DECISION SCIENCES INSTITUTE

An Evaluation of Machine Learning Methods and Visualization of Results to Characterize Large Healthcare Document Collections

(Full Paper Submission)

Vivian L. West
Duke University, Center for Health Care Informatics
vivian.west@duke.edu

David Borland
Renaissance Computing Institute, The University of North Carolina Chapel Hill
borland@renci.org

David West
East Carolina University
westd@ecu.edu

W. Ed Hammond
Duke University, Center for Health Care Informatics
william.hammond@dm.duke.edu

ABSTRACT

This research is an exploratory analysis of the abilities of machine learning algorithms (namely text mining) and interactive visualization to analyze large collections of health care research documents. Preliminary results from the analysis of 391 documents describing research in health care information visualization are presented.

KEYWORDS: Healthcare, text mining, machine learning, data visualization

INTRODUCTION

In 2004 a presidential executive order, ‘Electronic Health Records (EHRs) for All Americans’, laid out tenets to improve the quality and efficiency of healthcare. One of its goals was to have accessible EHRs for most Americans within 10 years (Bush, 2004a; Bush, 2004b). In September 2009, years of research and policy work culminated in the Health Information Technology for Economic and Clinical Health Act (HITECH Act), allocating $19.2 billion in incentives to increase the use of EHRs by hospitals and health delivery practices. The Office of the National Coordinator (ONC) reports that at the end of 2014, 95% of acute care hospitals have EHRs that meet federal requirements for certification of EHR systems, and 75% of federal non-acute care hospitals have installed EHRs with core functionality (Charles et al, 2015).

With the burgeoning amount of electronic data, the field of biomedical informatics is also growing. In 1996 Ben Shneiderman aptly stated what remains true today: “Exploring information collections becomes increasingly difficult as the volume grows” (Shneiderman, 1996). There are calls for research using informatics tools and approaches to develop methods to extract and use this large amount of data, herein referred to as big data. With the significant potential for knowledge discovery, researchers are challenged to find innovative and effective ways to use the information from the rapidly growing big data now available from EHRs.

Graphical visualization is an effective tool that has been used since the later part of the 18th century to communicate information about data. A plethora of scales, shapes, and colors have been used with both small and large datasets rendered as visual diagrams such as bar charts, line graphs, scatterplots, and pie charts to reveal patterns leading to knowledge discovery. Industries such as finance, accounting, and the petroleum industry routinely use information visualization, defined as ”interactive, visual representations of abstract data to amplify cognition” (Card et al, 1999) using innovative approaches that account for both the volume and complexity of the data. In the healthcare field, however, applications of advanced visualization techniques to large and complex datasets are limited.

We are currently working on a Department of Defense funded research project entitled Novel Visualization of Large Health Related Data Sets, exploring visualization techniques using big data from EHRs to discover what the data contain. With any research project using visualization, researchers will seek knowledge from published research on its use. Accordingly, we completed a systematic review of the visualization literature in May-June 2013 using primarily PubMed, the most frequently used reference database in health and indexed using the National Library of Medicine (NLM) controlled vocabulary (Medical Subject Headings, or MeSH) and Web of Knowledge. Of the volume of articles returned in the literature search, the focus was quite diverse, leaving us with the daunting and time-consuming task of manually reviewing the articles to select those of relevance. As interest in and publication about visualization of health related information increases, a tool to easily identify the topics covered by the literature would make this task more manageable.

The purpose of this research is to evaluate the effectiveness of using text mining and visualization techniques to explore and understand what has been reported in the health care visualization literature. With many articles and the large number of terms machine learning identifies for each, the ability to discriminate and identify significant articles related to our interests is better but still overwhelming. We hypothesize that 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.

LITERATURE REVIEW

Text mining in the field of biomedical informatics has become of great interest to researchers (Chaussabel, 2004; Hur et al, 2009; Shatkay & Feldman, 2003; Labaer, 2003; De Bruijn & Marin, 2002).Text mining biomedical literature is the topic for the March 2015 issue of Methods, a journal that focuses on experimental biological and medical sciences (Navarro & Iratxeta (eds.), 2015). The issue describes a sample of the text mining methods used in the field today. The concept of generating hypotheses from potential links in various publications using text mining, or literature-based discovery, has been used by a number of biomedical researchers (Srinivasan, 2004) to examine such topics as drugs (Androniz et al, 2011; Agarwak & Searls, 2008; Shetty & Dalal, 2011; Bellis et al, 2011), viruses (Hu et al, 2005; De Chassey et al, 2008; Their et al, 2012), and genetics (Hu et al, 2005; Papanikolaou et al, 2014; Poos, 2014; Xiang et al, 2013; Jung et al, 2014).

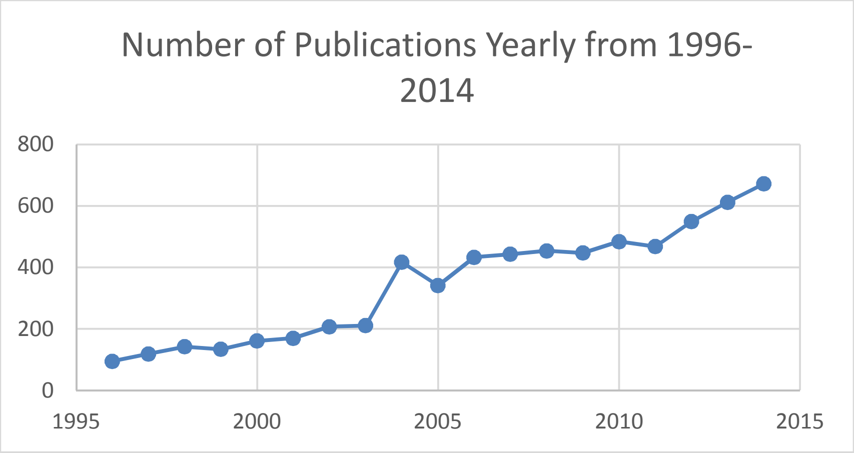
There are also examples of text mining used with scientific publications and visualization of results (Erten et al, 2004; Faisal et al, 2007; Fox et al, 2006; Synnestvedt et al, 2005). Stapley and Benoit constructed a prototype for genome information retrieval from 2,524 documents and visualization, linking terms of statistical significance (Stapley & Benoit, 2000). The authors state the graphical representation offered by their prototype provides researchers with the ability to intuitively assess the information presented and determine its value. A drawback they identified was that the volume of information represented by the combination of colors, graphics, and codes allowed the user to see the structure of the links but not the content. Allendoefer et.al. (Allendoerfer et al, 2005) use bibliometric analysis to create a visualization using nodes and clusters of networks showing similar nodes. They state that the layers created by their visualization have the potential to hide data, particularly if the user is not familiar with the database used. Andronis et al. (Andronis et al, 2011) describe several studies using ontologies with literature mining and visualization of results (primarily heat maps and graphs) for drug repurposing applications (new uses for existing drugs). They conclude that biomedical literature mining is an effective technique to generate hypotheses for drug repositioning, and clustering algorithms to bring similar concepts visually together is an effective method to provide additional insight into potential discoveries.

