

Visualizing Keyword Searches Using the Venn Diagram

John W. Burris

Southeastern Louisiana University, Hammond, LA, USA

Email: jburris@southeastern.edu

Abstract—This paper reports how the Venn diagram can be used to effectively explore and visualize search results using keywords of academic citations. By using a familiar and intuitive diagram for set relations, the results could provide not only the retrieval of a document, but also context for its relevance to the search. This work focuses on the searching of citation indexes due to availability of accurate and descriptive keywords. This work includes a proposal for the retrieval process as well as a prototype for searching using a limited dataset. This is a “work-in-progress” paper that will provide a limited implementation and a framework for evaluation.

Index Terms— visualization, information retrieval, search

I. INTRODUCTION

A comprehensive literature review is an essential part of the research process. Researchers must seek out as much relevant work as possible to ensure that their research is based on the complete current knowledge of the research topic as well as ensuring that the work has not been previously published. Researchers cannot be expected to maintain an exhaustive knowledge on the advancement of a topic by following the literature being published due to the sheer volume of information being published. With search engines, digital libraries, and citation indexes, the literature review has been reinvented and greatly simplified.

One common tool used for academic literature review is Google Scholar (<http://scholar.google.com>). This search engine-styled tool has been useful since its release in November 2004. Google Scholar works by indexing as many sources as possible and then providing ranked solutions to natural language search queries. Google Scholar is capable of indexing major digital libraries such as IEEE Xplore and ACM Digital Library as well as papers published on the web pages of individuals [1]. One possible limitation of Google Scholar is that it has a similar search interface as used by the Google web search engine. This interface makes the search process intuitive and simple. The simplicity makes such a tool meet the needs of a general audience, but not the needs of highly educated researchers who are knowledgeable of the search process. Another possible limitation of Google Scholar is that it does not appear to exploit the keywords used by authors and publications to classify the work.

Thomson Scientific’s Web of Science is a citation database that is designed to aid academic researchers in the search and identification of related scholarly work. Focusing on the citations allows the tool to exploit both the knowledge of the user (researcher) as well as the keywords used to describe the work. The interface to Web of Science relies on researchers to select the most relevant results. This interface could be extended to include features such as ranking and bibliometrics to increase the precision of search results. [2] Bibliometrics is used to analyze documents. An example of a simple bibliometric would be to say: “documents that share citations are more likely to be related”. The challenge of bibliometrics is integrating them into the search results in a way that does not complicate the search process or eliminates relevant results.

The use of the Venn Diagram in the search process is well established in previous work. VQuery is a graphical user interface for Boolean query specification. VQuery used the Venn Diagram to create the queries that would be used to search a digital library [3]. Although a usability study has not been published, the effectiveness of the diagram is well accepted and show to produce faster and more accurate results than boolean query syntax [4]. The Venn Diagram Interface (VDI) uses a similar approach to building queries (and subqueries) for general web search and showed a significant advantage in usability compared to Google’s standard search interface [5]. Approaches such as VQuery provide results that show strict relevance to the query, but do not provide a ranking of the results. WebSearchViz uses a similar visualization to provide a ranking of results. The visualization uses a single 2D circle (rather than a Venn diagram) with the center representing the point of most relevance and each result displayed at a distance away from the center that visualizes the relevance of the result [6].

This work proposes a method for using an intuitive selection interface that also provides a visualization for relevance when searching citation indexes for relevant publications. The initial search query is made up of multiple keywords or keyphrases. A Venn diagram is displayed and allows the researcher to select regions of the diagram to perform query refinement. The results of the query are displayed beneath the diagram and are analyzed by the researcher using the same diagram to determine relevance. This method hopes to increase the simplicity, recall, and precision of the search process by

exploiting the accuracy of keywords, including a selection interface for query refinement, and creating a relevance judgment that uses the same mathematical diagram used for query refinement. Fig. 1 shows the design concept of this method.

II. A MODEL FOR VISUALIZING SEARCH RESULTS USING THE VENN DIAGRAM

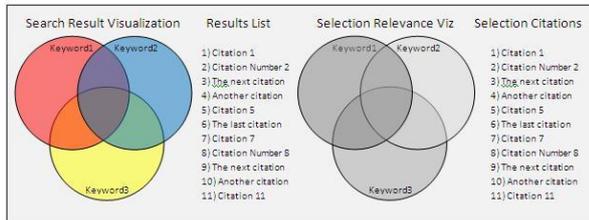


Figure 1. Design concept.

The query requires the user to select 2-4 keyword phrases from a list of phrases used in the database. The result is the first visualization and the results list.

