InfoVis + VAST Keyword Co-Occurrence Network Analysis

Fan Du (fan@cs.umd.edu)

Introduction

Keywords represent the topics of a paper. The co-occurrence of keywords reveals the overlap of multiple topics. By analyzing the keyword co-occurrence network, we can gain insights into the grouping of keywords, the popularity of keywords and the evolution of keywords. The IEEE InfoVis and VAST are two key conferences in the research field of information visualization. In this homework, I used NodeXL to explore the keyword co-occurrence network of the two conferences, and presented three headlines of findings.

Dataset

The two conferences’ publication datasets were combined and used to build the keyword co-occurrence network. Both of them were downloaded from Jigsaw [1]:

- Dataset 1: InfoVis 1995 - 2012, 479 articles.
- Dataset 2: VAST 2006 - 2012, 179 articles.

Keywords (nodes) were extracted from each article’s keyword list. Two keywords co-occur (have an edge between them) if and only if an article contained both of them in the keyword list. The resulting network was an undirected graph. In order to reduce the network size, I filtered out nodes with a degree less than 2. After the data preprocessing, the dataset was containing:

- 226 distinct keywords (nodes).
- 541 keyword co-occurrence records (edges).

Headlines

- Headline 1: Six Major Groups, Three Major Keywords.
- Headline 3: One Clique, Twelve Connectors.
Headline 1: Six Major Groups, Three Major Keywords

In figure 1, the keyword co-occurrence network was divided into 12 groups using Clauset-Newman-Moore algorithm. I laid out each of the groups in a box, with its group name shown at the top left corner. The boxes were sorted by group size. In each group, I chose the name of the highest degree node as the group name. The size of each node represented its degree, and colors were used to represent different groups. In figure 1, we can find six groups each occupies about 1/8 of the whole area: Geovisualization, Interaction, Focus+Context, Information Visualization, Visualization and Visual Analytics. Also, we can find that there are three nodes that are enormously bigger in size than the others: Information Visualization, Visualization and Visual Analytics. This network visualization presents an overview of the information visualization keywords.

Figure 1. Keyword co-occurrence network grouped by cluster using Clauset-Newman-Moore algorithm. Each of the groups is laid out in its own box, with its group name at the top left corner. In each group, the name of the highest degree node is selected as the group name.

In figure 2, I laid out the nodes of the keyword co-occurrence network in a Year-Degree coordinate. The year of a keyword was decided by its earliest occurring year. The nodes with the highest degree each year were highlighted in red color. Four nodes with highest degree can be found in this view: Visualization in 1995, Information Visualization in 1996, Geovisualization in 2003 and Visual Analytics in 2006.

Figure 2. Nodes were laid out in a Year-Degree coordinate. The year of a keyword was decided by its earliest occurring year. The nodes with the highest degree each year were highlighted in red color.
With the help of the Internet, I tried to search for potential reasons to explain the four “outliers”:

- **Visualization (1995):** The history of the *IEEE Symposium on Information Visualization* can be found at DBLP [2]. The first InfoVis was held in 1995 in Atlanta. This explains why the keyword *Visualization* emerged in 1995.

- **Information Visualization (1996):** The concept *Information Visualization* occurred in 1995 at InfoVis (e.g., in the title of [4]). However, as I did not collect keywords from either the title or the content of the articles, its occurring year was delayed mistakenly to 1996.

- **Geovisualization (2003):** The history of *Geovisualization* can be traced back to 1953 [5], and it occurred for the first time in the *keyword field* of InfoVis papers in 2003 as shown in figure 2. Although *Geovisualization* was not a new concept in that year, it became very popular since then and was the forth highest degree keyword in the dataset. Why *Geovisualization* has a high degree? A potential explanation could be found in figure 1. In figure 1, the group labeled *Geovisualization* was shown as the biggest group with the most distinct members (keywords). This indicates that *Geovisualization* has a wide range of applications and / or consists of a wide range of other concepts. As a result, *Geovisualization* would potentially attract researchers from a wide range of fields to study on it and make a contribution to its degree.