We could find no reports of text mining of complete articles as part of the visualization literature. Nunes et al. describe BeCAS, the Biomedical Concept Annotation System, developed for biomedical concept identification of PubMed abstracts that link to the reference databases (Nunes et al, 2013). In the systematic literature review we conducted in 2013, we found many concepts in the literature. We eliminated the vast majority of the articles from our final set based on information from the abstracts. With the final set, we read each article; 47% of the final set were eliminated (West et al, 2014). This raises the question of the value of the articles we initially rated based on the abstract: Would our results have been different had we had the means to more closely evaluate the complete text of each? Can text mining be used to easily identify the relevant terms in the literature, and would visualization of the results from text mining more accurately define the visualization literature?

METHODS

We completed a systematic review of the literature in May-June 2013 using PubMed and Web of Knowledge (West et al, 2014) as part of our research evaluating visualization of large health related data sets. The PubMed review was supplemented using citation searching and gray literature searching. Reference lists from highly relevant articles were also reviewed to find additional articles. We restricted our literature search to articles published since 1996. A query using PubMed for “information visualization” returned 6,559 articles, surprising since information visualization is a fairly recent approach in health care research and not part of the lexicon for PubMed. As Figure 1 shows, the number of articles on information visualization has grown significantly since 1996.

Figure 1. The number of publications in a PubMed search using the term “information visualization”. Adapted from Results by Year: http://www.ncbi.nlm.nih.gov/pubmed/?term=information+visualization.



In conducting the literature search, we wanted to find articles about the use of EHR data using innovative visualization techniques or describing techniques that could be applied to EHR data. We define EHR data as data in electronic clinical records that contain clinical information (e.g., demographics, problems, treatments, procedures, medications, labs, images, providers) collected over time that can be shared among all authorized care providers. We define innovative visualizations as visualizations other than standard graphs traditionally used for displaying health care information (e.g., bar charts, pie charts, or line graphs) that use complex data, which we define as data with multiple types of variables and many data points resulting in an exceptionally large amount of data, such as that in an entire EHR system. Interaction with the data is a key characteristic of information visualization, e.g. zooming or mouse-over to show features of the visualization. We were interested in articles describing any innovative visualization techniques for vast amounts of information that might be the foundation for an interactive system, however, so also included articles describing static visual representations of large amounts of EHR data. Articles were excluded if they related to animals or plants, were position papers describing the need for visualizing data or ideas for techniques in visualization, or did not describe specific techniques used for the visualization or have figures showing the results from visualization (West et al, 2014).

A total of 847 references were finally retrieved from our initial search of PubMed and Web of Knowledge. After a search of the gray literature and hand-searching references from articles, an additional 44 papers resulted in 891 articles to begin the review with. Using the abstracts from the articles, we found the literature replete with articles on visualization in genetics, syndromic surveillance, and geospatial environmentally-aware data. There were many articles on the technical details related to visualization techniques. We excluded 666 articles because the visualizations discussed were diagnostic, did not relate to EHR data, focused on animals or plants, used genomics data, discussed geospatial data or syndromic surveillance, were position papers suggesting the need for visualization or describing a potential visualization technique, or were primarily discussions of the technical details of visualization (West et al, 2014).

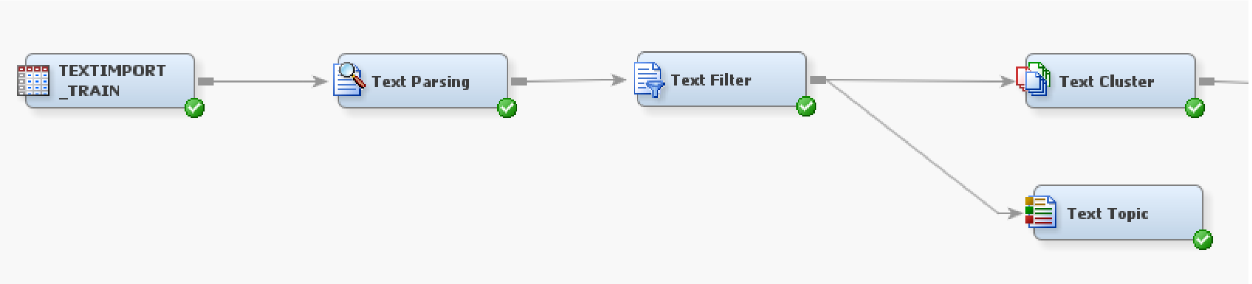
Sample

To assure we were familiar with the articles for this research using text mining, we drew from the literature we knew would be most similar to our interests in interactive health care information visualization. For our prototype and the visualization results we discuss later, we used the PDF files of 70 articles; this group of articles was included in the final abstract review for our systematic literature review in 2013. Our second experiment includes an additional 321 relevant articles, or a total of 391 articles. This combined group of 391 documents represents publications on visualization methodologies, tools, and applications to EHR data. The PDF of each document was used for this research. For our final exploration, we plan to use all 891 articles that were included in our systematic literature review conducted in 2013.