The first visualization shows how the citations that are returned by the query are distributed among the keywords. The darkness or intensity of the color is an indication of the relative number of citations published that use that particular keyword. The intersections of the circles have their own darkness or intensity related to the relative number of citations that use the keywords represented by the intersecting circles. This diagram is selectable. Each circle and their respective intersections represent a boolean query, and as such, each area represents a different set of results.

Upon selection of an area in the diagram, a list of citations is displayed to the right of the visualization. This list is the set of citations that are described by the keywords for the selected region of the diagram. This list is also selectable. The selection of a citation generates a display of the next visualization.

The second visualization is a Venn diagram that analyzes the works cited within the selected citation. Each keyword used in the initial query is represented in the diagram and the darkness or intensity of each region is determined by the relative number of citations that are contained in that set. This diagram is selectable as well. Upon selection of a region, the final results list shows the citations included in that set. This list is not selectable.

III. IMPLEMENTATION

Fig. 1 shows the original concept, for which a limited prototype has been developed. The prototype was implemented in PHP with a MySQL data source and is shown in Fig. 2. Since the low level procedures were unimportant to this work, the indexing, storage, querying, and retrieval were left to existing MySQL capabilities. The system will be described in terms of the data, the Google API that was used, and the procedures for generating the results/visualizations.

The data source used was the Cognitive Science ePrint Archive (cogprints). This reason that this data source was used is its availability of keywords that accurately

describe the documents held in its library. It also has a limited number of documents that it indexes, allowing for better evaluation against the existing search system. The primary disadvantage of this data collection was that the references stored by the system are simply the text of the citations. This introduces a severe limitation to the analysis of the documents retrieved by the system since the keywords that would be used to describe these papers are unavailable. The data was extracted from the cogprints website in Microsoft Excel format and then converted into a format which was imported into a MySQL database.

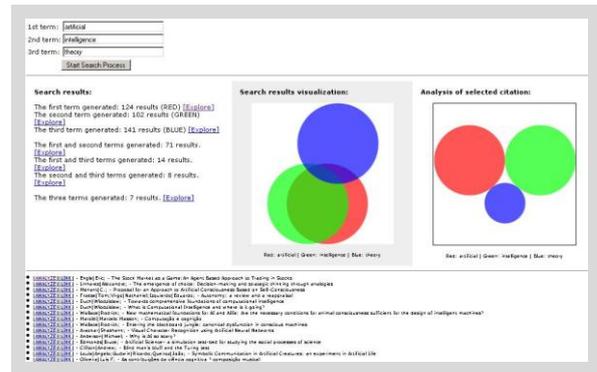


Figure 2. Design concept.

The main feature of the prototype is the querying and visualization. The novel part of this feature is not how the data is queried, but how the results are selected. Relevant queries were generated in PHP and executed using MySQL. As can be seen in Fig. 1, the HTML form has 3 text inputs that allow any strings to be entered and allows for the use of regular expressions as allowed by the SQL specification. (This is not mentioned in the interface because it may add unnecessary complexity to the search process.) Only one of these inputs is required. The query is generated by a PHP method and executed on the MySQL database.

The first set of results is the number of results to the search query for each combination of terms: Term A, Term B, Term C, Terms A AND B, Terms A AND C, Terms B AND C, and Terms A AND B AND C. Two critical parts of this list are the option to select/explore a particular result and the cardinality of the set as represented by the darkness of the corresponding section in the Venn diagram. The size of the list does adjust according to number of search terms. (See Fig. 3)

The prototype visualizations were created using the Google Charts API. The interface was chosen for usability and the provision of sufficient features for this project, but has limitations. The first limitation to using this API is significant. The Venn Diagrams produced by the API are not "clickable" meaning that the results of the query according to the sets that they belong must be displayed textually as well. The second limitation is that the labeling and naming features must be omitted as well.

In Fig. 2 and Fig. 3, the first generated visualization is called the "Search results visualization". The colors are consistent for the order used in the query. (The first term will always be red, second will be green, and third will be

blue.) The visualization is designed to be useful for the researcher to determine the number of publications that fall into each subcategory of research. The usefulness is dependent on how the searcher utilizes it. For instance, Fig. 1 shows terms 1 and 2 as “artificial” and “intelligence”. Note the high level of intersection between the sets shown. This gives the indication that the separation of the two search terms is unlikely to generate different results as well as indicating that the use of them as a single term will not significantly reduce the precision of the query.

