other non-InfoVis papers. The metadata includes title, authors and publication date.

The remaining 4722 references are to non ACM papers. The metadata includes title, authors, and in most cases the publication date.

3 Personal Expertise

I have done introductory work in document classification using k-means and EM techniques. This experience will be useful for identifying appropriate document clustering method.

4 Previous work

The tasks outlined in the previous sections are a subset of the tasks assigned for the 2004 InfoVis contest.

A particularly effective visualization created by Ke, Borner and Viswanath uses a node-link graph layout to present a paper citation network. The size of each node is proportional to the number of times the corresponding paper was referenced, while the color identifies its date. The application filters the data set removing papers with very few citations, which makes it easier to view the core publications and their relationships.

To show co-authorship relationships the contestants created a similar none-link graph. Each author is represented with a node, where node color encodes the number of citations he/she received and node size indicates the number of papers written. The thickness of edges between two nodes is proportional to the number papers the corresponding authors have written together.

My visualization follows a similar approach while adding category encoding to the co-authorship network view and the paper citation view.

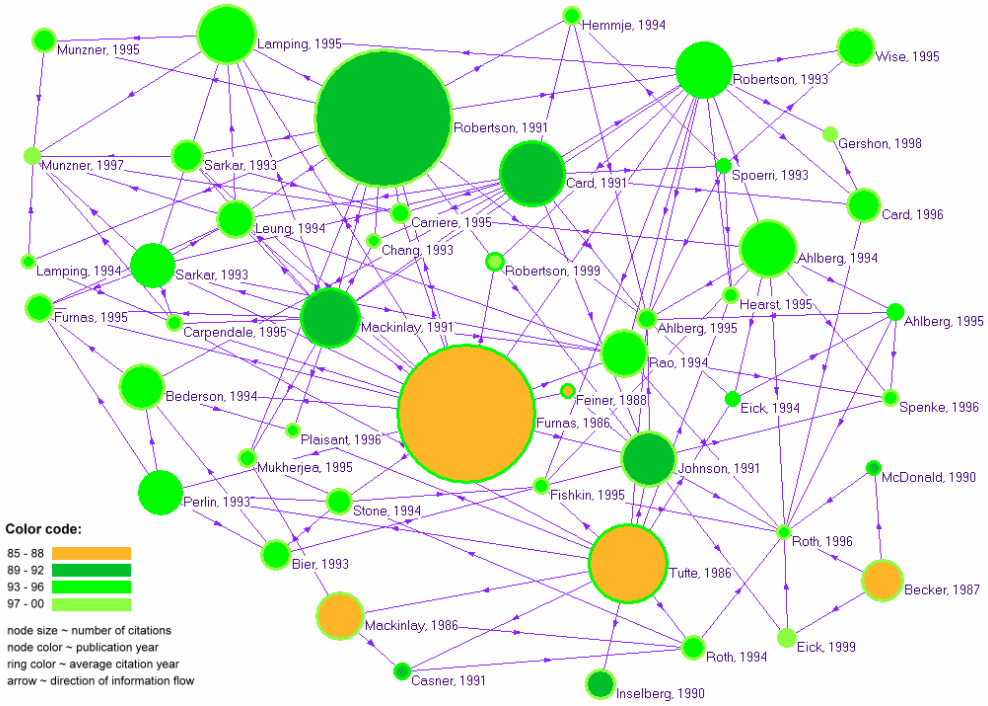

Figure 4.1 InfoVis 2004 contest visualization by Ke, Borner and Viswanath

5 Proposed Solution

With a good visual map of a research field a user should be able to answer all questions posed in Section 2.2. Questions 2.2.1 and 2.2.2 can be answered using the two visualizations presented in Section 4.

To answer questions 2.2.3 and 2.2.4 we need to calculate category information from the dataset and then embed it in the plots.