Text Mining

SAS Enterprise miner is the primary software used to process the documents. The text mining functionality of this software includes the ability to convert a collection of Adobe pdf documents into a SAS data set and then to parse, filter, and analyze the textual data, grouping similar documents as clusters and similar terms within the documents as topics. The experimental process sequence is shown in Figure 2.

Figure 2. Experimental Methodology Process Flow. Source: SAS.



The first node in Figure 2 is the conversion of the documents to a SAS data set. The SAS data set captures several properties (size, file location, etc.) of the document, and the document text is represented in a single variable by one long string.

The parsing node performs a number of processing functions on the text variable to identify a collection of terms in each document. Terms can be a single word or a multi-word expression. The single word terms can have parent-child relationships that account for synonyms and stemming. For example, stemming would establish a single term “zoom” to include the following: zooms, zooming, and zoomed. Synonyms can be identified from a database of predefined synonyms or by a user defined set. For example teach, instruct, educate, train could be a single document term. The basic concept is that parsing converts the string of text into a collection of terms that can be expressed as a term-by-document frequency matrix (see Table 1 for a simple generic example). The entries in this matrix are the raw counts of the number of occurrences of each term in each document. This data structure (with some refinement) is the basis for determining the relative degree of document similarity.

Table 1. Example of terms-by-document matrix: Adapted from Source: SAS.

TERM	DOC 1	DOC 2	DOC 3	DOC 4	DOC 5	DOC 6	DOC 7	DOC 8	DOC 9	DOC 10
color	4	5	1	3	2	1	0	3	2	1
heat map	2	3	0	1	0	2	0	1	0	0
graphs	1	2	2	0	2	3	1	0	1	2
information	0	2	4	2	3	1	2	4	1	2
medical	1	2	0	1	3	2	1	0	2	1
chart	1	2	0	0	1	0	0	2	2	0
modeling	0	1	0	0	1	0	1	1	0	0
computer	1	2	0	1	3	0	0	1	2	2
interface	2	2	3	1	1	3	2	1	0	0
geographic	0	1	2	1	0	1	0	2	1	1
genes	1	0	2	1	3	2	1	0	3	1
phenotype	4	0	0	1	1	3	0	1	1	1

The matrix generated in this research is large and sparse with 20,000 terms (rows) and 391 documents (columns). There is an obvious need to reduce the number of terms and the dimensionality of the matrix prior to analysis. This is accomplished by the text filter. The filter node eliminates very common terms like "the", the most frequently used word in the English language, and also eliminates terms rarely used in the document collection.

The text filter node enhances the information content of the raw term by a document frequency matrix, first weighting the raw frequency counts to reduce the effect of frequently occurring terms and then weighing the resulting terms to establish their ability to discriminate between documents in the collection being analyzed. The weighting function used to transform the raw frequency counts is a log function defined in Equation 1. The raw frequency counts, *fij* are the total number of occurrences of term i in document j.

$$g(f_{ij}) = \log_2(f_{ij} + 1) \quad (1)$$

The log weighting function dampens but does not eliminate the effect of terms that occur many times in a document. Alternatively, a binary function can be used to completely eliminate the effect of multiple term occurrences. The resulting matrix would then have an entry of 1 if there are one or more occurrences of the term and 0 if there are no occurrences.

The terms that are most effective at categorizing documents in a collection are those terms that occur in only a few documents but many times in those documents. The term weights are designed to identify these terms. There are two term weighting algorithms that are relevant for this research, entropy and inverse document frequency. Entropy is an information theory construct defined as follows where the entropy weight for term i, *wi*, is a function of the frequency, *fij,* the number of times the term i occurs in the document collection, *gi*, and *n*, the number of documents in the collection.

$$w_i = 1 + \sum_j \frac{\left(f_{ij/g_i}\right)*log_2(f_{ij})/g_i}{log_2(n)} \quad (2)$$

Inverse document frequency weights are determined as follows where *P(ti)* is the proportion of documents that have the term i.

$$w_i = log_2\left(\frac{1}{P(t_i)}\right) + 1 \quad (3)$$

A weighted term-by-document frequency matrix is calculated by first applying the frequency weighting function to the raw frequencies (Table 1) and then scaling that result by the term weighting function. Table 2 shows the transformation of the matrix in Table 1 by applying the log frequency weight function and the entropy term weight function.

Table 2. Term-Document Frequency from Text Miner. Term Weight=Entropy, Frequency Weight=Log. Adapted from Source: SAS.

Term	**DOC 1**	**DOC 2**	**DOC 3**	**DOC 4**	**DOC 5**	**DOC 6**	**DOC 7**	**DOC 8**	**DOC 9**	**DOC 10**
color	0.271	0.301	0.116	0.116	0.116	0.116	0.116	0.116	0.116	0.116
heat map	0.114	03447	0.265	0.265	0.296	0.114	0.114	0.1144	0.114	0.114
graphs	0.046	03079	0.079	0.079	0.079	0.183	0.183	0.183	0.125	0.125
information	0.000	03502	0.000	0.202	0.202	0.320	0.320	0.202	0.523	0.470
medical	0.000	0.000	0.397	0.397	0.000	0.000	0.000	0.397	0.000	0.000
chart	0.522	0.000	0.522	0.000	0.000	0.000	0.000	0.000	0.000	0.522
modeling	0.000	0.522	0.000	1.522	0.000	0.000	0.000	0.000	0.522	0.000
computer	0.000	0.000	0.522	0.000	0.000	0.522	0.000	0.522	0.000	0.000
interface	0.000	0.000	0.000	0.000	0.522	0.000	0.000	0.528	0.528	0.000
geographic	0.723	0.000	0.000	0.000	0.000	0.000	1.146	0.000	0.000	0.000
genes	0.698	0.000	0.000	0.000	0.000	0.000	0.698	0.000	0.000	0.000
phenotype	0.000	0.000	0.000	0.698	0.698	0.000	0.000	0.000	0.000	0.000

