An Information Visualization Approach to Summarization of Bibliometric Data:

An analysis of visualization software

Eric Dillalogue*, Santhosh Krishnamurthy, Neil Vapiwala and Michael Zankowski*

Abstract—This paper seeks to evaluate how information visualization of bibliometric data can be used to improve summarization, ultimately aiding users in finding trends in data and improving resource discovery. Using a dataset created from ISI's *Web of Science* database, we evaluate visualizations of over 13,000 records from articles appearing in the journal *Science* between 2003 and 2009. As a baseline, the features of *Web of Science*'s user interface for analysis of citations are considered. This paper examines visualizations from four separate software platforms: Pajek, CiteSpace, ManyEyes, and TouchGraph. The analysis spans several types of visualizations—from network visualizations to tag clouds—to determine which types work best for this dataset. Also considered is the usability of the software platform as well the usefulness of the actual visualizations for the user.

Index Terms—information visualization, summarization, visualization taxonomies, bibliometrics, network visualizations.

---

1 Introduction

The task of data summarization is an increasingly pressing one for users. As more and more resources become available electronically, users are faced with the ever-present problem of information overload. Traditional search engine interfaces are able, with some success, to show the most relevant or popular results; however, they are less adept at providing users with the overall "big picture" from a corpus of text or documents. This is especially troubling for user groups such as scholars and researchers who need to understand the network dynamic within a community of contributors. Our paper looks at the information visualization approach to summarizing data as a possible alternative to the more traditional text-based approach.

Using articles that appeared in the journal *Science* from 2003 to 2009, we created a dataset using ISI's *Web of Science* database. This dataset of over 13,000 records include bibliometrics such as the number of times an article is cited and the number of references in an article. This gave us a large network of related articles and authors from which to create visualization summarizations.

Our task was to determine (a) if information visualization is a more useful approach for this type of dataset and user need and (b) what types of information visualizations work best for this scenario. In the process, we also evaluate various information visualization software platforms with an eye for usability, as this ultimately will help determine if these technologies can be of use for summarization in the future. Finally, we utilized information visualization taxonomies to help determine what visualization components work for this project and whether or not the visualization software in question follow established best practices.

2 Methodology

In this paper we first examine how the *Web of Science* handles the task of summarizing data. We consider usability and the overall usefulness of the citation reporting features for this database. This creates a baseline from which we can evaluate the information visualization approach to summarization. After establishing the advantages and limitations of this approach, we then examine each of the four information visualization software platforms used to create

visualizations—Pajek, CiteSpace, ManyEyes and TouchGraph. First, using Chi's Data State Model [1] and Shneiderman's taxonomies [2], we seek to understand how these visualizations fit into the overall schema. We then look at the usability and the overall usefulness of each visualization and system for data summarization. Finally, we compare the both the text-based *Web of Science* results with the different visualizations created to determine which visualization types and software work best for this task.

3 Web of Science

3.1 Overview

ISI's *Web of Science* is a scientific database that incorporates bibliometric measures of journal articles such as the number of references and the number of times cited. The inspiration behind this database is that users can understand the scientific community's research trends more clearly by exploring the network of relationships between authors or articles: who is citing whom, which articles are more frequently cited, etc. *Web of Science* includes a search engine interface in keeping with the standards in information retrieval, with a results page listing surrogate records for articles, ranked in a specific order, and including pertinent metadata (in this case, bibliometric data).

Going through all of the surrogate records manually is not very conducive to summarization of the data. *Web of Science* correspondingly does include a feature called "Citation Report" to analyze the bibliometrics. Interestingly, this feature does include some degree of visualization with bar graphs for published items and citations in the year. Figure 1 illustrates the results of a "Citation Report" for our dataset. While it includes some visualization of data, the core component of the analysis remains a list of surrogate records describing each article in the collection. The report is rich with metrics, such as average citations per item and citations per year.

3.2 Usability

Web of Science's search interface and citation report both score well in terms of usability. It has the added advantage of being a familiar sight—the search engine interface—to most users. Labeling is clear, tasks are easily identified, the user is constantly kept abreast of the systems status, and navigation is relatively straightforward.

3.3 Usefulness

Although *Web of Science* does provide some very informative metrics, it leaves much to be desired in terms of summarizing the data and allowing the user to quickly grasp the "big picture." From the citation report it is difficult to understand the relationships between different authors, articles, or groups of articles or authors. Where this interface excels is in listing the top articles by certain metrics. However, the user is left to determine how the article in question is related to others in the collection. If the user task is simply to find the article most frequently cited, this approach solves the problem efficiently; however, if the user wants to view a community of scholars who, say, are citing each other on the subject of protein synthesis, this task is not well supported. It should also be noted that *Web of Science* does not support the "Citation Report" feature for datasets over 10,000 records. For this project, the dataset needed to be trimmed down to meet the system requirements.

