Overview

Seeing beyond reading: a survey on visual text analytics

Aretha B. Alencar, Maria Cristina F. de Oliveira and Fernando V. Paulovich*

We review recent visualization techniques aimed at supporting tasks that require the analysis of text documents, from approaches targeted at visually summarizing the relevant content of a single document to those aimed at assisting exploratory investigation of whole collections of documents.Techniques are organized considering their target input material—either single texts or collections of texts—and their focus, which may be at displaying content, emphasizing relevant relationships, highlighting the temporal evolution of a document or collection, or helping users to handle results from a query posed to a search engine.We describe the approaches adopted by distinct techniques and briefly review the strategies they employ to obtain meaningful text models, discuss how they extract the information required to produce representative visualizations, the tasks they intend to support and the interaction issues involved, and strengths and limitations. Finally, we show a summary of techniques, highlighting their goals and distinguishing characteristics. We also briefly discuss some open problems and research directions in the fields of visual text mining and text analytics.

How to cite this article:
WIREs Data Mining Knowl Discov 2012, 2: 476–492 doi: 10.1002/widm.1071

INTRODUCTION

Textual documents are widely available in digital format and provide a rich source of data and information. Nevertheless, accessing and interpreting such information poses a major challenge to human analysts working in a variety of domains and situations. Even the lay person faces difficulties in identifying, handling, and selecting relevant material from the many sources available. This scenario motivates a rising number of text analytic applications that embed visual representations to assist humans in tasks that require inspection of textual material. In this article, we review recent visualization techniques being applied in this context. We provide an overview of visualizations aimed at supporting a variety of tasks, from approaches targeted at displaying the relevant content information in a single document to those aimed at displaying document collections. We briefly discuss issues involved in obtaining representative visualizations, as well as the strengths and limitations of specific approaches.

Visualization techniques vary in how they preprocess and represent text. Many techniques adopt the standard 'bag-of-words representation' from information retrieval,[1] which models text content as a set of words (or terms), each with an associated frequency count. For single documents and simple tasks, this straightforward vector representation suffices to create appealing visualizations. It is also adopted in many techniques that display document collections, as it allows inferring document dissimilarity based on comparing shared word frequencies. Other techniques extract topics or other entities with semantic meaning, which typically requires more elaborate and computationally expensive preprocessing. Moreover, a text also embeds structural organization at multiple levels and has associated attributes, or metadata, describing additional properties. Structural information is sometimes employed to realize visualizations that attempt to convey semantically richer information, whereas many visualizations focusing on document relationships or their evolution along time usually consider metadata such as authorship, citations, or publication date.

*Correspondence to: paulovic@icmc.usp.br
Instituto de Ciências Matemáticas e de Computação (ICMC), Universidade de São Paulo (USP), São Carlos, SP, Brazil

DOI: 10.1002/widm.1071

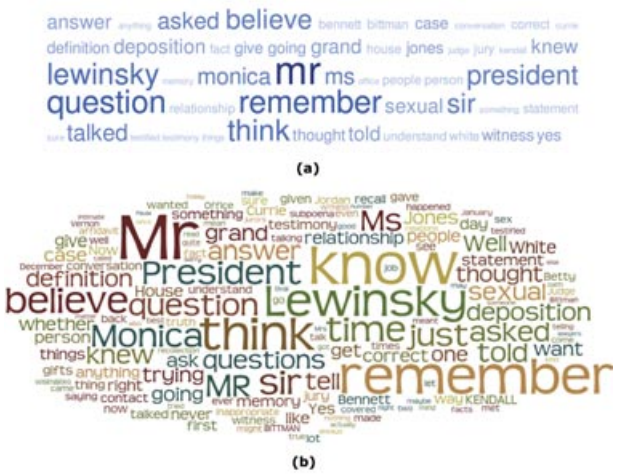

FIGURE 1 | Tag-cloud visual metaphor for the testimony of William Jefferson 'Bill' Clinton on his impeachment trial. (a) *TagCrowd* visual representation. (b) *Wordle* visual representation. The size of the font maps the frequency of the corresponding term occurring in the testimony, with larger fonts indicating more frequent terms. Images generated with the IBM Many Eyes visualization system (http://www-958.ibm.com) accessed on November 7, 2011.

VISUALIZING DOCUMENTS

Many simple visualizations of a single document simply show relevant words, or terms, considering frequency of occurrence as a relevance measure. Tag-clouds are currently a very popular visual metaphor. It presents a list of frequent terms in alphabetical order, with term frequency mapped to font size—as exemplified, e.g., by the *TagCrowd* Web application.[2] An improved visual representation, *Wordle*,[3,4] adopts a heuristic to optimize usage of the available visual area. Seifert et al.[5] also introduce an approach to render compact visualizations, in this case constrained to the interior of convex polygons of arbitrary shapes. Figure 1 shows visualizations obtained employing *TagCrowd* and *Wordle* on the text of the testimony of William Jefferson 'Bill' Clinton, former President of the United States, on his impeachment trial in 1999.

This simple approach does not guarantee, however, that sequences of related words will be placed close or sequentially in the visual representation. In *ManiWordle*[6] users are given flexible control of the layout produced by *Wordle* by supporting custom manipulations. Alternatively, a clustering algorithm has been employed to identify groups of similar terms, given by their co-occurrence in the text, and then create a visual representation that shows these clusters explicitly.[7]

On a different line, Oelke and Keim[8] propose a strategy suitable to explore extracted or calculated features that characterize documents, such as vocabulary richness or sentence length. These features represent documents at multiple levels of detail, from words to sentences and chapters. The visual representation is very simple: parts of the text (e.g., words or sentences) are mapped to screen pixels, with pixel color indicating the value of their associated features. Tests have shown that such simple visualizations result in text 'fingerprints' that are very useful to characterize texts and identify authorship.

Approaches based on term frequency, albeit appealing, cannot convey semantic relationships among terms. Several alternative visualizations attempt to overcome this limitation, e.g., representing a text as a tree that is rendered so as to enable fast content exploration. This is the underlying rationale of *Word Tree*,[9] which creates a tree with nodes representing terms and branches linking sequential terms, called a 'suffix tree'. Users can navigate on a text by selecting a word or groups of words, and checking all sentences that include them, enabling rapid exploratory queries. Figure 2(a) presents a *Word Tree* representation of the contexts including the word 'sexual' in Clinton's speech. Similarly, *DocuBurst*[10] adopts a radial space-filling layout to show semantic relations among terms, additionally mapping term frequency to font size.

Aimed at supporting more detailed analyses *Phrase Nets*[11] builds a graph where nodes represent the words and edges represent some user-specified relationship between them, defined either at the syntactic or the lexical levels. Figure 2(b) presents the visual outcome of Clinton's speech considering the clause 'is' as the target relationship between words. Font size is proportional to the number of word occurrences in a match; the thickness of an arrow between two words is proportional to how many times they occur in the same phrase. Darker font colors indicate a word more likely to be found in the first slot of a pattern. Rusu et al.[12] rely on natural language processing tools to create a directed graph that embeds semantic information, thus extending the tree representation. This solution shows existing relationships between words at a more refined level.