The second list is the display of the selected query results. The results are ranked in this list. The ranking is determined by the number of times the keywords are used in the publication. The ranking is limited in this work since the search terms are limited to keywords. This is ineffective for ranking since it is unlikely that keywords will be repeated despite being more relevant to that particular keyword. However, in an ideal implementation, the search terms would be extended to all fields.

The second visualization is the analysis of a selected citation. In Fig. 1, this is labeled as “Selection Relevance Viz” and is located to right of the search results. It represents the number of times that the keywords are used in the “reference text” information for the selected citation. The cog prints data collection does not have unique identifiers for references, so the keywords must come from the text used in the title of the document and publication. Ideally, this would be the frequency of keywords used to describe each of the works cited. The Venn diagram does not overlap. This is also not ideal. It would be better to know if the related documents are likely to be related to all keywords or just subsets of those terms. The usefulness of this visualization is to determine if the selected result is heavily related to a particular search term.



Figure 3. Results for one and two search query terms.

IV. EVALUATION DESIGN

Only a limited, qualitative evaluation of this method has been performed. A more extensive evaluation is proposed. One test was proposed to measure the effectiveness of the citation evaluation mechanism, One test was designed to measure the overall quality of the system. No test was designed to measure the effectiveness of the query refinement, since similar systems have already been evaluated [5].

Test 1: An empirical study was designed to determine the total time spent to determine the documents returned were relevant. This would determine the usefulness and usability of the citation evaluation. The steps of the test are:

Step 1: Enter three one word search terms.

(Results should be identical)

Step 2: Determine and record if each result is relevant.

Step 3: The total times of the two relevancy evaluations are compared.

Step 4: The actual relevance of the source is compared to the determinations.

The subset of the cogprints data collection was selected as the best accessible data to base the prototype on. However, the number of citations included in the prototype is too small. A meaningful quantitative evaluation must have a larger index.

Test 2: A simple qualitative study used to gauge the overall usefulness of the system. A group of three researchers were asked to search for relevant documents based on predetermined search criteria. They were instructed to become familiar with the existing search interface for the cogprints data collection. The subjects were then recorded for 10 minutes evaluating both the cogprints textual interface and the prototype developed to visualize the results using the Venn diagram. The subjects were instructed to “think out loud” as they searched for publications related to their research. They were then told the prototype’s purpose was:

“to create an easier way to search for relevant citations by providing a mechanism for judging the relevance within the selection interface. It does this by allowing for query refinement and relevance judgment that uses the same mathematical diagram used for query refinement”

No further instruction was given. The following are representative excerpts from the qualitative evaluation:

- “Clunky interface, but has potential”
- “Big learning curve”
- “Far superior to the limited categories offered at the site.”
- “Liked the idea of exploring relevance without reading the abstract.”
- “Room for improvement: HCI”

The criticisms focused on how the user interacts with the system. Particularly, the meaning of the Venn diagrams is not immediately clear to a new user. However, the positive feedback when the subjects compared the prototype to the existing textual interface showed that the system may have promise.

V. CONCLUSION

This work has proposed a method for exploring and visualizing academic citations using the Venn diagram. This method aids researchers in finding relevant publications by using the same diagram for interactive query refinement and result evaluation. A limited prototype has been developed to demonstrate the concept. Initial qualitative evaluation is encouraging, but future work will focus on increasing the functionality of the prototype so a meaningful evaluation of the method can be done.

REFERENCES

- [1] M. Shultz, “Comparing test searches in PubMed and Google Scholar,” *Journal of the Medical Library Association*, vol. 95, no. 4, pp. 442–445, 2007.

- [2] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas, "Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses," *FASEB J.*, vol. 22, no. 2, pp. 338-42, 2008.
- [3] S. Jones, "VQuery: A graphical user interface for boolean query specification and dynamic result preview," *International Journal on Digital Libraries*, vol. 2, pp. 207-223, 2003.
- [4] M. A. Hearst, *Search User Interfaces*, 1st ed., New York, NY, USA: Cambridge University Press, 2009.
- [5] L. Langer and E. Frøkjær, "Improving web search transparency by using a Venn diagram interface," in *Proc. 5th Nordic Conference on Human-Computer Interaction: Building Bridges*, New York, NY, USA, 2008, pp. 249-256.
- [6] T. N. Nguyen and J. Zhang, "A novel visualization model for web search results," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 981-988, Sept.-Oct. 2006.



John W. Burris is an Assistant Professor of Computer Science at Southeastern Louisiana University. He received his B.S. in Computer Science from Louisiana Tech University in 2003 and his Ph.D. in Computer Science from Louisiana State University in 2012. His research areas include software engineering, networking and high performance computing.