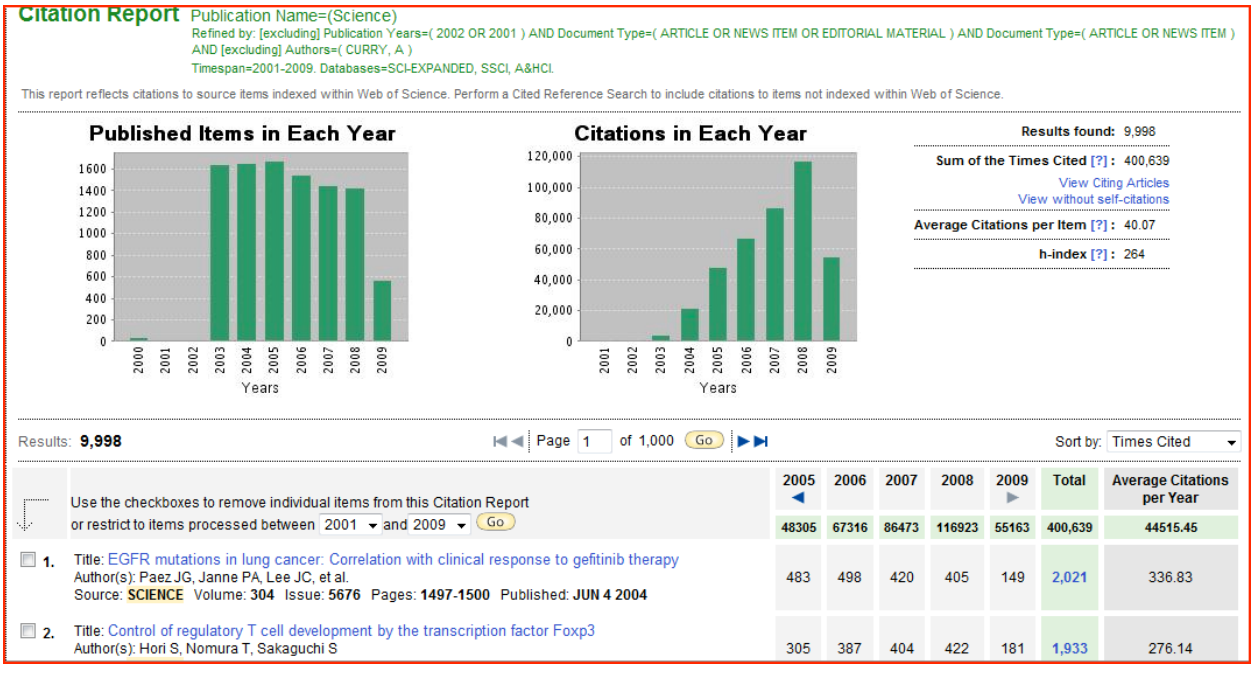

Figure 1: ISI *Web of Science*

4 Pajek

4.1 Overview

Analysis of Figure 2 starts with the largest cluster on the upper left-hand side of the figure. It has a bottom portion that is separated from its upper half by a line in between.

First, analysis of the bottom half: right-clicking on a node in the middle of the cluster, we find several references. A much-referenced article here is an E. S. Lander article from 2001 entitled "Initial sequencing and analysis of the human genome." Additional analysis of other articles shows that this cluster deals mostly with the human genome and related human genetic topics.

The midpoint of this cluster, which seems to separate one half from the other, is an article by Victor Ambros, "The Functions of Animal MicroRNAs." Further analysis of the top half shows that many of the articles again deal with RNA and further discussion about protein–coding genes and their influence on various organisms.

4.2 Taxonomy

By looking at the Pajek visualizations in terms of Chi's taxonomy and Shneiderman's Task by Data Type taxonomy, we get a clearer understanding of how effective this information visualization tool can be.

First, in terms of Chi's model,

1. **Value** – The raw data, which is loaded as a .net network file
2. **Data Transformation** – The visualization separates these records into nodes and vectors (if selected) showing the connectivity of the loaded network
3. **Analytical Abstraction** – The data about the data, or in this case references, which can be displayed by right clicking on the nodes, filtering is done by removing the nodes from the drawing
4. **Visualization Transformation** - Exists in a few ways by choosing different layouts of energy and dimensions using the drop-down menu on top
5. **Within View** – Functions allow for scrolling, rotating, and labeling of certain clusters with vector information and size distortion for nodes

Shneiderman's taxonomy breaks down Pajek's visualizations into two simpler aspects: tasks and data types. Pajek gives the user a few options in terms of data types even though it needs a network type to begin. Once a network is loaded, however, the ability to view in 2D and 3D forms is available.