Another focus for text analysis is on detecting changes in the narrative flow. Miller et al.[13] address this issue considering a textual document as a signal defined by its terms. A wavelet transform is applied to this signal, and the visual outcome is a *wave* layout that can support the identification of thematic changes. Mao et al.[14] also represent documents as curves that summarize sequential trends. Abrupt changes within documents may be identified inspecting their curvatures, thus overcoming the lack

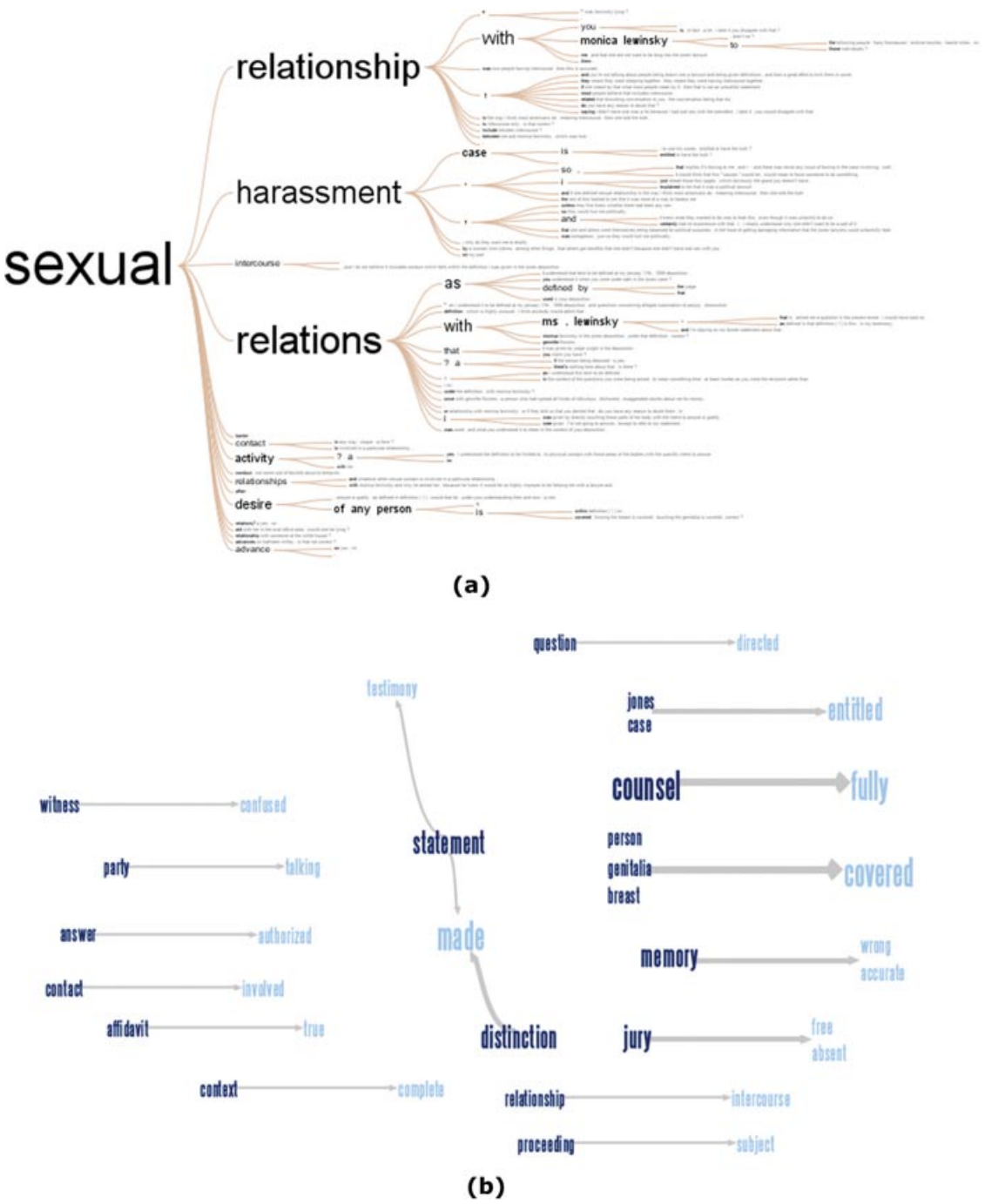

FIGURE 2 | Different visualizations that convey semantic relationships among terms occurring in the testimony of William Jefferson 'Bill' Clinton on his impeachment trial. (a) *Word Tree* representation. (b) *Phrase Net* representation. In the *Word Tree*, sequential terms in the text are linked, enabling users to navigate in the text by selecting words and checking all sentences in which they occur. The *Phrase Nets* creates a graph where nodes correspond to terms and edges correspond to user-specified relationships. In this example, the clause 'is' defines the relationship connecting the terms. Images generated with the IBM Many Eyes visualization system (http://www-958.ibm.com) accessed on November 7, 2011.

of sequential information incurred when representing documents as simple word histograms.

There are also contributions concerned with conveying document modifications along time. Information on when and how documents are created and edited is important to understand collaborative dynamics within communities, as in *Wikipedia* – the largest public wiki meant as a free online

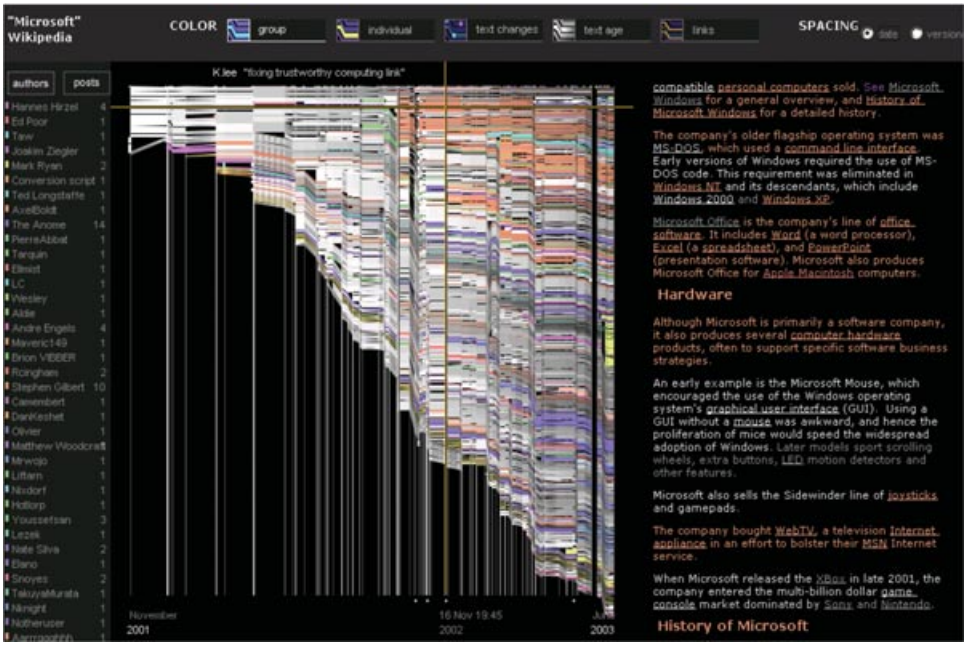

FIGURE 3 | *History flow*: this visualization highlights the temporal patterns of editions made by different authors in the *Wikipedia* entry about *Microsoft*. It shows each version of the target document as a vertical 'revision line', formed by several colored sections and with length proportional to the length of the corresponding text. Each author has been assigned a different color, and the sections of each revision line are colored according to their original author. Text sections that have been preserved across consecutive versions are visually linked. (Reproduced with permission from Ref 15. Copyright 2004 ACM.)

encyclopedia. Viégas et al.[15] proposed the technique *history flow* to highlight editions in a page, emphasizing what survives (or not) along time. A particular version of a document is represented by a vertical 'revision line' with length proportional to the text length. Authors are identified by colors, with the 'revision line' formed by sections colored to reflect their original authors. Text sections that have been preserved across consecutive versions are visually linked. In the visualization shown in Figure 3, a user has selected part of a 'revision line', and the linked text panel at the right-hand side shows the text of the corresponding version, highlighting the contributions by the author.

VISUALIZING DOCUMENT COLLECTIONS

When visualizing collections of documents, rather than individual pieces of text, document maps are a popular metaphor. Document maps are visualizations that spatially reflect some relationship among documents, providing a navigation interface useful to access information and improve human capability of solving real knowledge-management problems.[16] Map-based metaphors are appealing because they somehow mimic cartographic maps, intuitive to most users.

Two solutions for displaying document collections that rely on the familiar map metaphor are the *Cartographic Maps*[17] and *Galaxies*[18] systems. The former generates a visualization similar to a geographic map, whereas the later incorporates visualizations that resemble a night sky, for a global view. Both metaphors are available in the *IN-SPIRE*^{TM}[19] document visualization system.

From the existing methods to create documents maps, multidimensional projection techniques are possibly the most common.[20] Projection maps represent documents by graphical markers arranged in the visual space so that their proximity reflects the content similarity of their corresponding documents: close markers indicate similar documents, distant ones indicate contentwise uncorrelated documents. Projection techniques usually take as input a frequency-based vector space model of the collection, or a vector describing topics or other extracted features. Alternatively, some techniques only require a matrix describing dissimilarities, or distances, among all document pairs.