Higher term weights are associated with terms that are more effective at categorizing documents. A cutoff value for the term weight is used to eliminate ineffective terms and reduce the dimensionality of the weighted term-by-document matrix. The analysis of the document collection begins with cluster analysis using hierarchical or expectation maximization algorithms. In preparation for cluster analysis, Singular Value Decomposition (SVD) is used to reduce the dimensionality of the input data (Trefethen et al, 1997). SVD is similar to principal components analysis and factors the weighted term-by-document frequency matrix into two orthogonal matrices U and V and a diagonal matrix ∑.

$$\mathbf{A = U\sum V} \tag{4}$$

SVD calculates only the first k columns of these three new matrices. Higher values of k explain more of the variance in the document collection. K must be large enough to capture the meaning of the document collection but not so large that it captures the noise. The projection of the original weighted term-by document matrix is projected onto the first k columns of U. Each row (term) of the term-by-document matrix is projected onto the first k columns of V.

The text topics and text clusters nodes characterize the text document collection. Each and every document is assigned to one and only one text cluster base on a distance metric calculated by Wards algorithm.

Experimental Design

Using our prototype with 70 documents, our second experiment using 391 documents, and third experiment using 897 documents, four document analyses per experiment explore the capabilities of machine learning methods to process and categorize documents. We will characterize the differences in results between the two term weighting algorithms: (1) entropy and (2) inverse document frequency. Each is applied to two different sets of document terms: (A) those selected by the machine algorithm from natural language and (B) a dictionary of terms defined by experts in health care visualization. Results from the four analyses (1A, 1B, 2A, and 2B) will each generate a complete set of results that include cluster assignments and document collection topics. These results will be analyzed using evaluation methods and a quantitative assessment.

An evaluation of the clusters and topics generated will be conducted using a panel of three experts in health care visualization, people who are the likely users of the technique we propose. A questionnaire to guide the panel of experts' feedback will be used to evaluate the panel's perceptions and opinions on the adequacy, effectiveness, and usability of results from each of the four cases. Results will be correlated and single metric measures like the Rand index or F-measures will be used to score the four experiments.

These results will be assessed quantitatively by a random subset removal technique. This consists of five trials where a random subset of the documents is removed from the data set. The analysis is then re-run for all four experiments and the percentage of documents in each cluster is contrasted in the before and after subset removal cases. If the clusters fundamentally represent the information in the document collection, the subset removal should not create major changes in cluster size or composition.

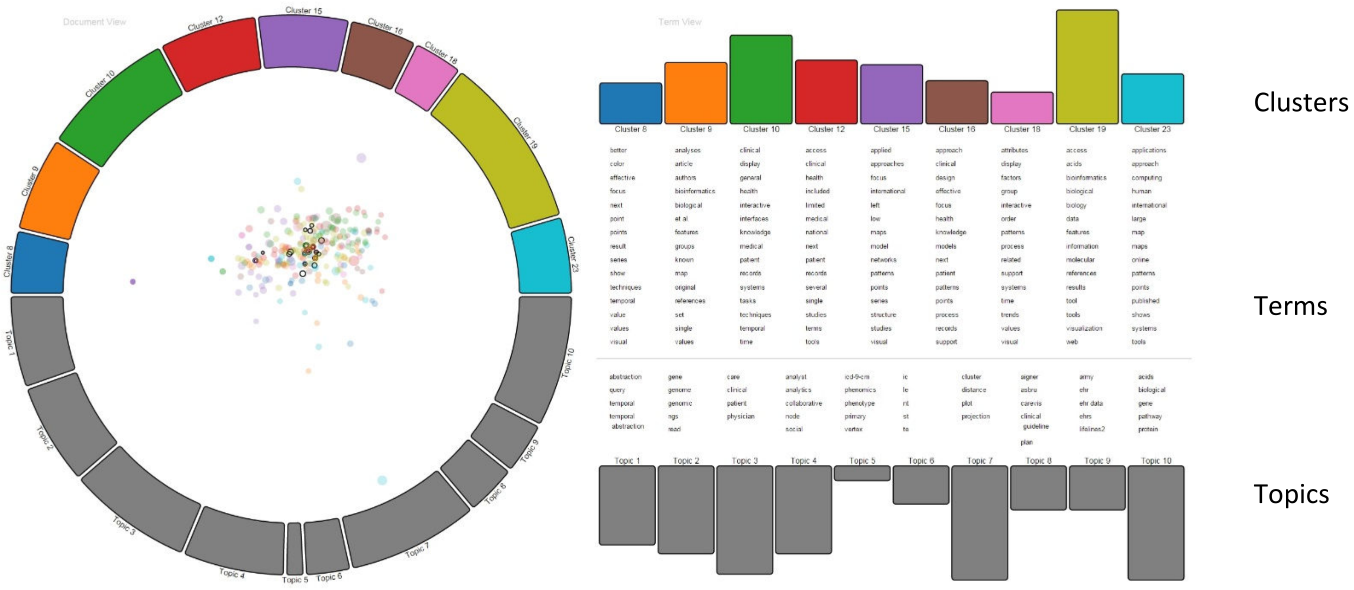
RESULTS AND DISCUSSION

This research is in progress, therefore our results section is not complete. The findings below include (1) the qualitative analysis of 70 documents and application of the findings for the prototype interactive visualization and (2) the quantitative analysis of the our second experiment using 391 documents .The expert panel evaluation of the and interactive visualization will be complete by the time of the DSI annual meeting.

Prototype and Interactive Visualization

For our prototype, we included 70 documents from the health care visualization literature as described earlier. Using results of the qualitative analysis as previously described, text was converted to numerical data using entropy. We have developed an interactive visualization tool using the D3 (Bostock et al., 2011) JavaScript library to display the resulting clusters, terms, and topics. Results of the visualization are shown in Figure 3. By visualizing these clusters and topics, we can evaluate those prevalent in the literature and those missing.