- **Visual Analytics (2006):** In figure 2, *Visual Analytics* was shown as a new keyword in 2006. By searching on the Internet, I found that the *IEEE VAST* was held for the first time in 2006, which was an international symposium dedicated to the advances in visual analytics science and technology [3]. This event could reasonably explain why Visual Analytics emerged as a new star in 2006. Since then, Visual Analytics has been widely studied and co-occurred with a wide range of other keywords. It is the most significant keyword since 1997 (based on its degree).
Headline 3: One Clique, Twelve Connectors

In figure 3, I grouped the keyword co-occurrence network by motif, and used Sugivama method to lay out the nodes. The position of motifs were adjusted manually to avoid visual clutter. There is only one clique in this network, which is a 5-clique (in figure 3, CQ 1): Information Visualization, Visualization, Visual Analytics, Interaction and Multivariate Data. This clique indicates that those five keywords are closely related to each other and form a tight “topic community”. Besides, there are twelve connector motifs in the network. All of them are 2-connectors. Take CT 1 (in figure 3) as an example. The keyword Visual Analytics and Geovisualization are connected by the keywords Prototypes, Movement Data and Time Series Analysis. This indicates that the research of Visual Analytics shares those three keywords / topics with the research of Geovisualization. And this suggests that Visual Analytics researchers could potentially collaborate with Geovisualization researchers in those topics.

Figure 3. Keyword co-occurrence network group by motif (N-connector >= 2 and N-clique >= 4). Nodes are laid out by Sugiyama method, and the position of motifs are adjusted manually to avoid visual clutter. Label format for connector (CT): [node A][connector nodes][node B].
NodeXL Critique

My experience with NodeXL was pleasant. NodeXL was easy to download and install, and the developers were very passionate in answering questions from users. After installation, NodeXL became a part of Excel, and I could hardly tell if I was using Excel or NodeXL. There were many great aspects of NodeXL to talk about:

- **NodeXL is seamlessly integrated into Excel.** Most of the Excel operations can be used as usual, especially the “Formulas”. In NodeXL, I can apply my Excel skills to handle the data, and use the powerful NodeXL functionalities to visualize the network.
- **NodeXL is rich in functions.** In NodeXL, I can try different network layout algorithms, I can calculate the network metrics and autofill the node size by degree and I can group the nodes by different clustering algorithms and show groups in boxes, all by just clicking on the menu. In practice, most of the functions ran very fast on my dataset and the results were fantastic.

As a free open-source product, NodeXL is one of the best to me. I used NodeXL very carefully and took notes of the problems I encountered. Also, I tried my best to provide suggestions for improvements:

- **The menu is not intuitive.** Some options are hidden deeply in the hierarchal menus / dialogs and options for different types of objects are mixed together. It cost me a few hours to find the target options, e.g., laying out groups in boxes, modifying the font size of labels, showing axises, counting duplicate edges, etc. I think a powerful search box is necessary, and the options would be easier to find if NodeXL collects and shows the options for edges, labels, axises, vertices, etc. separately.
- **The vertex label layout need further design.** In practice, I could only make the label show in the positions of top, left, right or bottom. In some cases, especially on large datasets, the labels overlap with each other seriously.
- **Need more warnings.** Warnings show up when users try to autofill the node size with degree before calculating the metrics, try to show group in the box before executing the grouping functions, etc. Without warnings, users may be confused and waste time on figuring out the problem.
- **The original undo and redo buttons become non-clickable.** When I wrongly modified the configuration options or the graph, the undo and redo buttons were non-clickable, and I had to manually undo the change and refresh the graph. This may cause disasters when users forget what they have changed, and the unnecessary refresh may cost a long time when the dataset is large in size.
- **Need support for exporting as PDF.** After I finished my network visualization, I seek for functions that could help me to export my results into vector graphics. However, I sadly found that NodeXL only supported exporting images like PNG or JPEG, and the Excel itself only supported exporting data sheets into PDF. I strongly suggest the NodeXL developers to add PDF exporting features. PDF allows editing in a lossless way and would not blur after magnification.

References