Many issues must be considered when deriving low-dimensional spatial layouts to display document collections, handled in different ways by numerous

techniques available. The *least square projection* (LSP)[20] adopts a strategy of seeking to preserve local data neighborhoods instead of pure dissimilarity between documents. LSP builds and solves a Laplacian system to place each document within the convex hull of its nearest neighbors. For high-dimensional sparse spaces, which is typically the case in vector space (bag-of-words) representations, LSP has been shown to be more effective in revealing groups of similar documents, as compared to distance-preserving approaches.

Figure 4(a) shows a map, obtained with LSP, of a collection of scientific papers—content considered includes title, authors, abstract and references—in four different areas (indicated by the colors), namely, *case-based reasoning*, *inductive logic programming*, *information retrieval*, and *sonification*. The map has been annotated with informative labels obtained with an automatic topic extraction technique based on term covariance.[21]

Automatic topic extraction is, indeed, a critical problem, as text visualizations must display informative labels. This may be addressed with data mining algorithms, e.g., Lopes et al.[22] propose an association rule mining strategy to identify meaningful term associations indicative of relevant topics. The strategy is a good example of coupling text mining with visualization: in a similarity-based map, users brush the visualization to delimit a group of documents. These are input to the rule mining algorithm, which in turn outputs meaningful term associations for labeling the selection.

Although document maps can speed up tasks that require interpreting document collections, they face some critical problems, such as the overlapping of graphical elements and the cognitive overload faced by users in layouts that show many documents at once. Hierarchical strategies have been developed to handle such limitation, allowing users to view maps at multiple levels of detail, departing from large clusters of similar documents and gradually drilling down and navigating until reaching small groups and individual documents.

InfoSky[23] offers an interesting approach for hierarchically organized document collections. A recursive Voronoi subdivision of the visual space is used to display the hierarchy and users can zoom in or out at certain areas of the projection, analogous to operating a telescope. For collections with no hierarchical structure, the *hierarchical point placement* (HiPP)[24] projection employs a recursive partitioning process to automatically infer a cluster tree from the data. Tree nodes are projected to create a multilevel visualization of groups and subgroups of documents depicted as circles within circles. Placement of circles in the visualization reflects the overall similarity of their containing documents. Figure 4(b) shows a document map created with HiPP for the same scientific paper collection depicted in Figure 4(a).

As a matter of fact, collections of scientific papers provide an ever growing body of data for visualization and pose challenges of their own. The body of techniques suitable to visualize the domain structure of scientific disciplines is generally known as 'knowledge domain visualizations'.[25] Visual representations range from the already mentioned content-based document maps to graph layouts depicting authorship or citation networks, enriched with domain-specific interaction functionalities, as discussed later on.

Other visualizations to support exploratory analysis of document collections focus on properties and attributes other than those considered in creating document maps. *Document Cards*[26] presents a quick overview of either collections or single documents aimed at enhancing browsing capability on display devices of different sizes. The technique adopts the rationale of top trumps game cards, which use expressive images and facts to provide a combined overview of an object. With a similar intent, *Document Cards* visualizations highlight important key terms and representative images extracted from a document. This solution is suitable to provide a compact visualization of a large document collection, nonetheless it fails to show interdocument relationships.

Visualizing Document Collections over Time

Time-related attributes also establish relevant relationships in collections such as news corpora, email archives or scientific articles, which have an associated time stamp informing the date/time a news piece was reported, an email sent, or an article was published. Although often ignored, the temporal component is critical for understanding and analyzing topical changes in such time-stamped document collections. This is a difficult problem that has been attracting increased attention over the last years.

Several contributions attempt to adapt existing document visualizations to handle time-stamped collections, e.g., time-oriented variations of tag clouds: *SparkClouds*[27] display a sparkline (a minimal simplified line chart) under each term to show its frequency variation over time; Cui et al.[28] introduced a visualization method that couples a trend chart with tag clouds at each time point, trying to preserve semantic coherence and spatial stability of the terms. The trend chart shows the significance of each tag cloud

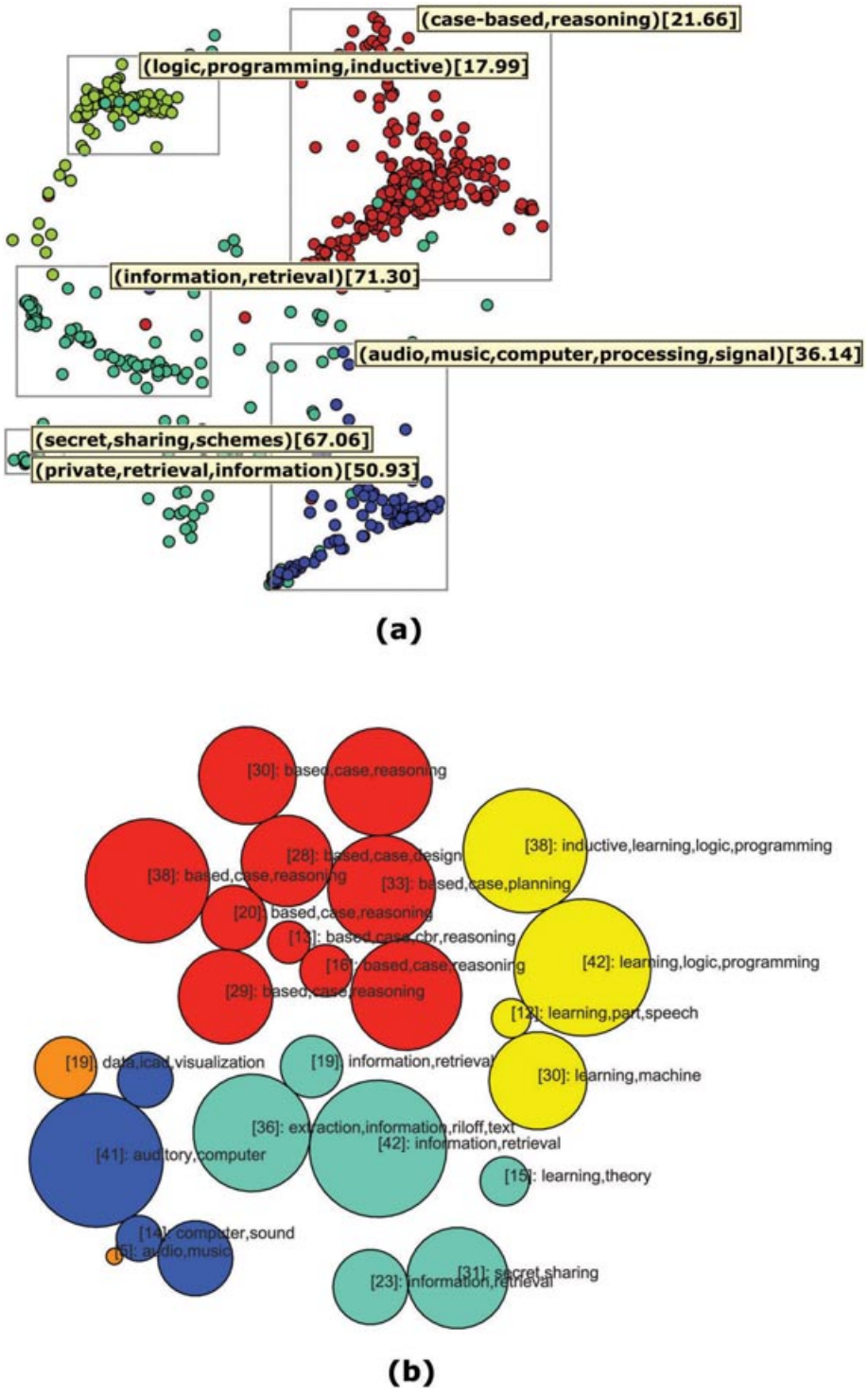

FIGURE 4 | Document maps of a collection of scientific papers obtained with multidimensional projection techniques. (a) *Least square projection*(LSP) representation. (b) *Hierarchical point placement*(HiPP) representation. On LSP, circles represent documents and are placed so that circle proximity is proportional to the similarity among the corresponding documents. On HiPP, the circles represent groups of similar documents and proximity maps the similarity between the groups. Both maps are annotated with automatically extracted topics, and the colors reflect an existing classification of the documents. (Reproduced with permission from Refs. 20, 24 Copyright 2008 IEEE.)