1. **Overview** – Pajek supports an overview in the sense that you can see the relation (or connectivity) in a picture view; however, the user cannot get an overview of one particular selection of interest in one or two clicks
2. **Zoom** – This function is not well supported
3. **Filter** – Can be done by deleting uninteresting nodes which are then removed from the visualization
4. **Details-on-Demand** – Provided by right clicking on nodes thus showing connectivity between referenced materials
5. **Relate** – Essentially shown in the way in which nodes are connected but only enhanced once the details-on-demand functionality is used
6. **History** – Tasks supported by clicking "Previous," "Next," and "Redraw" in order to shuffle through actions and refine the visualization progressively
7. **Extract** – Supported by exporting the visualization or certain aspects of it in 2D, 3D or appending it to an existing Pajek project file

4.3 Usability

Pajek is not completely intuitive in its initial use. That is to say, a user must experiment with various different functions before discovering the most beneficial visualization for a given set of data. However, once explored, Pajek shows its value (especially for larger networks) in the ability to view relationships among networks.

4.4 Usefulness

Our goal for this visualization was to separate clusters and identify related topics within them. After breaking the visualization into portions that were easily differentiated, we could right-click on the nodes and look up the topics of the material. This, we believe, was the most useful aspect of the visualization. It allowed us to get a general idea of what kinds of topics were discussed and which articles were most often referenced in the area of discussion. The more a particular source is referenced, the more likely it is to have some significant research findings accompanied by notoriety within its particular field.

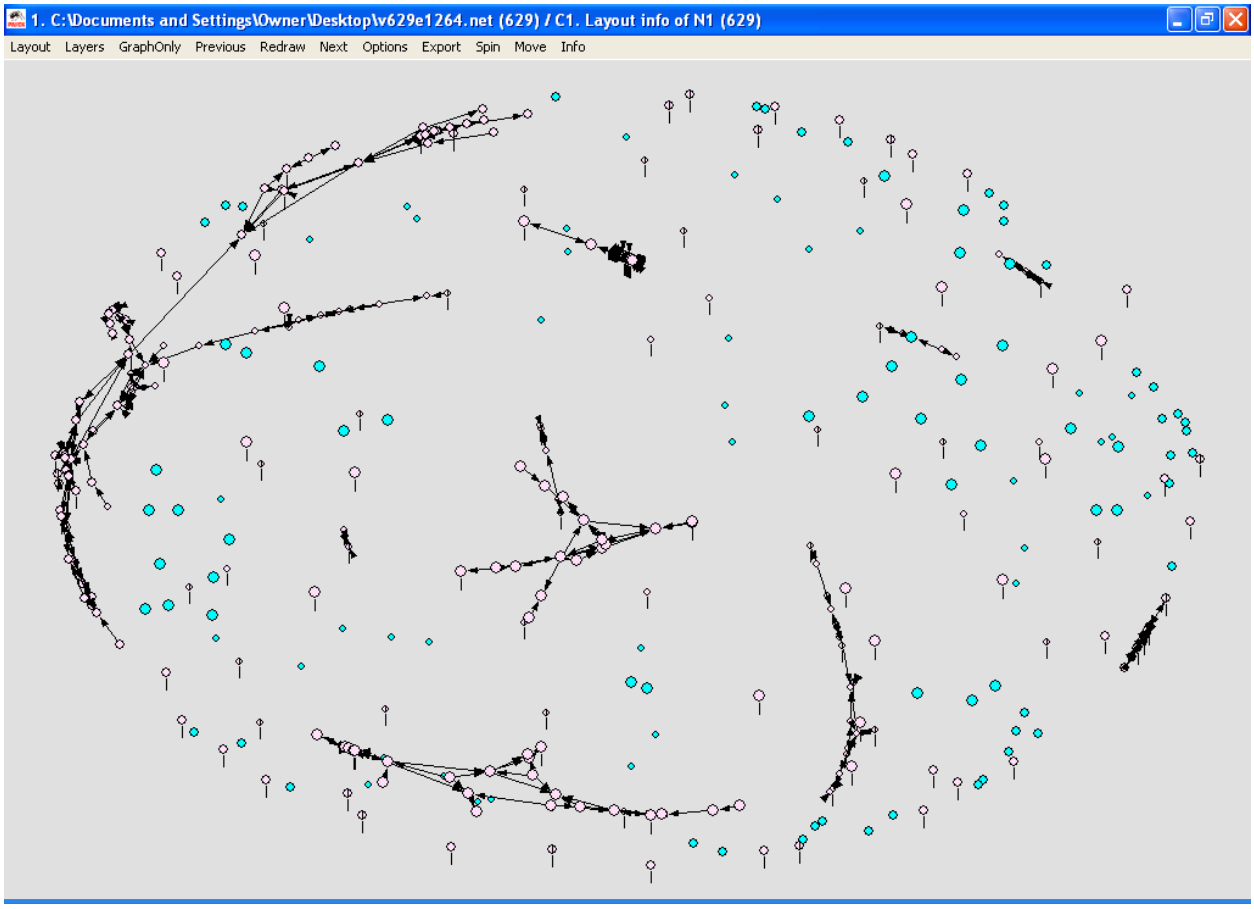

Figure 2: Pajek

5 CiteSpace

5.1 Overview

The CiteSpace visualization can appear cluttered at first– often the initial attempt results in a mess of nodes and links. To make it more understandable visually, the data was limited to the top 1% of the most cited articles and, of that selection, only 25 nodes per year. The resulting visualization is sparser, but not necessarily clearer. We can see in Figure 3 that there is one heavily linked cluster in the center, a few other clusters with a moderate amount of linking, and several nodes with either one link or no links at all. However, since the top 1% of cited articles were chosen, some of the lone nodes are quite large (indicating a high number of citations to that article).

