LIS9721/LIS9821/CS9639 Information Visualization

Dr. Kamran Sedig

Fall 2015

**Information Visualization System Design:**

**Google Scholar**

Final Report

November 30, 2015

**Team Members:**

Tanner Bohn

Zhao (Joe) Lin

Lei Pu

Brent Davis

Sze Hang Lee (Liz)

**Information Visualization System Design: Google Scholar**

Final Report

**Executive Summary**

**0.5 page**

**Description of the problem**

Google Scholar is one of the most popular database used to locate scholarly articles. However, locating the useful articles could be a time consuming and mentally exhausting task. With numerous articles returned by Google’s search engine in response to any keyword query, it requires a lot of time and cognitive effort for users to sift through a long list of search results to get an overview of the results and to find the relevant and useful articles.

Researching a subject is not a trivial task. When trying to find academic research, a number of problems present themselves. One of the most obvious and prevalent problems is finding current, reliable information. As techniques in research change, or new technologies become available, or even just as old hypotheses are revised, the current information changes. This is not to say that looking at the current information is the solution either. New information has yet to be subject to widespread analysis and criticism from peers and may not be correct or useful. There are also metrics available that help decide which papers to look at, such as how many times a paper has been cited, which influence the papers to be examined. Another aspect of finding useful information is the availability of it. On almost any given subject that someone is interested in, there is a wide breadth of knowledge. There needs to be some way to sort the information provided so that relevant results are more easily accessible.

This is not a problem that has been left alone, either. There are various academic search tools available - Google Scholar, PubMed, PubChem, Microsoft Academic Search - that are designed to assist in this task. Each of them returns a large number of results for everything but the most specific search results. They can be of considerable assistance when the user has a sharp, focused idea of what they are looking for but can quickly overwhelm them and present a tide of seemingly identical results. Metrics are often included but are not presented in a way that makes them easily comparable. It is difficult to navigate the information presented in a meaningful way besides individually accessing the papers and reading their abstracts.

The problem then is finding relevant, useful and reliable information in a large mass of it. Our design does not aim to completely solve the problem, as proper analysis of any of these articles requires at least reading the abstract. The intention, however, is that we accelerate the task of finding the articles worth more than a casual glance.

The objective of our visualization tool is to address the problems identified in Google Scholar and to help users better comprehend search results and accomplish research goals more efficiently and effectively.

**Analysis of the information space (e.g., data set), representations, presentation techniques, and interactions that you have used in your system**

**1-2 page?**

Information Space:

It is safe to limit the scope of our information space to that of academic papers. Their contents are included in our domain but are beyond what is required for our visualization and are hence omitted. The central entity of our information space is the documents themselves. Other entities include the authors of the paper and the publishing journals.

A given paper has at least one and potentially many authors and typically exists in one journal. An author can have many properties, but we are not interested in all of them given the scope of this project. The relevant author properties for our project include all of the papers the author contributed to, their field of study, *their college/university*, and some abstract sense of ‘reliability’. Journals have only a few properties that are relevant: Field; Issue number; and Date.

Papers have a number of properties that are more important to us. The authors are a property of the paper, as is the number of citations a paper has received. Papers all have a publication date. Papers also have a field. There is also a title used to identify the paper and the citations that the paper itself makes to other research. Additionally, there are a few desirable properties for a paper that ‘exist’ in an abstract sense but are not so easily identified. The importance, reliability and impact of a paper are all things that are at the center of our problem but are not easily extracted from the infospace. These may be more entangled in the relationship the paper has with other entities than actual information.

Given that we have three entities with a set of properties, there are a corresponding set of relations between them. There is some sort of hierarchy present in our data, which comes from how a paper is produced by an author and then collected into a journal. So the final paper is a result of the author’s work, and a journal is a collection of papers. However, a journal is not the result of a single author, or a single set of authors from a singular paper. It would seem from the analysis of the information space that a journal is a logical space to begin an examination of a topic, but a few things prevent this. Before getting to that, there are some common properties amongst the entities. Each entity possesses a ‘field’ which it belongs to. This is a categorical clustering, but it is not a completely distinct one. A paper will often blend fields and is difficult to assign unless the authors have set one themselves. An author may also study more than one field, adding to the difficulty of their categorization. Thus, this is a necessary relationship to include but one to be careful with as it is an imperfect categorization.

[identify entities, properties and relationship]

Brent

Representations:

While it is almost impossible to show all of the information on one VR, we aim to show the most important information that can help users locate the relevant articles on the first VR. Then, a sub-VR is designed to provide additional information to the users.

Presentation Techniques:

Our design has applied the ADIVSeR (Sedig & Parsons, 2015) and EDIFICE-AP (Sedig & Parsons, 2013) frameworks with different visual variables, techniques and patterns to encode the information of the search results.

Coordinate pattern is used to encode the score of Web of Science and h-index with different colours. Token pattern located at the centre of the coordinate structure is used to represent each article. The radius (size) of the circle including the outer two rings represents Google rank of the article. Link pattern is used to represent the connection and relationship between articles in terms of citation. In addition, sparkline charts based on the fusion pattern is used at the bottom of the VR to indicate amount of articles over time. Another sparkline chart based on the fusion pattern is used on the left hand-side of the VR to indicate the citation count of the articles. Thus, this VR is (TK • CC??????? • CR • LK • FS)-based.

Interactions:

Brent

**Outline of the concepts, ideas, techniques, and IV systems that you have considered when designing your system**

**2-3 pages?**

This design is made with the consideration of the way how people conduct online search, their cognitive process, and what types of information are crucial to them.

Complex visualization display, such as three-dimensional graphs, has been considered. However, it may be difficult for users to understand and analyze the information.