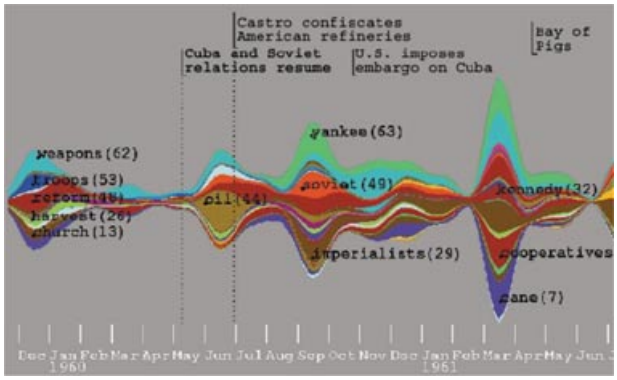

FIGURE 5 | *ThemeRiver*: visualization showing documents about the *Cuban Missile Crisis*, from December 1959 through June 1961. In this representation, the major topics addressed in the document collection are shown as colored 'streams', with stream width indicating the topic's strength at a certain moment. (Reproduced with permission from Ref 29. Copyright 2002 IEEE.)

along time, which is higher when the tag cloud conveys more information by itself with less information shared by surrounding tag clouds.

The 'river' metaphor is often applied in time-oriented visualizations with information flowing from left to right through time. The *ThemeRiver*[29] is a visualization that indicates temporal variation adopting this metaphor. It is intended to display temporal thematic changes in a document collection by highlighting selected topics represented by single words. Individual topics are visually represented as colored 'streams' within the river. Flow width indicates topic strength, and the width of the river at a specific time instant depicts the collective strength of the selected topics.

Figure 5 shows a visualization created with *ThemeRiver* from documents related to the Cuban missile crisis. This visual representation includes a river of topics (words), a timeline below the river, and markers manually added by the authors along the top to identify related historical events. The *TIARA* system (*text insight via automated responsive analytics*)[30] adopts a similar metaphor to depict temporal evolution of the topical content in collections of news or emails. *TIARA* relies on a more sophisticated strategy to identify topics, based on *latent Dirichlet allocation* (LDA),[31] a statistical approach applicable for summarizing texts into topics (represented as vectors of weighted words), and deriving time-sensitive keywords.

The same metaphor is employed in *TextFlow*,[32] now in a more complex scenario in which topics events—such as topic birth, split, merge, and death—are detected and visualized. In this technique, first a set of topics, along with merge and splitting relationships, are automatically extracted with an incremental hierarchical Dirichlet process.[33] Given this first output, topic events such as birth, death, splitting, and merging are detected. To help users to better identify topic content and understand the major reasons behind critical events, the system also detects keyword correlations by extracting terms from each document, counting their co-occurrences and displaying the top frequent keywords. This information is then visually presented in a river flow layout, formed by three layers: flows that represent the topics; glyphs that represent critical events, overlaid at the time points where they occur; and threads (blue lines) that represent the detected keywords. The choice of meaningful threads is tricky: a lot of information may be hidden if just a few keyword threads are included; on the other hand, showing many keywords results in a cluttered visualization.

Figure 6 shows *TextFlow* applied to scientific articles published in the *IEEE Information Visualization (InfoVis)* conference from 2001 to 2010. Keyword pairs pointing to flows were labeled manually by the authors. For instance, the critical event *d* indicates that the topic *document/temporal* (characterized by keywords *explore* and *document*) has become a major topic in *InfoVis* around year 2009.

The 'river' metaphor has also been employed to assist analysis of news corpora. News pieces are typically a consequence of relevant events occurring, and the underlying rationale in *EventRiver*[34] is to identify clusters of news and map them to real-life events. A temporal-locality clustering technique is applied to group news that are both similar in content and adjacent in time. Each cluster is assumed to represent an event, and events are semantically represented by extracted keywords. The proposed visual layout resembles a river of events flowing over time. Each event is shown as a bubble, for which the vertical dimension maps the number of its documents and the horizontal dimensional maps its duration. Events with the same color and place adjacent to each other are closely related and construct a long-term story, i.e., a group of events with close content. The visualization is enhanced by different interaction techniques that allow, e.g., to search for events by keywords.

An alternative strategy is to create new multidimensional projection techniques, or adapt existing ones, to handle the time attribute explicitly so as to convey temporal changes in the similarity relationships among documents from a collection. This relates to the problem of computing layouts that evolve over time to reflect changes in the data set. For instance, the *Visone* tool[35] has been employed to generate several

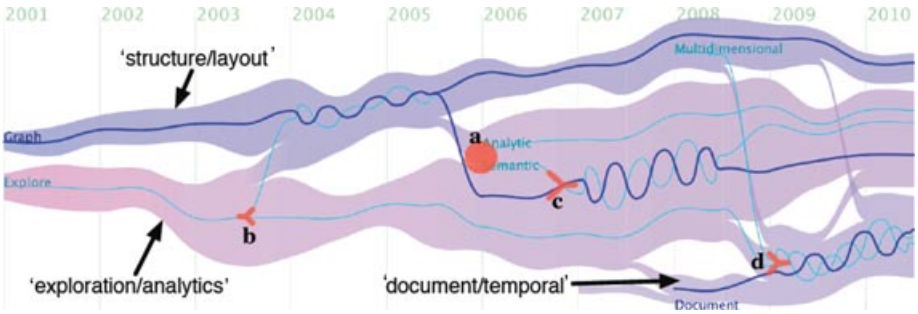

FIGURE 6 | *TextFlow*: topic flows for scientific articles published in *IEEE InfoVis* from 2001 to 2010. Similarly to *Theme River*, *TextFlow* employs a metaphor based on river 'streams' to represent the strength of different topics varying over time within a document collection. It adds extra visual marks to represent events associated with topics, such as topic birth, split, merge and death. In this example, the event marked as **d** indicates that the topic *document/temporal* has turned into a major topic in this collection around year 2009. (Reproduced with permission from Ref 32. Copyright 2011 IEEE.)

time-based networks from collections of scientific articles, including journal citation networks and heterogeneous networks that have title words, authors, and journals as nodes. The authors claim these time-based networks are good indicators of structural changes on the underlying data. *Visone* relies on an MDS (multidimensional scaling) algorithm to dynamically layout a sequence of networks by optimizing a stress measure over the current, previous, and subsequent maps. This modified stress function penalizes drastic movements of a node from a map to the next. In this manner, stability and consistence are preserved along a sequence of layouts, thus avoiding user confusion due to sudden unexpected layout changes. However, stability is not dictated solely by the data, but by a parameter in the stress function.

The *time-based least square projection* (T-LSP)[36] adapts the already introduced LSP multidimensional projection to show temporal evolution. Given a collection of time-stamped documents split into a list of batches according to some temporal property, it operates backward to generate a temporal sequence of similarity-based maps. First, the entire collection is projected using LSP generating a final layout. Then, starting from the last, each batch is processed: the documents in this batch are removed from the subsequent layout. As a result, documents in the high-dimensional neighborhoods of the removed documents must be reprojected in order to update the map. The documents that were not in the neighborhood of removed documents will remain at their current position—these stable documents are taken as 'control points' for LSP in reprojecting the others. An intermediate layout is thus generated for each batch, and finally a smooth animation is created to display the series of layouts so obtained, in the correct order. The technique seeks to maintain local accuracy and global spatial coherence throughout the sequence of maps, and unlike in *Visone* the degree of stability is dictated by the data.