5.2 Taxonomy

Applying the taxonomies of Chi and Shneiderman to CiteSpace gives a better understanding of the capabilities of the system.

Under Chi's Data State Reference Model, CiteSpace has these characteristics:

1. **Value** – Consists of .csv datasets from the ISI *Web of Science* that include the authors, journals, article titles, and co-citations from a selected search
2. **Data Transformation** – Extracting the article information from the .csv files (such as author, title, journal, etc.) and the connections between these articles through their citations
3. **Analytical Abstraction** – Involves dynamic value-filtering and the visualization transformation
4. **Visualization Transformation** - Creating a colorful network from the datasets wherein the articles become nodes of varying sizes (based on how many other articles cite to them) and links form between these nodes (to note citations)
5. **View** – Options to manipulate the visualizations that include scrolling, zooming, dynamic view-filtering, changing colors, and moving the networks

CiteSpace falls under the category of a "network," in Shneiderman's taxonomy, using the familiar construct of nodes and links. The system fulfills nearly all of the

required tasks that Shneiderman outlines for an information visualization tool.

1. **Overview** – The first view is of all of the clusters and their links
2. **Zoom** – There is an option to zoom in on any of the clusters
3. **Filter** – Possible through a variety of controls in the right-hand menu or drop-down menus
4. **Details-on-Demand** – Revealed by clicking on nodes to reveal a particular paper's author, year of publication, and title
5. **Relate** – Designed to show the relationships between articles, representing co-citations as links and clustering related nodes together
6. **History** – There is the one task that is not well supported as there is simply no easy way to undo an action once it has been processed, though the original search window remains open for re-processing the visualization
7. **Extract** – Data about the material can be extracted from the system through a variety of means: by data tables, by Pajek .net networks, and by HTML links to specific articles via services like Google Scholar

5.3 Usability

CiteSpace is moderately complex to use initially, but quickly becomes simpler with more experience. The system does a good job of labeling the controls, though the number of options can be overwhelming at first. The visualizations open in a new window–a smart design choice since the user can easily look back at the parameters they set or begin a new model. Manipulating the actual visualization is a bit more difficult, as some of the functions do not automatically produce an effect and only become active after manipulating other controls (such as the cluster labeling slider function). A few of the controls seem to have little to no effect (like the "relaxer" slider) and others can result in transformations that are difficult to stop or undo (like listing LSA ranked terms).

5.4 Usefulness

The original idea to use this dataset was to see if there were trends or interesting developments in the articles published in the journal *Science* over a little more than five years (2003 – 2009). The CiteSpace visualization, since it relies on the links between articles, certainly shows us what important topics were discussed during that time period. For example, we can use the link walkthrough option and see that the largest cluster is from 2004 and consists of two groups – one that discusses the "gusev crater" and the other the "meridani planum." Those terms come from the topography of Mars and coincide with the landing of the Mars Rovers, Spirit and Opportunity, early in 2004.

