Visualization of the Health Care Visualization Literature

Vivian L. West, PhD, MBA, RN (1); David Borland, PhD (2); David A. West, PhD (3); W. Ed Hammond, PhD (1)

(1) Duke University, Center for Health Informatics (2) Renaissance Computing Institute, The University of North Carolina at Chapel Hill (3) East Carolina University

Duke Center for Health Informatics

Abstract

Text mining has been used in a variety of applications to discover information in unstructured textual data. Information visualization can help users see things in data that would otherwise not be evident. We combine the two techniques to better understand what the healthcare visualization literature contains.

Background

Information visualization as a means to help users of very large data sets understand what is in the data is of increasing interest in health care researchers eager to use this big data to discover ways to improve health care outcomes. With any research project using visualization, researchers will seek knowledge from published research on its use. A literature search in MEDLINE or PubMed returns thousands of published articles, surprising since information visualization is a fairly recent approach in health care research.

Visualization, information visualization, and data visualization are not recognized terms in the lexicon of the National Library of Medicine's Medical Subject Headings (MeSH). Of the volume of articles returned in a literature search using any of these terms, the focus is quite diverse, leaving the researcher with the daunting task of manually reviewing the articles to select those of relevance. Text mining helps this process by using natural language processing (NLP) to provide the researcher with key words or terms in the unstructured data. With thousands of articles and the number of terms NLP identifies for each, the task is better but still overwhelming. Representing key words and terms visually can quickly help users identify the most frequently used terms, relationships, and correlations, and address the problem of information overload. The purpose of this research is to evaluate the effectiveness of using text mining and visualization to understand what has been reported in the health care visualization literature.

Methods

The PDF files of 70 articles previously retrieved from a systematic review of the health care visualization literature[1] were imported into SAS Text Miner. NLP was used to parse each document and its constituent words and terms. One of four different frequency weighting algorithms identified a set of key terms (number determined by user–from 2,000 to 5,000) that efficiently discriminate between the documents. Singular value decomposition converted the weighted term matrix into a numerical vector for each document. The document vector was used for distance calculations and the formation of clusters and to form four higher order themes for the document population. This process is as follows.

Text Parsing

- Identification of synonyms
- Multi-word terms
- Start / stop list: terms to include or exclude from analysis
- Verb and noun stemming, i.e. matrix, matrices; criteria, criterion

Text Filter converting words to number

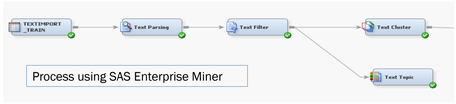
- Spell checking
- Identification of words/terms that discriminate between documents
- Result: document frequency matrix

Weighted term matrix

- Frequency that term *i* appears in document *j*
- Frequency that term *r* appears in document collection
- Number of documents in collection
- Number of document in which term *I* appears

Singular value decomposition (SVD)

- Similar to principle components analysis
- Reduces dimensionality of term document weight matrix
- Used to identify concepts and topics



Process using SAS Enterprise Miner

After converting text to numerical data, visualization techniques were applied to the terms and concepts.

Results

Text mining enabled us to identify the clusters and concepts from the health care visualization literature. By visualizing these clusters and concepts, we can determine those prevalent in the literature and those missing.

Using a DOCUMENT VIEW CIRLCE, **clusters** are colored arcs, with size proportional to the number of documents in that cluster. The **topics** are represented by the grey arcs, with size proportional to the number of documents with that topic. (Figure 1)

In the DOCUMENT VIEW SCATTER PLOT, **documents** are represented as pale circles colored by the document cluster, with size proportional to the number of topics. Documents belong to a single cluster but may have multiple topics (or no topics). The position is taken from the first two SVD components. The **cluster centers** are represent by colored donuts, with the radius proportional to the number of documents in the cluster. The **topic centers** are black donuts, with radius proportional to the number of documents with that topic. (Figure 1)

The TERM VIEW is another visualization of the same information. **Clusters** are represented by colored bars, with the height proportional to the number of documents in that cluster. **Topics** are grey bars, with size proportional to the number of documents with that topic. **Terms**, which are taken from the cluster description and topic terms, are drawn beneath or above each cluster/topic. (Figure 1)