Figure 3. Document View (left) and Term View (right) linked visualizations of clusters and topics in the health care visualization literature.



In the Document View circle on the left, the top half are clusters represented by colored arcs; the size of each arc is proportional to the number of documents in the cluster. Gray arcs in the bottom half of the Document View circle are topics, with size proportional to number of documents with that topic. Inside the Document View circle is a scatter plot. Documents are displayed as pale-colored circles, with each color a cluster and the size of the circle proportional to the number of topics. Documents belong to a single cluster but may have multiple topics (or no topics). Colored rings are cluster centers with the radius proportional to the number of documents in the cluster. Black rings are topic centers with the radius proportional to the number of documents with that topic. In the Term View to the right of the Document View, the colored bars on the top are clusters, with height proportional to the number of documents in that cluster. The gray bars on the bottom are topics, with height proportional to the number of documents with that topic. Terms, taken from cluster descriptions and topic terms are noted in text beneath/above each cluster/topic. It is easy to determine that Topic 7 and 10 have the greatest number of documents, and Topic 5 has the least. Cluster 19 has the most documents and Cluster 18 has the fewest.

Interaction with the visualizations provides the user with a quick and easy way to gain a great deal of information about the literature. The user can explore the literature about the topics and articles published to date by mouse clicking various sections of the visualizations. It is possible to identify closely related clusters themes identified by the topic, their relationship to each other, the clusters and topics shared with common terms, the terms that comprise each cluster of documents, and the various authors of associated documents (Figures 4 and 5). This is a powerful way to determine like topics, the authors associated with a particular cluster who have published on specific topics, terms that might be used in a more targeted literature search on a given topic, and what topics have the greatest or least number of publications. The results from visualization of the output from machine learning applied to documents has the potential to become a useful way to determine what the literature contains, identify publications that might be similar to those the user is already familiar with, and identify like researchers in a given field and their area of concentration.

Figure 4. Mouse-over of Topic 10 (far right) displays closely related cluster and topic themes (terms), denoted by the box around each.

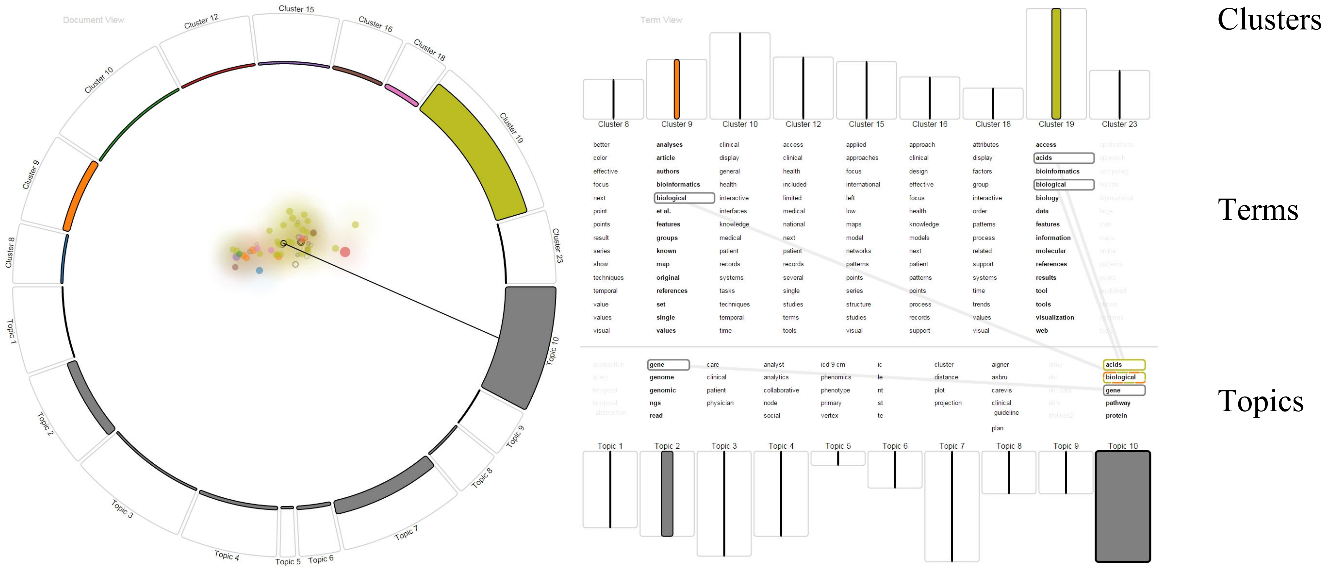
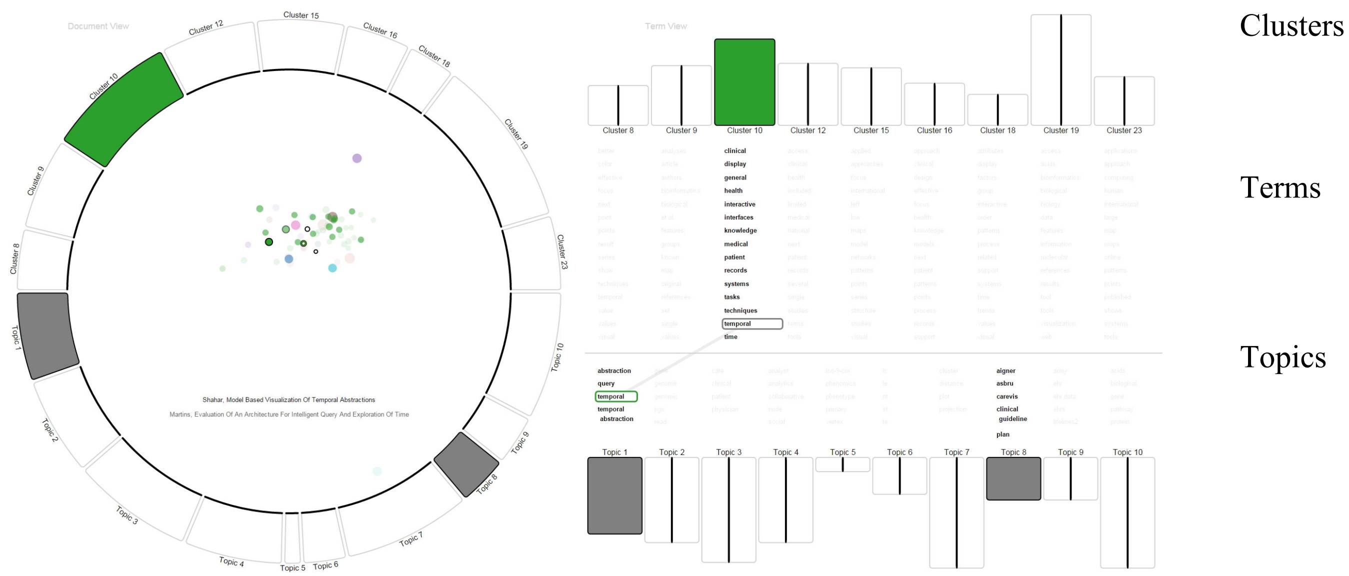