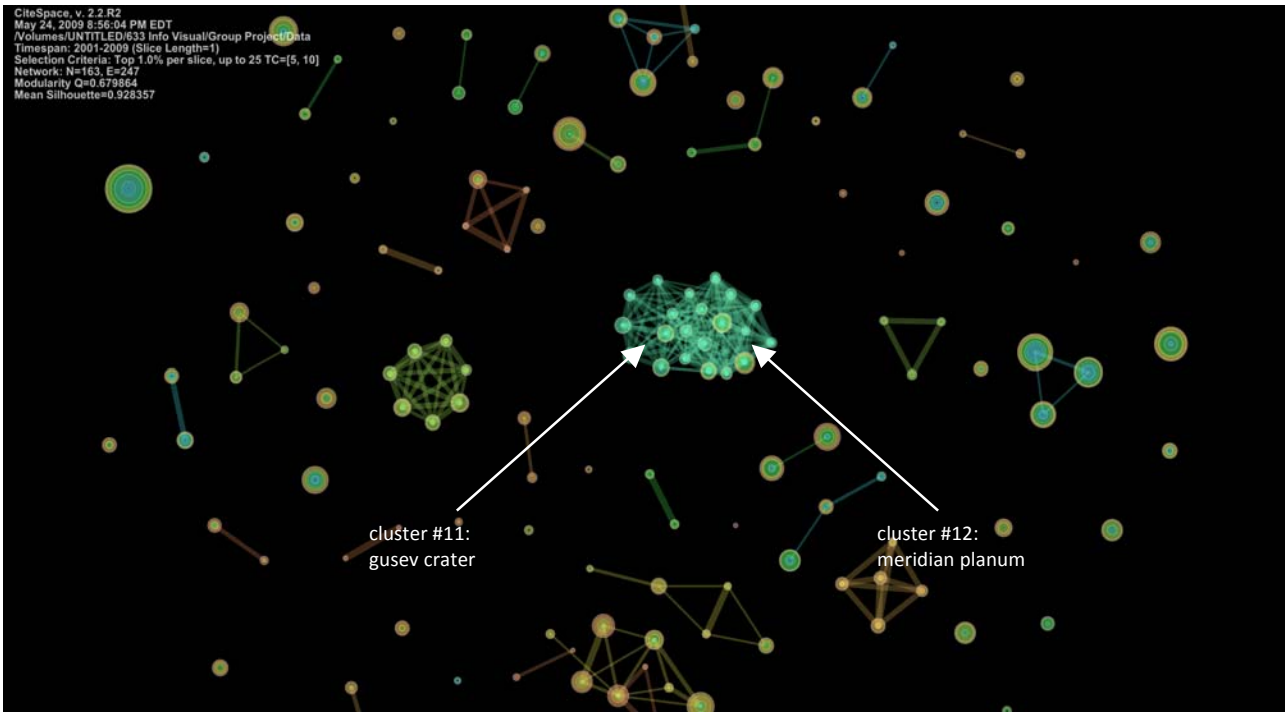

Figure 3: CiteSpace

Another interesting result from the visualization is to see the re-emergence of a topic. In the timeline view of the data, the temporal relationships of the clusters become apparent. One of the clusters (number four) uses the concept of a "ring species." A ring species is a biological term to describe "a ring of populations in which there is only one place where two distinct species meet" [3]. This particular cluster of articles links back to a piece by J.B.S. Haldane from 1922, one of the oldest papers in the entire dataset. This article sets forth "Haldane's Rule" which "has important implications for understanding the evolution of interspecific reproductive isolation" [4]. In the cluster view, this temporal relationship would not have been readily apparent since the nodes and links are only color-coded for the years 2003 to 2009.

CiteSpace easily pulled out the most frequent terms used in the *Science* articles. Using the top 1% of cited works, the terms ranged from "circadian rhythm" to "human chromosome" to "recent ice-sheet growth." Perhaps the most useful take away from the visualization, however, is that many of the journal articles are distinctly part of their own domains. That is, there was very little crossover in citations of the top 1%, with nearly all of the links existing between articles that are very similar to one another. This shows the trend of modern scientific literature towards becoming narrower in focus and topic, rather than more general and wide-ranging.

6 ManyEyes

6.1 Overview

ManyEyes is a web-based information visualization platform that enables a community of users to upload datasets and create pre-determined visualizations. It can be thought of in two parts: (a) contributing datasets to the community of users and (b) manipulating the data into visualizations and making these visualizations available to the user community. For the first part, ManyEyes accepts most tab-delimitated forms of data, with some metadata to describe the fields. For the purposes of this project, an Excel spreadsheet was used to upload the data. ManyEyes has a five-megabyte upload limit, so our original Excel file needed to be altered to meet the system requirements. This was accomplished by removing several redundant or irrelevant fields from the original *Web of Science* citation output data. Once uploaded successfully, the dataset remains accessible to all users, who may create visualizations based on this dataset at will. It allows for a multitude of visualization types, including: bar charts, histograms, and scatter plots. ManyEyes also encourages tagging and other social networking tools to increase the exposure of the dataset to the user community. Once uploaded, there are several options available for the users to create visualizations. Many of the visualizations, such as Wordle and TagCloud, are intended to use free text datasets. There are, however, other visualizations that can accommodate more structured datasets, such as the one used in this project.

During the visualization process, ManyEyes allows for various levels of customization of the visualization, depending on the visualization being used. Although ManyEyes does support many type of visualizations and is highly usable, its interface is fairly static once the visualization has been created and the features limited once the user has selected a visualization to create.

6.2 Taxonomy

Using Chi's Data State Reference Model, we can categorize ManyEyes as a spreadsheet information visualization adhering to the following data stages and transformation operators:

1. **Value** – The text from *Web of Science*
2. **Data Transformation** – Editing metadata fields to structure tabled-based dataset
3. **Analytical Abstraction** – Transformed Excel spreadsheet
4. **Visualization Transformation** - Customization of visualization and features
5. **Visualization Abstraction** – Revised dataset
6. **Visual Mapping Transformation** – This is largely invisible to the user and handled behind-the-scenes by the system
7. **View** – Final created visualization, largely static in ManyEyes

Shneiderman's taxonomy is more difficult to apply to ManyEyes given its how many different types of visualizations are offered by ManyEyes. Some visualization are one-dimensional while others are two or three dimensional in nature. We can, however, evaluate ManyEyes against Shneiderman's seven-tasks:

1. **Overview** – This is perhaps ManyEyes's biggest strength, with easy summarization of data and readable, informative visualizations
2. **Zoom** – This varies from visualization to visualization, but is generally not well supported
3. **Filter** – This is supported to some degree during the creation of visualization, but less so once the visualization has been created
4. **Details-on-Demand** – This is generally well-supported and useful with regard to resource discovery of articles and authors