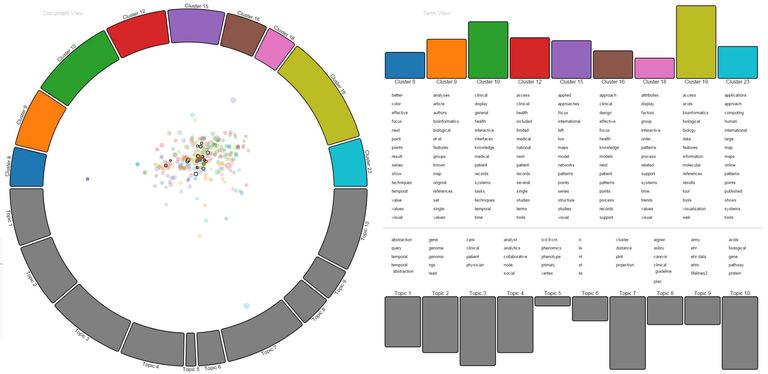


Figure 1. Document/Term View without any selection.

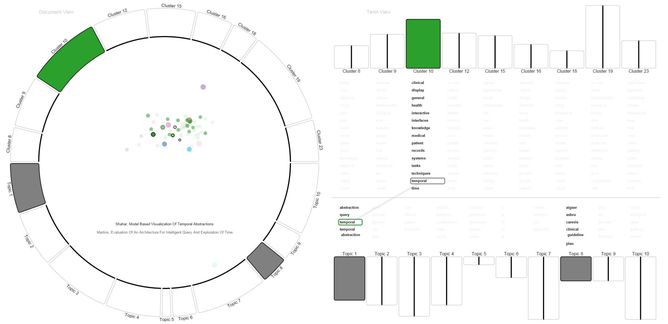


Figure 2. Authors: selected author (Shahar) and mouse over author (Martins). Both contain 'temporal' and are associated with Topics 1 and 8.

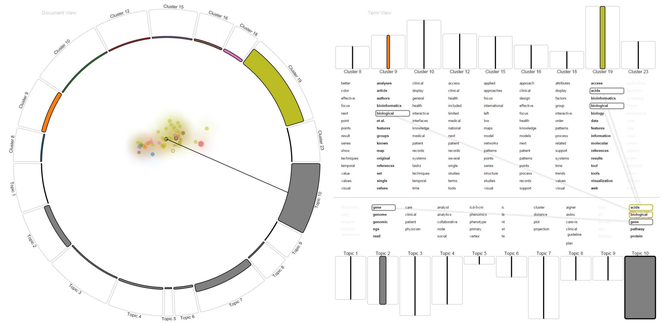


Figure 3. Selected Topic 10: shows all terms shared with other clusters/topics and all documents with that topic. Can also select three terms to show relationships.

Conclusions

In this pilot study, we used text mining with the unstructured data in many publications about information visualization in health care to identify the clusters and concepts from the health care visualization literature. Once converted to numerical data, we then visualized the results. This is a powerful way to understand what this literature includes. We found genetics to be an area in the visualization literature that has the greatest number of documents.

Visualization of the clusters and concepts enables us to determine those prevalent in the literature and those missing, and provides a method for understanding numerous terms and their relationship with each other It also helps to understand where further research is needed,

The visualization we used is two dimensional, which makes the data overlap and more difficult to see in the document view. With a three dimensional visualization, it would be easier to see results. We are now working with a much larger data set and will use three dimension visualization.

References

((1) West VL, Borland D, Hammond WE. Innovative information visualization of electronic health record data: A systematic review. *J Am Med Inform Assoc*. Published Online First: 23 Oct 2014. doi: 10.1136/amiajnl-2014-002955

Contact

Vivian West, PhD, MBA, RN
Associate Director, Operations
Duke Center for Health Informatics
Hock Plaza, 2424 Erwin Rd
Suite 9002, Room 9021
Durham, NC 27705
919.668.189
vivian.west@duke.edu