Figure 5. Two authors and their publication titles are displayed in the inside bottom of the Document View circle when highlighting a cluster that also displays a related topic and terms.



Document Collection Analysis

Our second experiment includes 391 documents. The output for one of our four analyses is described below. This analysis uses log weighting for terms, the inverse document frequency algorithm, and a set of 373 terms specific to the field of healthcare visualization developed by experts in the field.

Singular value decomposition transforms the document collection from text into a vector of numbers that can be analyzed by quantitative algorithms. Documents are clustered into 11 groups based on the distance between the documents. The results of Ward's hierarchical clustering algorithm are shown in Table 3. The (+) sign in front of some of some of the cluster description terms indicates the presence of synonyms or stem terms. The number of documents in each cluster is listed in the Frequency column and shown graphically in the bar chart of Figure 4.

Table 3. Clusters of similar documents.

Cluster ID	CLUSTER DESCRIPTION TERM	Frequency
9	fisheye spatial metadata 'computer graphics' +shape 'information visualization' visualization information visualizations +mapping multidimensional +focus +size interactive +cost	44
10	geographic spatial environmental +frequency space +shape +complexity +size times +space +evaluation novel computational +focus +overview	32
11	mappings environmental geographic 'user interfaces' system spatial multidimensional +flow +mapping visualizing phenotypes +phenotype costs times 'time-oriented clinical data'	7
13	genomics phenotypes +phenotype large-scale +visualization novel +information +system +scale +size technologies +complexity systems environmental times	34
19	'time-oriented data' temporal intelligent multidimensional 'information visualization' spatial visualization interactive visualizing 'data mining' 'data analysis' +aggregation 'time-oriented clinical data' +complexity information	15
20	genomics metadata computational +shape multidimensional large-scale +zoom 'data visualization' +scale +size +mapping 'data analysis' +visualization +space	34
22	'time series' +zoom +size 'data mining' +technology large-scale +frequency 'data analysis' +space +visualization +cost +information +scale visualizations +overview	17
24	genomics novel computational +mapping multidimensional +visualization programming +information large-scale +size 'data analysis' +space technologies +complexity 'data mining'	83
25	'knowledge base' intelligent 'time-oriented clinical data' temporal 'time-oriented data' +aggregation 'medical informatics' 'data mining' evaluation 'information visualization' 'time series' knowledge multidimensional computational spatial	28
26	temporal healthcare 'health care' 'medical informatics' +timeline 'decision making' +cost intelligent time +evaluation information +overview times +system 'information visualization'	75
28	evaluation 'data analysis' novel costs technologies +cost design +flow effectiveness +technology 'information visualization' environmental geographic visualization knowledge	22

Figure 6. The number of documents within Document clusters labeled with Cluster ID Number