5. **Relate** – The question of relating items to each other is largely dependent on the type of visualization in question when using ManyEyes
6. **History** – This functionality is not well-supported
7. **Extract** – Printing or downloading records is not well-supported

6.3 Usability

Evaluation of ManyEyes usability needs to be considered with regards to two user groups. First, what is the usability of the system for users creating visualizations? In this regard ManyEyes excels since it involves a relatively effortless process to create the visualization. Tasks are simple and understandable. The use of icons, prevention of error, and other usability best practices are employed. The success in terms of usability is most attributable to the lack of options and features for ManyEyes. Once the user selects the visualization, there is little the user can do to augment or customize the visualization. It should also be noted the system does a poor job of providing feedback and explaining why a visualization will not load properly. Where ManyEyes can be faulted is in lack of documentation, especially in the area of uploaded datasets to the site.

The second user group to be considered is the users viewing and interacting with the visualizations. Since the options are so limited and the visualization so static, it is difficult to make a usability assessment, but for the options ManyEyes does provide, it performs well.

6.4 Usefulness

Overall, ManyEyes produces some very useful visualizations for the task of summarization. Some visualizations work better than others. For example, while the system can visualize the project dataset into Wordle, more structured datasets work better in other visualizations, such as bar graphs or treemaps. Take for instance the visualization of the *Web of Science* data into a scatter plot, as seen in Figure 4.

This visualization enables the user to understand quickly the relationship between cited articles and number of references. It also allows for easy resource discovery of pertinent author or article information from a collection of several thousand. In this regard, ManyEyes succeeds in providing summarization to the user. The user can quickly analyze trends and patterns of the *Web of Science* data that would be much more laborious in a more traditional texted-based approach to summarization. However, the visualizations in ManyEyes for any given dataset are very hit-or-miss. Unfortunately, ManyEyes does not provide much documentation to aid users in deciding which visualization to use, opting instead to have users experiment with the system.