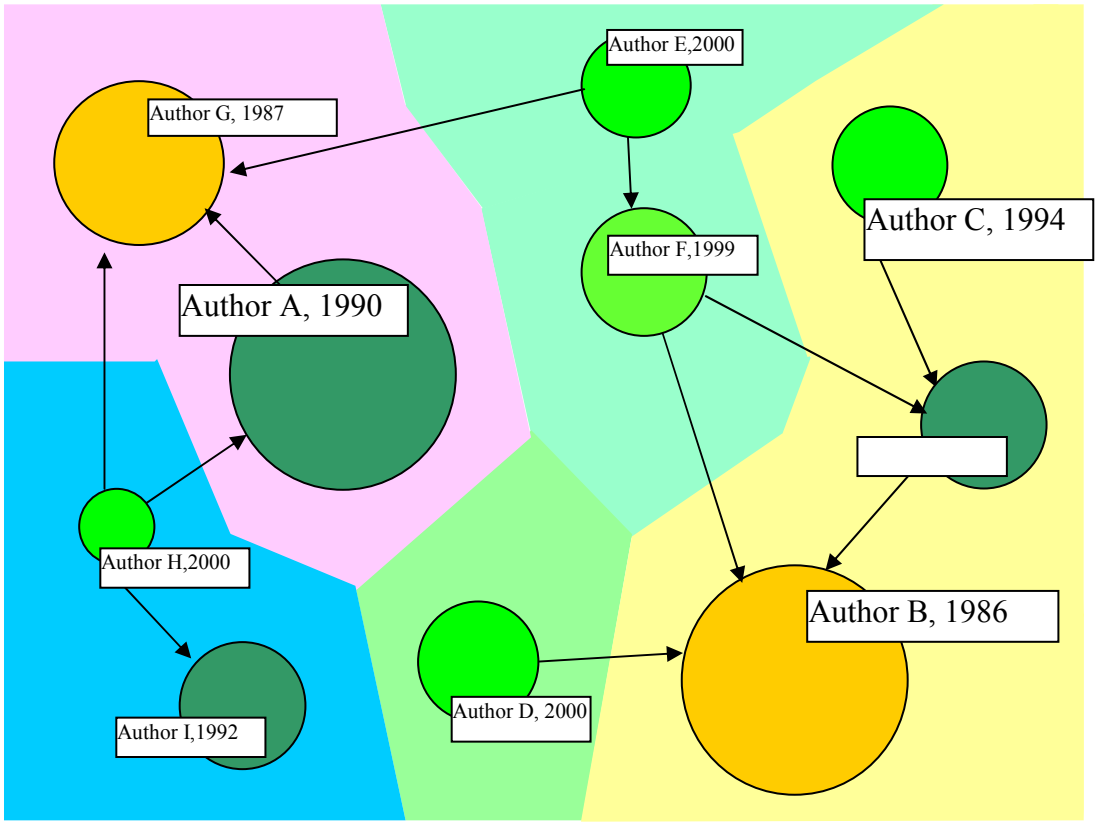

Figure 5.1 Category embedding in a node-link graph of paper citations

5.1 Calculating publication categories

We can extract paper categories by clustering metadata until a small, balanced number of groups is reached. While keywords may be the most natural to use for grouping into categories, they may be too noisy or generic. Better results may be achieved using titles and abstracts. Clusters can be extracted by using k-means or, if more granular division between clusters is found useful, EM.

The primary category of research for a particular author can be acquired by averaging the categories of his/her papers over a specified time period.

5.2 Visualizing categories

To visualize category information I plan to highlight the background around each node with a color corresponding to the category of the node. Because color cue is also used to represent the number of received citations I will use very light background color. Figure 5.1 shows an example of this color coding. The division of space will be accomplished using a Voronoi diagram. Because the segmentation is continuous it is easy to see groups of nodes with the same category. Because the category colors are light, user focus can be easily shifted to other features of the graph.

5.2 Graph Layout

The layout can be accomplished entirely using topology information. Category information can then augment the graph revealing groups of collaborating researchers. However, depending on the connectivity of the graph it may be required to augment the layout algorithm to favour positioning connected nodes with the same category close to each other.

6 Scenario of Use

The user is a graduate student working on an InfoVis project. He/she is looking to become familiar with a particular area within InfoVis as a part of the preparation step. Having reviewed all materials linked directly to a key paper, the user opens the InfoVis publication browser. The user selects the publication view (rather than the author view) and identifies the cluster which he/she has been researching. However, the user also notices an independent cluster with the same category highlighting. The two groups of papers are not related through citations, but have the same theme. The user continues to investigate the second group of papers to complete his/her project.

7 Implementation

The graph visualization will be accomplished using Prefuse toolkit in Java. In order to produce a Voronoi diagram of the publication graphs I will use CGAL C++ toolkit.

8 Milestones

- **Nov 3** – Document clustering investigation completed
- **Nov 10** – Node-link citation graph for papers completed
- **Nov 12, 14** – Update presentation
- **Nov 16** – Space partitioning and category coloring implemented
- **Dec 2** – Node-link co-authorship graph completed
- **Dec 7** – User interface controls for various types of category clustering, visualization options finalized
- **Dec 12** – Draft paper write up
- **Dec 12** – Final presentation
- **Dec 14** – Post presentation revisions, final paper write up

References

[1] Fekete, J.-D., Grinstein, G., Plaisant, C., IEEE InfoVis 2004 Contest, the history of InfoVis, www.cs.umd.edu/hcil/iv04contest (2004)