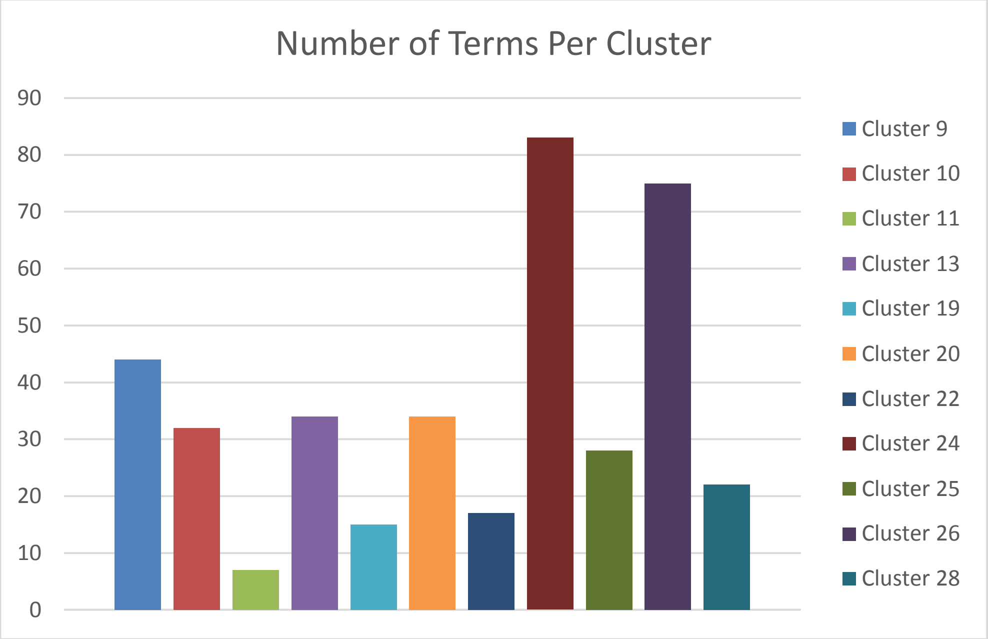
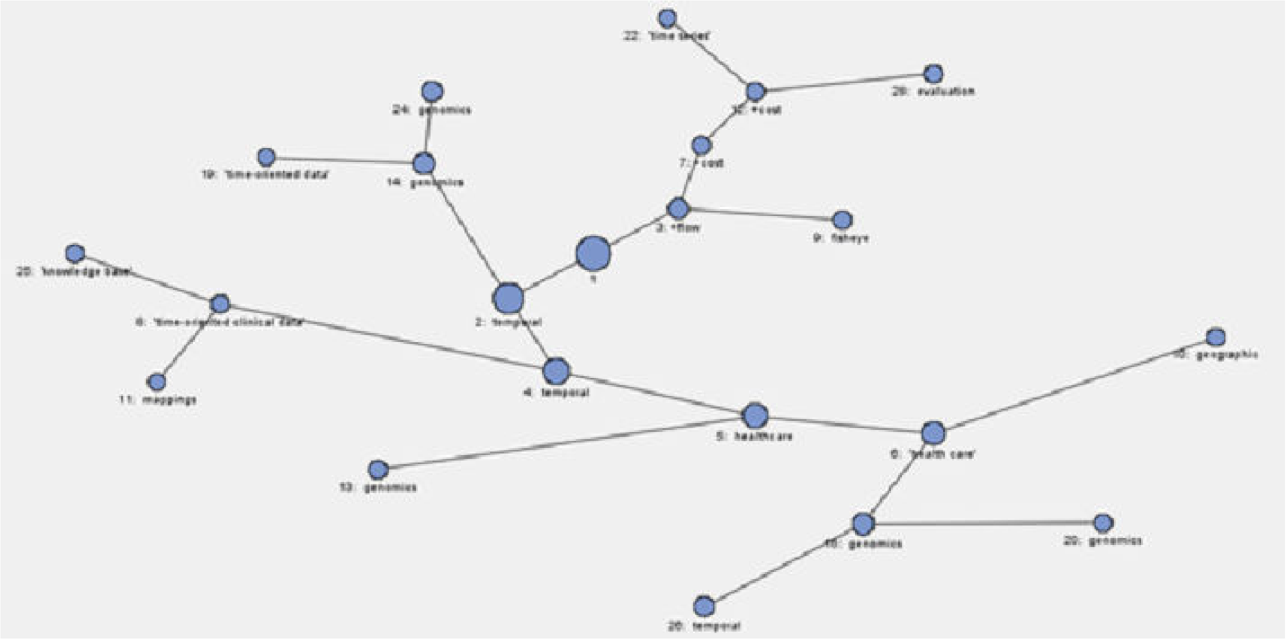


Figure 5 shows the hierarchical development of the document clusters. The largest circle, Cluster 1, is the entire collection of 391 documents. The collection is first split into two groups. The larger, identified as Cluster 2, is to the left and consists of 308 documents distinguished by the labels temporal visualization and genomics. The smaller cluster to the right of Cluster 1 is Cluster 3 and has 83 documents identified as flow and spatial visualization. Successive levels of cluster formation lead to the terminal clusters in Table 3.

Figure 7. Ward's cluster hierarchy.



The text mining algorithm also extracts topics from the document collection. Topics are common themes that occur as subsets of the document collection. The topics identified by the text mining algorithm for this experiment are summarized in Table 4.

Table 4. Document collection topics

Topic #	Topics				
1	"+visualization	+information	+system	+time	+technology"
2	"time-oriented data	intelligent	knowledge base	time-oriented clinical data	semantics"
3	"information visualization	design	+cognition	visualization and computer graphics	visualization"
4	"healthcare	healthcare	healthcare	health care	biomedical informatics"
5	"geographic	spatial	environmental	environmental	spatial"
6	"event sequences	visual analytics	EHR	visualization and computer graphics	+frequency"
7	"multidimensional	data visualization	+dimensionality	+mapping	+space"
8	"genomics	genomics	heatmaps	visual analysis	genes"
9	"metadata	biomedical informatics	metadata	+saturation	focus"
10	"+phenotype	+aggregation	genomics	intelligent	+dimensionality"
11	"time series	time series data	information visualization	dynamic query	time"
12	"fisheye	+zoom	+focus	zoom	fisheye"
13	"temporal	temporal	temporal	+timeline	artificial intelligence"
14	"environmental	+saturation	+flow	simultaneous	geographic"
15	"EHR	medical informatics	electronic health record	+timeline	EHR systems"
16	"imaging	medical imaging	spatial	grayscale	+flow"
17	"+precision	precision	+saturation	+vision	+frequency"
18	"+cost	health care	information technology	healthcare	+technology"
19	"exploratory data	data analysis	novel	programming	knowledge"
20	"medical informatics	decision making	decision support	EHR	artificial intelligence"

Results of document analyses will be presented to a panel of experts to assess the relative effectiveness of machine-based text mining analysis of research documents. Once there is concurrence on the effectiveness of the clusters and terms, we plan to conduct a third experiment using all 891 documents that comprised the initial set of documents from our 2013 literature review. These results will then be visually represented following the output of our prototype.

ACKNOWLEDGEMENTS
The literature review described in this paper was supported with funding from the US Army Medical Research and Material Command (USAMRMC), grant number W81XWH-13-1-006.

REFERENCES

Agarwal, P., & Searls, D.B. (2008). Literature mining in support of drug discovery. *Briefings in bioinformatics*, *9*(6), 479-492.

Allendoerfer, K., Aluker, S., Panjwani, G., Proctor, J., Sturtz, D., Vukovic, M., & Chen C. (2005, October). Adapting the cognitive walkthrough method to assess the usability of a knowledge domain visualization. In Information Visualization, 2005. INFOVIS 2005. IEEE Symposium, 195-202.