During earlier stages in the design, VR patterns were considered and evaluated in isolation and in different combinations to identify the blending structure that could best represent article features. Our objective is to design a structure that can represent several crucial attributes of the entity. After assessing and evaluating the suitability of the drawings of various forms of blending of VR patterns, the VR ϵ TK • CC • CR • LK • FS was chosen in our design for several reasons. First, these patterns are simple and easy to understand. Second, they have the potential to incorporate various structure for representing different features. Most importantly, our design applied the circular theme of patterns with colour scheme of red, green, yellow and blue taken from Google’s logo to maintain the theme of Google Scholar. These colours are also easily distinguishable from each other.

**[Tanner’s initial design idea]**

**[Brent’s idea]**

**[Joe’s idea]**

**[Lei’s idea]**

**Your justifications (based on what has been studied in the course) for your design choices and decisions**

Our VR followed the visualization pipeline model to transform data into visualization views (Card, Mackinlay, & Shneiderman, 1999). Different visual variables, including position, size, hue, brightness, focus, and connection, have been used to represent information in our design to maximize the effectiveness of our VR. Particularly, size is used to represent Google rank of the articles. While the amount and size of the circles can be visually differentiated, a user can quickly tell whether an article in the search result has higher relevance than other articles.

Furthermore, our VR followed the visual information seeking mantra: “Overview first, zoom and filter, then details on demand” introduced by Shneiderman (1996). Our VR started from an overview of the data according to the search query of the users. Then, the users can zoom into the dense areas to explore different levels of detail and filter out irrelevant information according to their needs. When it is necessary, the users can also select a specific article and drill down for further details. In addition, our VR allows users to adjust their search queries by using dynamic querying and instantly receive feedback. Multiple filtering options can be combined to provide optimal result. Our VR communicates the filtering results by dimming the irrelevant articles, which allows the users to focus on the relevant information while maintaining the overall context.

Both stepped interaction and continuous interaction have been employed in our VR to help users obtain relevant information. An example of stepped interaction is to click the buttons of the dynamic querying functions to filter the search results. On the other hand, dragging the sliders to adjust dynamic querying is an example of continuous interaction.

The radar view is used to provide users an overview in the direction of navigation, which contains articles that are off-screen. This allows the users to quickly navigate to potentially relevant articles. Furthermore, the local-edge lens is designed to allow users to uncover and investigate a specific article within a cluttered area. By dimming the articles that are not connected to the selected one, the users can focus their attention on that particular article.

Moreover, our VR show articles in their frame of reference according to time. An adjustable slider located underneath the VR with sparkline chart embedded to show article count is designed to help users filter the articles based on their selected temporal range.

[EditLens????]

Although the list view of search results currently used by Google Scholar is easy to use and widely accepted, it does not convey as much information as quickly as graphical visualization. Therefore, our design incorporated the strengths of both list view and graphical visualization.

To help users quickly gain an overview of the articles in the search results, our VR first provides the results based on the search query to assist users in judging which articles are potentially relevant and whether they should spend more time and effort for further exploration at first glance. Additional information, such as relationship between articles, Web of Science and h-index scores,  \_\_\_\_\_\_\_\_\_ are presented to help users make a better judgment in selecting relevant articles with high quality.

Although some information may not be represented in our design, this is a reasonable choice to allow the users to identify the most important information with the first glance of the VR. Otherwise, the VR may become too complicated and cause cognitive overload to the user (Ives, 1982)

**Your final thoughts on what you think is missing and how your design can be improved.**

**0.5-1 page.**

Not started / Brent

The design relies on existing metrics (citation count, h-index, year of publication, etc. ) as a way to show the data. There is some insinuation of quality inherent in the design by doing this and an improved design would use better metrics. However, the problem of sorting papers by quality and relevance is not a new problem and well beyond the scope of our project. The selection of metrics alone is up for debate and could potentially be improved or customized to a user’s particular preferences.

Our glyph was designed with quick comparisons between other glyphs present on an axis in mind, but the sensitivity encoding was something that frequently came up in the design iterations for it. Any tuning of the design should be mindful of new potentials to allow the glyphs to encode information more faithfully.

Another potential thing that could come from our design is the examination of meta-data about the papers searched. Given that our design does not include its own ranking algorithm for results it was not something we desired to do. However, users looking into trends about the articles in the field may be interested in looking at the geographical distribution of publishing authors (complicated by how the authors for one paper can be spread out globally). Or, they may want to look for evidence of ‘citation clubs’ as discussed in class; a visualization tool using some of the same ideas here would be useful for identifying cliques in academia.

A final thought on what is missing is the application of this technique to search results as a whole, which is complicated by the lack of structure to the data in that information space. In the scholar results, we can reliably expect certain properties to be accessible and reasonable. If we try to extend our visualization generically to all of google’s search results we lose our ability to sort in any meaningful way. We cannot use the date that a page was published or updated because it may contain information from manuscripts predating most civilizations. A similar problem emerges with all of our other metrics - not many generic search results have an h-index - and prevents any kind of sorting in the way we have done here. Translating the kind of visualization we have done here to generic results would be helpful, but is not a trivial task.

**References**

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in information visualization: Using vision to think*. San Francisco, Calif: Morgan Kaufmann Publishers.

Ives, B. (1982). Graphical user interfaces for business information systems. *MIS Quart. 6*, 15-47.

Sedig, K., & Parsons, P. (2013). Interaction design for complex cognitive activities with visual representations: A pattern-based approach. *AIS Transactions on Human-Computer Interaction*, *5*(2), 84-133.

Sedig, K., & Parsons, P. (2015). Design of visual representations for human-information interaction: A pattern-based framework. Morgan & Claypool Publishers.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In Proceedings of the IEEE Symposium on Visual Languages (VL). *IEEE Computer Society*. 336–343.