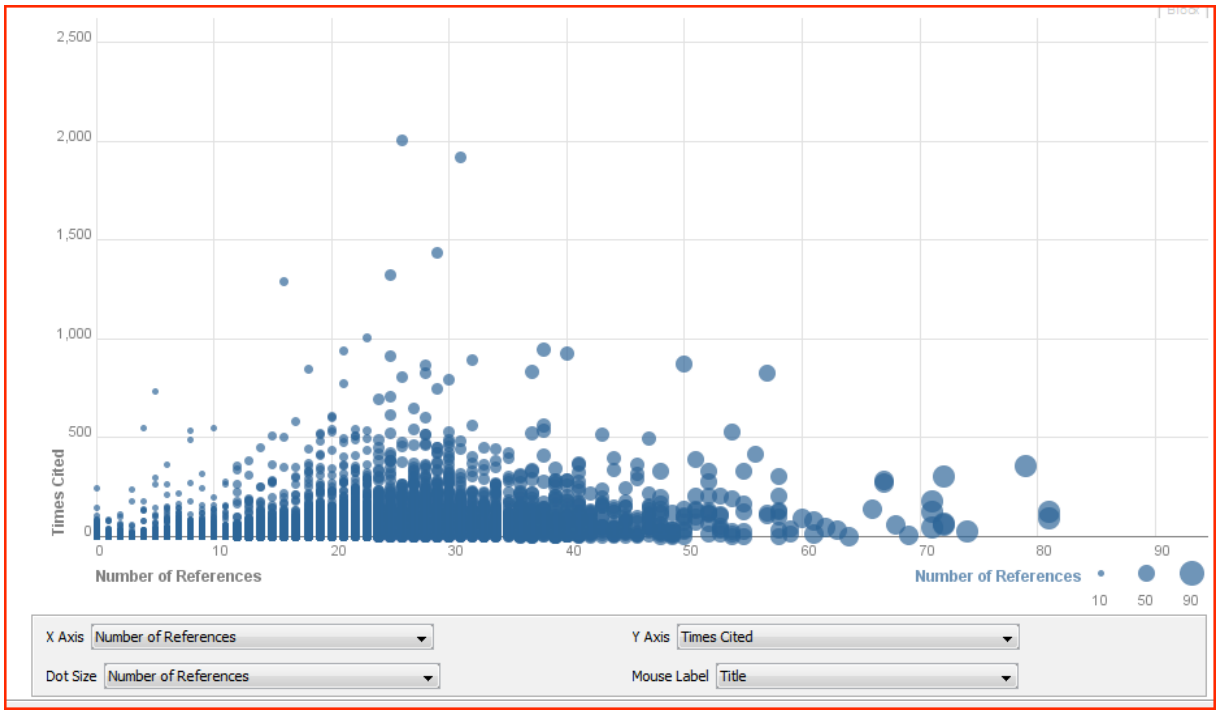

Figure 4: ManyEyes

7 TouchGraph

7.1 Overview

TouchGraph was founded in the year 2001 by Alex Shapiro. The company specializes in creating interactive visualization tools that convert information sources into more meaningful and interconnected representation of objects. An Excel dataset was used to create the visualizations in TouchGraph. Authors were classified as Items and the corresponding titles were the Nodes. The nodes of titles are clustered and edged with related authors. Figure 5 is a visualization based on five degrees of separation. More degrees of separation involve processing more data and hence more information being available. Due to very high processing time, we have limited the degrees to five. The halos are the round circles that appear behind every item. The halos, items, and nodes can be individually formatted for font size and color. The halo circle is larger if the number of relationship is higher. For example if title #1 has 50 authors and title #2 has 10 authors, then the halo for title #1 will be having a larger circumference than title #2.

7.2 Taxonomy

The TouchGraph tool overlays Chi's Data Reference Model very effectively:

1. **Value** – An excel file of the data from the *Web of Science*
2. **Data Transformation** – Supports data and visualization transformation via a wizard-like interface where items, nodes and attributes can be defined
3. **Analytical Abstraction** – Transformed Excel spreadsheet
4. **Visualization Transformation** - A snapshot of the transformation is displayed to the user based on the user's choice of categorization and grouping options
5. **Visualization Abstraction** – None supported
6. **Visual Mapping Transformation** – Performed by the cluster and the layout algorithm
7. **View** – The cluster algorithm identifies long edges and connects them with other clusters and the loading algorithm creates and positions the node, it also creates and recreates any animations related to the visualization when certain filters are applied

TouchGraph can be compared to Shneiderman's task by data type taxonomy:

1. **Overview** – TouchGraph provides an overview of all the connections ("edges") between the Title ("nodes") and the Authors ("items")
2. **Zoom** – Zooming can be performed from a macro level to a micro level involving a single node
3. **Filter** – Nodes that are of interest can be selected and the rest filtered
4. **Details-on-Demand** – On clicking any of the nodes you can view the full details
5. **Relate** – The left hand pane displays the relations between various items and the nodes
6. **History** – None supported
7. **Extract** – None supported

Multi-dimension visualizations are also possible through TouchGraph as Shneiderman mentions that "items with n attributes become points in a n-dimensional space."

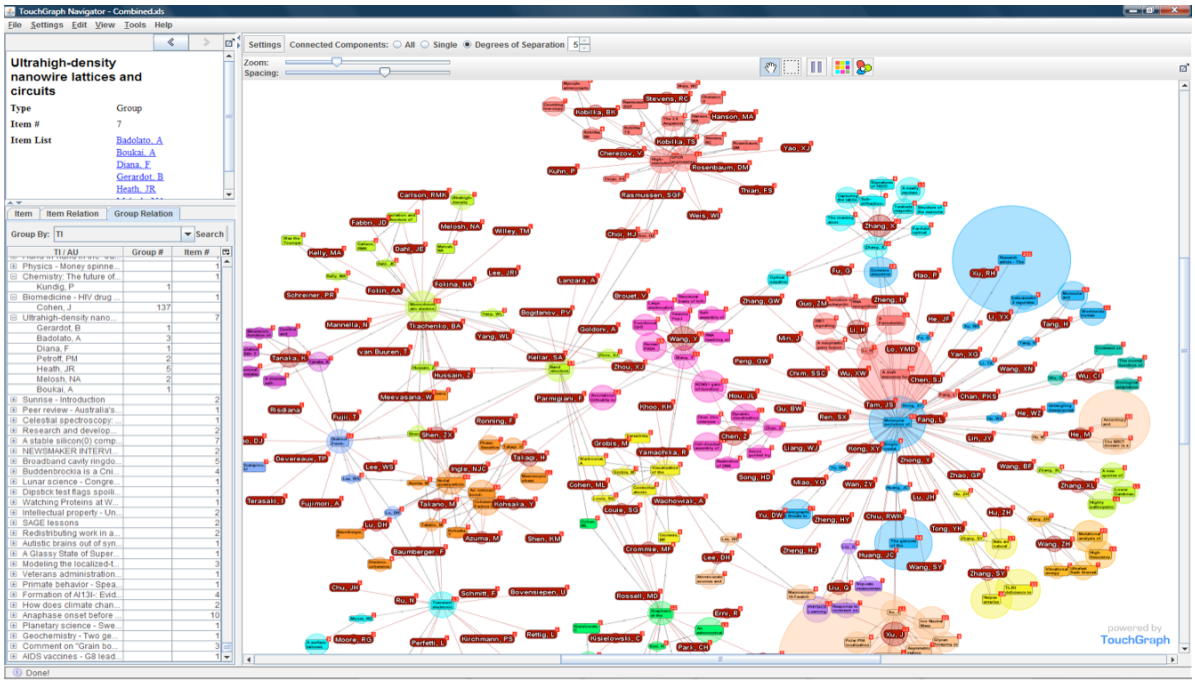

Figure 5: TouchGraph

7.3 Usability

The tool expects datasets like any other visualization tool, but it accepts datasets in most basic formats like Excel, .csv and XML files. The datasets are easy to prepare with the above three formats. Unlike Pajek, TouchGraph accepts only certain data formats that have to be converted using converters that are available with the tool or from CiteSpace. The tool has an easy to create, wizard-like visualization interface to create the visualizations. It also has a good help manual to reference on any topic. After creating the visualization, the user can change the settings to format everything.