A major drawback of the time-oriented techniques mentioned so far is their inability to handle document streams, such as newswires and blogs. Such techniques are not truly incremental because layouts, once obtained, cannot be rearranged to accommodate new incoming elements. The *Incremental Board* (*incBoard*)[37] handles this problem by placing a sequence of data elements (e.g., documents or images) over a two-dimensional grid of visual cells: data elements are placed incrementally and dynamically rearranged in the grid so as to reflect their relative similarity rankings, rather than a similarity metric. The solution adopted is inherently incremental, as the grid maintains a coherent disposition of elements along time while it is dynamically rearranged as elements are added or removed. Authors also extend the underlying principle to an *Incremental Space* (*incSpace*) that eliminates the grid.

With a different strategy, *Streamit*[38] visualizes text streams employing a dynamic force-directed projection. Force-directed projection techniques iteratively rearrange the data points approximating those projected too far away and repelling those projected closer than expected. This iterative process accounts for the dynamic behavior of the technique. A similarity grid over the current layout is employed to determine the initial placement of a new document: it is inserted in the center of the cell with more documents similar to the incoming one. Force-directed placement is not suitable for handling high-dimensional data, due to its quadratic complexity. Thus, the topic modeling technique LDA is employed to obtain a low-dimensional vector representation of the collection. However, because this system handles text streams,

FIGURE 7 | *Streamit*: dynamic document map for a collection of abstracts describing projects funded by the US National Science Foundation Information and Intelligent Systems award between March 2000 and August 2003, generated with a dynamic force-directed projection. Given latent Dirichlet allocation topics extracted in a preprocessing step, documents that match specific user-selected topics are presented as pie charts, with slice sizes indicating the topic's weight in the corresponding document. Circle sizes represent the amount of funding to the project. Topical events are discovered with a dynamic clustering approach: (a) September 2000—red pie slice represents topic 16 (*Query, Database, Data, XML, Stream, Edu*) and green slices represent topic 19 (*Data, Workflow, Privacy, Management, Web, Metadata*); (b) September 2001—clusters 1 and 2 from Figure 7(a) have merged into cluster 3. Clusters 4 and 5 are new. (Reproduced with permission from Ref 38. Copyright 2012 IEEE.)

LDA is actually applied to a very similar document collection to extract the topics, represented as feature vectors of terms probabilities. Each incoming document must be matched, according to its terms, to the topics extracted by LDA and then represented by a vector of the probable weights of its topics. The system also contains a dynamic clustering that automatically discovers clusters from the evolving instances and the corresponding merge and split events along time.

Figure 7 shows a *Streamit* visualization of 1,000 abstracts of projects funded by the US National Science Foundation Information and Intelligent Systems (NSF IIS) between March 2000 and August 2003. Figure 7(a) and (b) show the stream for the same month in two subsequent years. The size of circles representing the documents is proportional to the project's funding amount. The largest clusters identified are shown with background colored halos. Given the LDA topics, documents that match specific user-selected topics may be presented as pie charts with slice sizes indicating the weight of a topic in the document.

NETWORKS FOR VISUALIZING DOCUMENT RELATIONS

The techniques presented so far are mainly concerned with visualizing document content. However, documents commonly have associated properties represented as metadata. Scientific manuscripts, for instance, have properties such as title, authors and their affiliations, abstract, keywords, journal or conference name, references, and publication year. Some visualization techniques and tools have been proposed specifically to assist exploratory analysis and visualization based on such metadata. Most of these rely on network analysis,[39,40] with the units of interest—which may be papers, authors or institutions—represented as network nodes and their relationships as edges. One example are article citation networks[41] that model how articles cite others via references—articles are depicted as nodes and references as directed edges from the citing article to the cited article, indicating the information flow. Citation networks may also be built for authors, journals, and so on. The body of knowledge in complex network analysis provides a rich set of tools to characterize and understand the behavior of such networks.

The *Action Science Explorer* (ASE) environment[42] (see Figure 8) incorporates network analysis as part of its framework. This tool has been designed for exploration of collections of scientific articles through network visualization, statistics, citation context extraction, and natural language summarization. ASE partly integrates two existing tools: the *SocialAction*[43] network analysis tool and *JabRef*,[44] an open-source bibliography reference manager. The reference manager view includes the list of articles under analysis, whereas the network view includes a force-directed citation network visualization, plus functions for ranking and

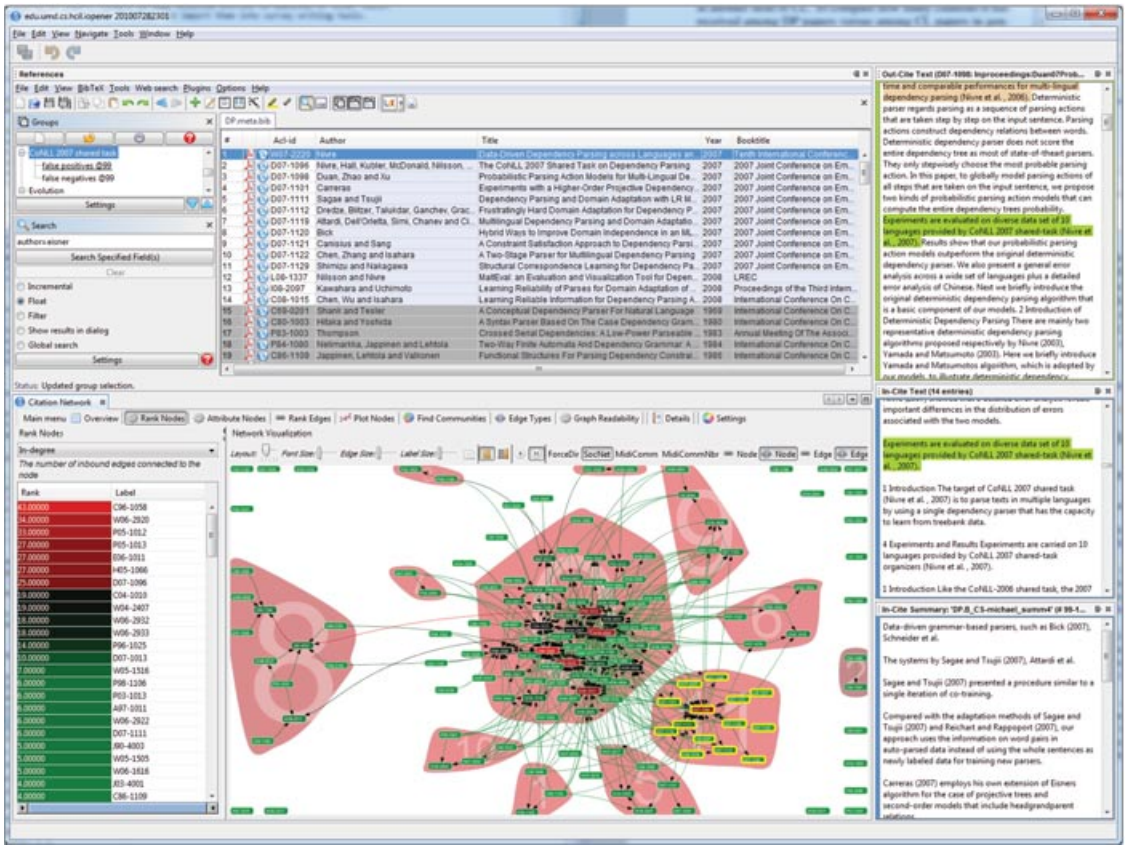

FIGURE 8 | *Action Science Explorer* (ASE): tool presenting multiple views of research papers on a particular field—tables of papers, full texts, text summaries, and visualizations of the citation network and its groups are shown. All data views are coordinated. (Reproduced with permission from Ref 42. Copyright 2012 American Society for Information Science and Technology.)

filtering papers by statistical measures, scatterplots of paper attributes and statistics, and automatic cluster detection on the network. A multidocument summarization view shows automatically generated summaries of selected articles. These views are linked and coordinated to help users discover unexpected trends, clusters, gaps, and outliers on the information flow.

The *CiteSpace II* tool[41] relies on article citation networks to visualize two related concepts: *research fronts*, a subset of highly cited articles that characterize the state of the art in a research field; and *intellectual bases*, articles heavily cited by *research front* articles. The tool builds a visual representation aimed at depicting the temporal evolution of research fronts and intellectual bases and their transient patterns, for a given research field.