Andronis C., Sharma A., Virvilis V., Deftereos S., & Persidis A. (2011). Literature mining, ontologies and information visualization for drug repurposing. *Briefings in bioinformatics*, *12*(4), 357-368.

Bellis L., Akhtar R., AlLazikani B., Atkinson F., Bento A.P., Chambers J., & Overington J. (2011). Collation and data-mining of literature bioactivity data for drug discovery. *Biochemical Society Transactions*, *39*(5), 1365.

Bostock, M., Ogievetsky, V., & Hear, J. (2011). Data-driven documents*. IEEE Transactions on Visualization and Computer Graphics, 17*(12), 2301-2309.

Bush G.W. (2004a) State of the Union Address, Promoting Innovation and Competitiveness, President Bush's Technology Agenda.

Bush G.W. (2004b) Office of the Press Secretary, the White House. Executive Order: Incentives for the Use of Health Information Technology and Establishing the Position of the National Health Information Technology Coordinator. Press release, April 27, 2004. http://www.whitehouse.gov/news/releases/2004/04/print/20040427-4.html2004 (accessed 23 Jul 2013).

Card S.K., Mackinley J.D., & Shneiderman B., eds. (1999) Readings in information visualization: using vision to think. Morgan Kauffman.

Charles D., Gabriel M.; & Searcy T., (2015). Adoption of Electronic Health Record Systems among U.S. NonFederal Acute Care Hospitals: 2008-2014. ONC Data Brief No. 23, April 2015. Accessed April 27, 2015 at http://healthit.gov/sites/default/files/data-brief/2014HospitalAdoptionDataBrief.pdf,

Chaussabel D. (2004). Biomedical Literature Mining. *American Journal of Pharmacogenomics*, *4*(6), 383-393.

De Bruijn B., & Martin J. (2002). Getting to the (c)ore of knowledge: mining biomedical literature. *International journal of medical informatics*, *67*(1), 7-18.

De Chassey B., Navratil V., Tafforeau L., Hiet M.S., Aublin-Gex A., Agaugue S., & Lotteau V. (2008). Hepatitis C virus infection protein network. *Molecular systems biology*, *4*(1).

Erten C., Harding P.J., Kobourov S.G., Wampler K., & Yee G. (2004). Exploring the computing literature using temporal graph visualization. *Electronic Imaging 2004, International Society for Optics and Photonics*, June, 45-56.

Faisal S., Cairns P., & Blandford A. (2007). Building for Users not for Experts: Designing a Visualization of the Literature Domain. *Information Visualization, 11th International IEEE Conference*, July, 707-712.

Fox E.A., Neves F.D., Yu X., Shen R., Kim S., & Fan W. (2006). Exploring the computing literature with visualization and stepping stones & pathways. *Communications of the ACM*, *49*(4), 52-58.

Hu X., Yoo I., Rumm P.,& Atwood M. (2005). Mining candidate viruses as potential bio-terrorism weapons from biomedical literature. In *Intelligence and Security Informatics*, Springer Berlin Heidelberg, 60-71.

Hur J., Schuyler A.D., & Feldman E.L. (2009). SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. *Bioinformatics*, *25*(6), 838-840.

Jung J.Y., DeLuca T.F., Nelson T.H., & Wall D.P. (2014). A literature search tool for intelligent extraction of disease-associated genes. *Journal of the American Medical Informatics Association*, *21*(3), 399-405.

Labaer J. (2003). Mining the literature and large datasets. *Nature Biotechnology*, *21*(9), 976-977.

Navarro,M.A., & Iratxeta, C.P. (2015). Text mining of biomedical literature. *Methods*, *74*, March, 1-106.

Nunes T., Campos D., Matos S., & Oliveira J.L. (2013). BeCAS: biomedical concept recognition services and visualization. *Bioinformatics, 29*(15), 1915-1916.

Papanikolaou N., Pavlopoulos G.A., Pafilis E., Theodosiou T., Schneider, R., Satagopam V.P., & Iliopoulos I. (2014). BioTextQuest+: a knowledge integration platform for literature mining and concept discovery. *Bioinformatics*,*30*(22), 3249-3256.

Poos K., Smida J., Nathrath M., Maugg D., Baumhoer D., Neumann A., & Korsching E. (2014). Structuring osteosarcoma knowledge: an osteosarcoma-gene association database based on literature mining and manual annotation. *Database, 2014*, 1-9.

Shatkay H., & Feldman R. (2003). Mining the biomedical literature in the genomic era: an overview. *Journal of computational biology*, *10*(6), 821-855.

Shetty K.D., & Dalal S.R. (2011). Using information mining of the medical literature to improve drug safety. *Journal of the American Medical Informatics Association,18*(5), 668-674.

Shneiderman B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *IEEE Symposium on Visual Languages*, September, 336-343.

Srinivasan P. (2004). Text mining: generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology, 55*(5), 396-413.

Stapley B.J.,& Benoit G. (2000). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pacific Symposium of Biocomputing, 5*, January, 529-540.

Synnestvedt, M. B., Chen, C., & Holmes, J. H. (2005). CiteSpace II: visualization and knowledge discovery in bibliographic databases. *AMIA Annual Symposium, 2005*, 724.

Thieu T., Joshi S., Warren S., & Korkin, D. (2012). Literature mining of host–pathogen interactions: comparing feature-based supervised learning and language-based approaches. *Bioinformatics*, *28*(6), 867-875.

Trefethen, L.N.; Bau III, D. (1997). *Numerical linear algebra.* Society for Industrial and Applied Mathematics, 361-369.

West, V. L., Borland, D., & Hammond, W. E. (2015). Innovative information visualization of electronic health record data: a systematic review. *Journal of the American Medical Informatics Association*, *22*(2), 330-339.

Xiang, Z., Qin, T., Qin, Z. S., & He, Y. (2013). A genome-wide MeSH-based literature mining system predicts implicit gene-to-gene relationships and networks. *BMC systems biology*, *7*(Suppl 3), S9.