7.4 Usefulness

When using regular search engines like Google or Yahoo, it is difficult to link the data and get more meaningful information. This tool links the data and provides information based on user selection. This gives more control to the user in choosing what kind of information they want to see. For example, in Figure 5, we can isolate one particular topic and identify the authors who wrote the book. This is at each node level. This information can also be clustered based on authors and the titles they are associated with. These are only two examples of information visualizations using this tool. There are more features available that can be explored with different datasets.

Some of the interesting aspects of TouchGraph are the ability to group similar results into one and the ability to interconnect and display the connections with similar results. The other features are the "Keep" and the "Expand" buttons. Since information overload may occur for the user, they can decide what elements are required and discard the rest. This provides the user with the freedom of narrowing down their search and, since they see only the relevant results, the visualization provides a much clearer summarization. Another feature is to color the search result clusters, wherein conceptually related websites are given one color. The user can customize the color choices if they desire.

8 Comparison

8.1 Information Visualization vs. Text-Based Approaches to Summarization

The task of summarization appears, overall, to be more successful using a visualization approach rather than a more traditional text-based approach. The *Web of Science* database provides a text-based tool for analysis of a large dataset; however, it remains difficult for the user to grasp trends—such as citation, authors, or topics—quickly. Curiously, *Web of Science* also does not support analysis for datasets with over 10,000 records, giving a clear advantage to the visualization softwares that can handle these larger datasets. Using Shneiderman's rubric, *Web of Science* does perform fairly well: it informs the user of the search history, allow for printing records, etc. It does not, however, do a particular good job at providing an overview. The visualization approach, on the other hand, does do a good job at providing an overview of the data, which is essential to meaningful summarization. The different softwares performed with various degrees of effectiveness. The most successful tools proved to be much more useful in aiding the user to summarize the data than using *Web of Science* alone.

8.2 Visualization Softwares

Visualization softwares and specific types of visualizations varied a considerably in terms of effectiveness for summarization. We can use the aforementioned taxonomies as a basis for understanding these tools and why these tools are effective (or not). The following table illustrates how well these tools adhere to Shneiderman's suggestions for information visualization systems (Y=fulfills requirement, N=does not fulfill requirement, ?=fulfills in some cases or partially fulfills requirement):

Tool	Overview	Zoom	Filter	Details-on-Demand	Relate	History	Extract
Pajek	Y	N	Y	Y	Y	Y	Y
CiteSpace	Y	Y	Y	Y	Y	N	Y
ManyEyes	Y	?	Y	Y	Y	N	N
TouchGraph	Y	Y	Y	Y	N	N	N

Table 1: Comparison of InfoVis Tools

We can see here that CiteSpace adheres most closely to Shneiderman's suggestions of functionality in a visualization system. Other tools, such as ManyEyes, meet some of the criteria but not all and suffered correspondingly in terms of usability, usefulness and overall effectiveness for summarization. Usability varied considerably among the tools. Less usable tools, such as Pajek, though very powerful and feature-rich, ultimately were not as useful since the user is unable to effectively manipulate the software and create understandable visualizations.

The most effective visualization tool, though not perfect, was CiteSpace. It should be noted that this tool adheres to Shneiderman's rubric, scores fairly well in terms of usability and was designed more specifically with this kind of dataset in mind.

9 Conclusion

The information visualization approach holds promise as a possible alternative for summarization to text-based interfaces currently employed by information retrieval systems. More specifically within the domain of cited reference searching, visualization enables users to graph underlying trends in research, patterns and network relationships more clearly and efficiently. Among the challenges for successfully implementing systems are improving usability and matching the visualization type effectively to the summarization needs.

References

[1] E. H. Chi, "A taxonomy of visualization techniques using the data state reference model," *Proc. of Information Visualization 2000*, pp. 69-75, Oct. 2000.

[2] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," *Proc. 1996 IEEE, Visual Languages*, pp. 336-343, Sept. 1996.

[3] D. E. Irwin, "Ring species: Unusual demonstrations of speciation," *ActionBioscience*, http://www.actionbioscience.org/evolution/irwin.html. 2002.

[4] J. A. Coyne, "The genetic basis of Haldane's rule," *Nature*, no. 314, pp. 736-738, 1985.

Eric Dillalogue is a master's student at Drexel's College of Information Science and Technology.

Santhosh Krishnamurthy is a master's student at Drexel's College of Information Science and Technology.

Neil Vapiwala is a master's student at Drexel's College of Information Science and Technology.

Michael Zankowski is a master's student at Drexel's College of Information Science and Technology