Health-related document collections also share relationships, in this case based on facets. For example, documents in *Google Health* describe diseases and include information on a number of facets: symptom, treatment, cause, diagnosis, prognosis, and prevention. A relationship occurs when two described diseases share, e.g., a symptom or a prevention. *FacetAtlas*[45] is an interactive visualization proposed to help users analyzing large multifaceted document collections with complex cross-documents relationships. Applied to health-related document collections, *FacetAtlas* helps answering complex questions such as 'Which diseases can lead to this set of symptoms?'. The technique employs a multifaceted graph to visualize local relations and a density map to convey a global context.

VISUALIZING QUERY RESULTS

Many users are familiar with handling a collection of document summaries, known as snippets, retrieved by a search engine in response to a query posed as a set of terms. Typically, the snippets are presented as a list ordered by their relevance to the query, as computed by the search algorithm. The list-based snippet metaphor is simple and intuitive, but recognized as

limited in many situations. Users feel overwhelmed, for instance, when too many hits are returned, or when they try to grasp a global view of the retrieved documents and how their content relates with the query and among themselves.

Several visualization techniques and systems have attempted to provide more flexible alternatives to users inspecting and navigating the result of textual queries. Up to 1995, most information retrieval systems focused on retrieving only document titles and abstracts. Hearst[46] argued in favor of full text document search, stressing that it should indicate, in addition to the strength of the match, the frequency and distribution of relevant (e.g., query) terms in the retrieved texts. The *TitleBars* visualization has been proposed to supply this information to users performing full text searches. It visually represents each document as a rectangle icon composed by colored bars – each bar represents a set of related query terms. The bars are visually displayed as squares spatially placed to indicate a text segment that addresses a particular topic (detected automatically by the *TextTiling* algorithm[47]). Squares shown in darker colors indicate higher frequencies of a particular query term set in that specific text segment.

Figure 9 shows the results of a query by a user interested in documents addressing computer aided techniques for medical diagnosis. The query has been formulated as a conjunction of three term sets: *(patient medicine medical) AND (test scan cure diagnosis) AND (software program)*, thus each rectangle is represented as three colored bars. The rectangles length indicates the document length, whereas the colored bars simultaneously indicates the frequency of the term sets in the document and their distribution relative to the document and to each other.

Later visual techniques contributed variations to the *TileBars* strategy of enriching the basic ranked list metaphor by adding term frequency and/or term distribution information, e.g., the work by Heimonen and Jhaveri[48]; the *HotMap*[49] and *WordBars*[50] visualizations, or the *PubCloud*[51] system that creates tag cloud visualizations of abstracts returned from searching the PubMed database.

Other techniques for visualizing query hits replace the textual snippets by summary thumbnails.[52–55] As the visual summaries may include representative text and/or figures, those techniques must access the full document content. There are also visualizations that replace or complement the snippet list with alternative metaphors, e.g., a solar system,[56] a spiral shape[57] or scatter plots.[58] A comprehensive review of these and other techniques for visualizing query results is beyond the scope of this paper. The interested reader is referred to Yao et al.[59] and to Hearst[60] for further information on visual interfaces to support general and textual search tasks.

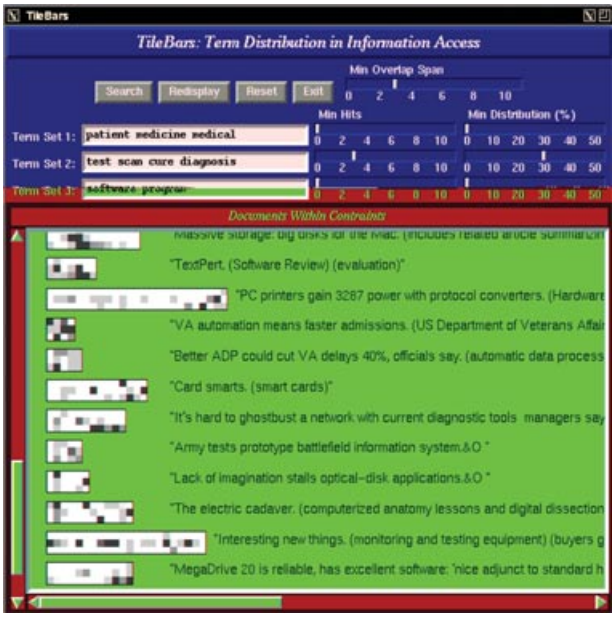

FIGURE 9 | *TileBars*: visualization of the results of a search on medical documents. Each document appears as a rectangular icon composed by colored bars spatially placed to indicate the frequencies and distribution of the query terms in the document. Squares in darker colors indicate higher frequencies of a particular query term set. (Reproduced with permission from Ref 46. Copyright 1995 ACM.)

SUMMARY, CRITICAL ANALYSIS, AND DISCUSSION

Tables 1, 2, and 3 show a summary overview of the main visualization techniques considered in this article, highlighting their underlying layout type, properties, choice of representation, tasks supported, publication year, and main reference work. They reveal a lively field with many techniques and systems proposed that rely on a rich variety of choices of visual representations and tasks to support, as well as of pre-processing and interaction strategies. One observes, however, that certain visual representations afford certain types of tasks, e.g., document maps are often employed to support tasks that require identifying content correlation among documents, and river flow metaphors are popular to convey temporal behavior. Indeed, it is remarkable that the increasing number of visualization techniques aimed at supporting analysis of temporal behavior of documents and document collections introduced over the past few

TABLE 1 | Summary of Visualization Techniques

Name	Layout Type	Properties	Representation	Tasks Supported	Year	Reference
			Techniques for Single Documents			
TagCrowd	Word cloud	Simple tag cloud	Bag-of-words	Visualize frequent terms	2011	2
Wordle	Word cloud	Heuristic to optimize area usage	Bag-of-words	Visualize frequent terms	2009	3
ManiWordle	Word cloud	Based on *Wordle*	Bag-of-words	Visualize frequent terms; custom manipulations	2010	6
Oelke and Keim	Text 'fingerprints'	Features mapped to screen pixels at multiple levels of detail	Feature values that characterize documents	Characterize texts according to the features; identify authorship	2007	8
Word Tree	Suffix tree	User specifies the term to search for contexts	Occurrences of a term, along with its following phrases	Visualize phrases (contexts) including a term	2008	9
DocuBurst	Radial space-filling layout	Visual summaries at varying granularity levels	Lexical tree structure	Show semantic relations among terms	2009	10
Phrase Nets	Network layout	User must specify the patterns	Pattern matches	Visualize patterns (relationships) between words	2009	11
Miller et al.	Wave layout	Document as signal, wavelet transform to identify changes	Narrative order of the words	Detect thematic changes in narrative flow	1998	13
Mao et al.	Curve layout	Drastic curve movements indicate thematic changes	Locally weighted bag of words representation	Detect thematic changes in narrative flow	2007	14
History Flow	Layout based on 'revision lines'	Targeted at versioned documents	Differences among version pairs	View history of versions of a document	2004	15

years. This scenario signals the great potential of visual techniques as valuable aids in text analytics tasks. Still, it is apparent that most solutions are being validated with case studies in limited scenarios, rather than with real users handling real tasks. In fact, user needs are inferred from the lack of proper support to certain tasks, and visual aids provided to fill these gaps, but little is known about how such tasks fit into the global analysis and decision making processes and the real needs of users from this wider perspective.

Considering that text analytics spans a wide variety of domains and goals, approaching the problem from a general perspective is hardly effective, and it is not surprising that many solutions reviewed focus on domain-specific tasks or on particular types of documents, as one observes from the tables. Even for visualizations targeted at a specific domain, providing the right support, requires knowing user intentions and goals, which typically vary. For instance, for users searching for a specific information, the traditional ranked list of snippets is a very appropriate viewing metaphor. A similarity-based document map might be preferable if the user wants to browse and correlate the results of an ill-posed query, for example. But then, these different needs should be detected so that a browser would switch between alternative representations, according to user convenience.

These inherent difficulties, added to the lack of more in-depth knowledge about the target audience, likely contribute for the low deployment of existing solutions to end users outside the visualization community. As such, user studies and further systematic investigation on evaluation and validation procedures are very welcome in the text analytics field. Also still lacking is a careful analysis of the low- and high-level cognitive aspects and perceptual

TABLE 2 | Summary of Visualization Techniques (cont.)

Name	Layout Type	Properties	Representation	Tasks Supported	Year	Reference
			Techniques for Document Collections			
Cartographic Map	Document map	Geographic map metaphor	Bag-of-words	View global document relationships; visually identify groups	2002	17
Galaxies	Document map	Night sky metaphor	Bag-of-words	View global document relationships; visually identify groups	1999	18
Least Square Projection (LSP)	Document map	Seeks to preserve local data neighborhoods	Bag-of-words	View global document relationships; visually identify groups; topic identification	2008	20
InfoSky	Layout based on recursive Voronoi subdivision	Hierarchical; targeted at hierarchical document collections	Bag-of-words	View global document relationships; visually identify groups	2002	23
Hierarchical Point Placement (HiPP)	Document map	Hierarchical; infers cluster tree to create hierarchical document map	Bag-of-words	View global document relationships; visually identify groups at multiple levels; topic identification	2008	24
Document Cards	Layout based on game cards	Suitable for large and small display sizes; does not show inter-document properties	Key term and image extraction	Highlight important key terms and representative images	2009	26
			Techniques for Document Collections over Time			
SparkClouds	Temporal tag cloud	Sparkline under each term shows temporal frequency variation	Term frequencies along time	Visualize trends across multiple tag clouds	2010	27
Cui et al.	Trend chart coordinated with tag clouds	Tries to preserve term semantic coherence and spatial stability along time	Term frequencies along time	Visualize trends between multiple tag clouds	2010	28
ThemeRiver	River layout	Topics represented by single words	Term frequencies along time	View temporal thematic changes	2002	29
TIARA	River layout	Topics represented as vectors of weighted words	Latent Dirichlet Allocation (LDA)	Depict temporal content evolution of topics	2010	30

processes involved in interpreting document visualizations, to guide developers into producing more effective visual solutions. For the body of techniques reviewed, no systematic studies or analyses have been reported, aimed at verifying how users perceive the relevant information and how they incorporate it into their overall decision making processes. Moreover, we identified no studies on how the short- or long-term memories are involved in handling text visualizations, nor to which extent perception of such visualizations is uncontrolled (preattentive) or controlled (attentive).

CONCLUSION

This survey has provided an overview of the lively field of visual text analytics. The variety of tasks

TABLE 3 | Summary of Visualization Techniques (cont.)

Name	Layout Type	Properties	Representation	Tasks Supported	Year	Reference
			Techniques for Document Collections over Time (cont.)			
TextFlow	River layout	Scenario with topic events: birth, split, merge and death	Incremental Hierarchical Dirichlet Process	Visualize topic evolution (events); view keywords correlated with each topic	2011	32
EventRiver	River layout	Targeted at collections of news	Keyword vectors	Identify clusters of news that can be mapped to real life events	2012	34
Visone	Dynamic network	Uses animation; based on *Multidimensional scaling* (MDS)	Citation matrices	Highlight structural changes on data	2008	35
Temporal-LSP	Temporal document map	Uses animation; based on *Least Square Projection* (LSP); seeks to maintain local accuracy and global spatial coherence	Bag-of-words	Highlight temporal changes in the similarity patterns	2012	36
Incremental Board (incBoard)	Dynamic grid layout	Streaming; uses animation; seeks to maintain local accuracy and global spatial coherence; Uses relative similarity	Bag-of-words	Highlight temporal changes in similarity patterns	2010	37
Streamit	Temporal document map	Streaming; animation; based on force-directed placement	Latent Dirichlet Allocation (LDA)	Highlight temporal changes in similarity patterns; dynamic clustering	2012	38
			Techniques for Documents Relations			
Action Science Explorer (ASE)	Network layout with coordinated views	Targeted at collections of scientific articles	Citation matrices	Multi-document summarization; cluster detection	2011	42
FacetAtlas	Graph and density map	Developed for health-related documents with facets	Multifaceted entity relational data model	Reveal multifaceted relationships within documents or cross the document clusters	2010	45
			Techniques for Query Results			
TileBars	Rectangles formed by colored bars	Targeted at for full text document search	Term frequencies within *TileBars*[47]	Visualize relative document length, query terms frequency and distribution	1995	46

and situations addressed introduces a demand for many domain-specific and/or task-oriented solutions. Nonetheless, despite the impressive number of contributions and wide variety of approaches identified in the literature, the field is still in its infancy. Deployment of existing and novel techniques to a wider audience of users performing real-life tasks remains a challenge that requires tackling multiple issues.

One issue is to foster tighter integration with traditional text mining tasks and algorithms. Various contributions are found in the literature reporting

usage of visual interfaces or visualizations to support interpretation of the output of traditional text mining algorithms. Still, visualization has the potential to give users a much more active role in text mining tasks and related activities, and concrete examples of such usage are still scarce. Many rich possibilities remain open to further exploration. Better visual text analytics will also likely require more sophisticated text models, possibly integrating results and tools from research on natural language processing. Finally, providing usable tools also requires addressing several issues related to scalability, i.e., the capability of effectively handling very large text documents and textual collections.

REFERENCES

1. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. *ACM Commun* 1975, 18:613–620.
2. Steinbock D. Tag Crowd (Home page). Available at: http://tagcrowd.com/. (Acessed November 7, 2011).
3. Viegas FB, Wattenberg M, Feinberg J. Participatory visualization with wordle. *IEEE Trans Vis Comput Graph* 2009, 15: 1137–1144.
4. Feinberg J. Wordle (Home page). Available at: http://www.wordle.net/.
5. Seifert C, Kump B, Kienreich W, Granitzer G, Granitzer M. On the beauty and usability of tag clouds. In: *International Conference Information Visualisation*. Washington, D.C.: IEEE Computer Society; 2008, 17–25.
6. Kyle K, Bongshin L, Bohyoung K, Jinwook S. ManiWordle: Providing flexible control over Wordle. *IEEE Trans Vis Comput Graph* 2010, 16:1190–1197.
7. Hassan-Montero Y, Herrero-Solana V. Improving tag-clouds as visual information retrieval interfaces. In: *International Conference on Multidisciplinary Information Sciences and Technologies*. Mérida, Spain:Open Institute of Knowledge; 2006.
8. Keim DA, Oelke D. Literature fingerprinting: a new method for visual literary analysis. In: *IEEE Symposium on Visual Analytics Science and Technology*. Washington, D.C.: IEEE Computer Society; 2007, 115–122.
9. Wattenberg M, Viégas FB. The Word Tree, an interactive visual concordance. *IEEE Trans Vis Comput Graph* 2008, 14:1221–1228.
10. Collins C, Carpendale S, Penn G. DocuBurst: visualizing document content using language structure. *Comput Graph Forum* 2009, 28:1039–1046.
11. van Ham F, Wattenberg M, Viegas FB.; Mapping text with Phrase Nets. *IEEE Trans Vis Comput Graph* 2009, 15:1169–1176.
12. Rusu D, Fortuna B, Mladenic D, Grobelnik M, Sipos R. Document visualizationbased on semantic graphs. In: *International Conference Information Visualisation*. Washington, D.C.: IEEE Computer Society; 2009, 292–297.
13. Miller NE, Chung WP, Brewster M, Foote H. Topic Islands—a wavelet-based text visualization system. In: *IEEE Conference on Visualization*. Los Alamitos, CA: IEEE Computer Society; 1998, 189–196.
14. Mao Y, Dillon J, Lebanon G. Sequential document visualization. *IEEE Trans Vis Comput Graph* 2007, 13:1208–1215.
15. Viégas FB, Wattenberg M, Dave K. Studying cooperation and conflict between authors with history flow visualizations. In: *Conference on Human factors in Computing Systems*. New York: ACM; 2004, 575–582.
16. Becks A. Benefits of document maps for text access in knowledge management: a comparative study. In: *Proceedings of the ACM Symposium on Applied Computing*. New York: ACM; 2002, 621–626.
17. Skupin A. A cartographic approach to visualizing conference abstracts. *IEEE Comput Graph Appl* 2002, 22:50–58.
18. Wise JA. The ecological approach to text visualization. *J Am Soc Inf Sci* 1999, 50:1224–1233.
19. PNNL. IN-SPIRE^{TM} Visual document analysis. Pacific Northwest National Laboratory (PNNL). Available at: http://in-spire.pnl.gov/. (Accessed October 10, 2011).
20. Paulovich FV, Nonato LG, Minghim R, Levkowitz H. Least square projection: a fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Trans Vis Comput Graph* 2008, 14:564–575.
21. Eler DM, Paulovich FV, de Oliveira MCF, Minghim R. Topic-based coordination for visual analysis of evolving document collections. In: *International Conference on Information Visualisation*. Washington, D.C.: IEEE Computer Society; 2009, 149–155.
22. Lopes AA, Pinho R, Paulovich FV, Minghim R. Visual text mining using association rules. *Comput Graph* 2007, 31:316–326.
23. Andrews K, Kienreich W, Sabol V, Becker J, Droschl G, Kappe F, Granitzer M, Auer P, Tochtermann K. The InfoSky visual explorer: exploiting hierarchical structure and document similarities. *Inf Vis* 2002, 1:166–181.

24. Paulovich FV, Minghim R. HiPP: A novel hierarchical point placement strategy and its application to the exploration of document collections. *IEEE Tran Vis Comput Graph* 2008, 14:1229–1236.
25. Börner K, Chen C, Boyack KW. Visualizing knowledge domains. *Annu Rev Inf Sci Technol* 2003, 37:179–255.
26. Strobelt H, Oelke D, Rohrdantz C, Stoffel A, Keim DA, Deussen O. Document Cards: a top trumps visualization for documents. *IEEE Trans Vis Comput Graph* 2009, 15:1145–1152.
27. Lee B, Riche NH, Karlson AK, Carpendale S. SparkClouds: Visualizing trends in tag clouds. *IEEE Trans Vis Comput Graph* 2010, 16:1182 –1189.
28. Cui W, Wu Y, Liu S, Wei F, Zhou MX, Qu H. Context-preserving, dynamic word cloud visualization. *IEEE Comput Graph Appl* 2010, 30:42–53.
29. Havre S, Hetzler E, Whitney P, Nowell L. ThemeRiver: Visualizing thematic changes in large document collections. *IEEE Trans Vis Comput Graph* 2002, 8:9–20.
30. Wei F, Liu S, Song Y, Pan S, Zhou MX, Qian W, Shi L, Tan L, Zhang Q. TIARA: a visual exploratory text analytic system. In: *ACM International Conference on Knowledge Discovery and Data Mining*. New York: ACM; 2010, 153–162.
31. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res* 2003, 3:993–1022.
32. Cui W, Liu S, Tan L, Shi C, Song Y, Gao Z, Qu H, Tong X. TextFlow: towards better understanding of evolving topics in text. *IEEE Trans Vis Comput Graph* 2011, 17:2412–2421.
33. Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet processes. *J Am Stat Assoc* 2004, 101:1566–1581.
34. Luo D, Yang J, Krstajic M, Ribarsky William, Keim DA. EventRiver: visually exploring text collections with temporal references. *IEEE Trans Vis Comput Graph* 2012, 18: 93–105.
35. Leydesdorff L, Schank T. Dynamic animations of journal maps: Indicators of structural changes and interdisciplinary developments. *J Am Soc Inf Sci Technol* 2008, 59:1810–1818.
36. Alencar AB, Paulovich FV, Börner K, Oliveira MCF. Time-aware visualization of document collections. In: *ACM Symposium on Applied Computing - Multimedia and Visualization Track*. Riva del Garda, Italy: ACM; 2012, 997–1004.
37. de Pinho R, Oliveira MCF, Lopes AA. An incremental space to visualize dynamic data sets. *Multimedia Tools Appl* 2010, 50:533–562.
38. Alsakran J, Chen Y, Luo D, Zhao Y, Yang J, Dou W, Liu S. Real-time visualization of streaming text with a force-based dynamic system. *IEEE Comput Graph Appl* 2012, 32:34–45.
39. Sci^2 Team. Science of Science (Sci^2) Tool. Indiana University and SciTech Strategies. Available at: http://sci2.cns.iu.edu.
40. Herr BW, Duhon RJ, Börner K, Hardy EF, Penumarthy S. 113 Years of physical review: using flow maps to show temporal and topical citation patterns. In: *International Conference on Information Visualisation*. Los Alamitos, CA: IEEE Computer Society; 2008, 421–426.
41. Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *J Am Soc Inf Sci Technol* 2006, 57:359–377.
42. Dunne C, Shneiderman B, Gove R, Klavans J, Dorr B. Rapid understanding of scientific paper collections: integrating statistics, text analytics, and visualization. *J Am Soc Inf Sci Technol* 2012; (to appear).
43. Perer A, Shneiderman B. Balancing systematic and flexible exploration of social networks. *IEEE Trans Vis Comput Graph* 2006, 12:693–700.
44. JabRef Development Team. *JabRef*. JabRef Development Team; 2010. Available at: http://jabref.sourceforge.net.
45. Cao N, Sun J, Lin Y-R, Gotz D, Liu S, Qu H. FacetAtlas: Multifaceted visualization for rich text corpora. *IEEE Trans Vis Comput Graph* 2010, 16:1172–1181.
46. Hearst MA. TileBars: visualization of term distribution information in full text information access. In: *Conference on Human Factors in Computing Systems*. Denver, CO: ACM; 1995.
47. Hearst MA. Multi-paragraph segmentation of expository text. In: *Proceedingsof the 32nd Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics; 1994, 9–16.
48. Heimonen T, Jhaveri N. Visualizing query occurrence in search result lists. In: *International Conference on Information Visualisation*. Washington, D.C.: IEEE Computer Society; 2005, 877–882.
49. Hoeber O, Yang XD. The visual exploration of web search results using HotMap. In: *International Conference on Information Visualization*. Washington, D.C.: IEEE Computer Society; 2006, 157–165.
50. Hoeber O, Yang XD. Interactive web information retrieval using wordbars. In: *ACM Conference on Web Inteligence*. New York: ACM; 2006.
51. Kuo BY-L, Hentrich T, Good BM, Wilkinson MD. Tag clouds for summarizing web search results. In: *International Conference on World Wide Web*. New York: ACM; 2007, 1203–1204.
52. Lam H, Baudisch P. Summary thumbnails: readable overviews for small screen web browsers. In: *Conference on Human Factors in Computing Systems*. New York: ACM; 2005, 681–690.
53. Li Z, Shi S, Zhang L. Improving relevance judgment of web search results with image excerpts. In:

International Conference on World Wide Web. New York: ACM; 2008, 21–30.

54. Teevan J, Cutrell E, Fisher D, Drucker SM, Ramos G, Andre P, Hu C. Visual snippets: summarizing web pages for search and revisitation. In: *International Conference on Human Factors in Computing Systems*. New York: ACM; 2009, 2023–2032.
55. Jiao B, Yang L, Xu J, Wu F. Visual summarization of web pages. In: *New York: ACM Conference on Research and Development in Information Retrieval*. New York: ACM; 2010, 499–506.
56. Nguyen TN, Zhang J. A novel visualization model for web search results. *IEEE Trans Vis Comput Graph*, 12:981–988, 2006.
57. Spoerri A. Rankspiral: toward enhancing search results visualization. In: *International Conference Information Visualisation*. Washington, D.C.: IEEE Computer Society; 2004, 208–214.
58. Nizamee MR, Shojib MA. Visualizing the web search results with web search visualization using scatter plot. In: *IEEE Symposium on Web Society*. Washington, D.C.: IEEE Computer Society; 2010, 5–10.
59. Jing TY, Orland H, Xue DY. *Supporting Web Search with Visualization*. London: Springer; 2010, 183–214.
60. Hearst MA,. *Search User Interfaces*. Washington, D.C.: Cambridge University Press; 2009.

