Document visualization: An overview of current research

ARTICLE · JANUARY 2014

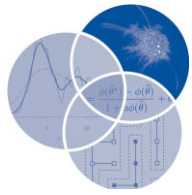
DOI: 10.1002/wics.1285

6 AUTHORS, INCLUDING:

Mingzhao Li
University of Tasmania
1 PUBLICATION **5** CITATIONS

SEE PROFILE

Overview



Document visualization: an overview of current research

Qihong Gan,[1] Min Zhu,[1]* Mingzhao Li,[1] Ting Liang,[1] Yu Cao[2] and Baoyao Zhou[2]

As the number of sources and quantity of document information explodes, efficient and intuitive visualization tools are desperately needed to assist users in understanding the contents and features of a document, while discovering hidden information. This overview introduces fundamental concepts of and designs for document visualization, a number of representative methods in the field, and challenges as well as promising directions of future development. The focus is on explaining the rationale and characteristics of representative document visualization methods for each category. A discussion of the limitations of our classification and a comparison of reviewed methods are presented at the end. This overview also aims to point out theoretical and practical challenges in document visualization.

How to cite this article:
WIREs Comput Stat 2014, 6:19–36. doi: 10.1002/wics.1285

Keywords: document; document visualization; information visualization; interactive graphics

INTRODUCTION

Recent advances in the creation, capture, and storage of information have generated an explosion of data, including a significant increase in the number of documents, such as literary works, papers, news articles, criminal case reports, web pages, emails, and so on. Few people have enough time to read everything, however, in many applications we often have to make critical decisions based on our understanding of documents. As the quantity of document information continues to increase, it is more urgent to scan, understand, operate, and navigate the enormous corpus of documents, and thus to efficiently acquire useful information and knowledge.

Document visualization is a class of the information visualization techniques that transforms textual information such as words, sentences, documents, and their relationships into a visual form, enabling users to better understand textual documents and to lessen their mental workload when faced with a substantial quantity of available textual documents.[1] Compared with other information visualization techniques, such as high dimensional data visualization that focuses on processing multidimensional data and social network visualization that focuses on visualizing the relational network among people in a representation form of node-link diagrams, document visualization pays more emphasis on visualizing textual document information. And compared with text visualization that aims to visualize information on the text level, document visualization concentrates more on visualizing documents that include attributes and metadata except the core textual contents.

Document visualization has significant advantages over helping people to analyze and control big quantities of textual information in many cases. For example, we can intuitively get access to (1) word frequency or distribution; (2) semantic content and repetition; (3) the topic or topics that define document clusters; (4) the core content of document;

*Correspondence to: zhumin@scu.edu.cn

[1]Vision-Computing Lab, College of Computer Science, Sichuan University, Chengdu, Sichuan, China

[2]EMC Labs China, Beijing, China

Conflict of interest: The authors have declared no conflicts of interest for this article.

(5) similarity among documents; (6) the connections among documents; (7) how content changes over time; and (8) information diffusion or other interesting patterns in social media, as well as improve text searches.

This overview article aims to provide a brief introduction to document visualization, representative methods, and challenges for future development. This overview is primarily intended for advanced students and researchers without a strong background in the field. Thus, we will not introduce or describe algorithms but rather focus more on discussions of the visualization design for each method. Our goal is to help readers quickly gain a systematic understanding of document visualization.

The remaining sections are organized as follows. The following section provides definition of a document and introduces the principles used in information visualization design. Next, we present categories of existing document visualization methods, and expound on the rationale and characteristics of representative methods on each category. We also summarize the advantages and disadvantages of various document visualization methods. Then, we discuss the limitations of our classification and the performance of the reviewed methods. Finally, we present several challenging and interesting directions for the field of document visualization.

DOCUMENT AND DOCUMENT VISUALIZATION DESIGN

As one of the most common methods to record information, the concept of a document has a long history. Generally 'document' is a textual record or physical form/representation of 'information'. The evolving notion of 'document' among Jonathan Priest, Otlet, Briet, Schürmeyer, and the other documentalists increasingly emphasized whatever functioned as a document rather than traditional physical forms of documents.[2] For example, Paul Otlet considers objects as documents. Suzanne Briet considers physical evidences as documents.[2] And with the development of digital technology, anything exists physically in a digital environment, such as a mail message or a technical report, could be considered as a document. In this overview, we mainly pay attention to the visualization of textual or text-like document.

Documents are often minimally structured and may be rich with attributes and metadata, especially when concentrated in a specific application domain. For example, documents have a format and often include metadata about the document (i.e., author, date of creation, date of modification, comments, size).[3]

Document visualization has significant advantages over showing the necessary content and knowledge from a collection of documents. It has become a powerful tool for helping people to analyze and control big quantities of textual information. A proper design principle for document visualization is needed. We may learn from the good practical guidelines to create an effective user interface for an interactive information visualization tool, as propounded by Ben Shneiderman who suggested in a form of mantra that an effective information visualization tool should follow the principle:

> Overview first, zoom and filter, then details on demand.[4]

The mantra is accompanied by a task taxonomy for information visualizations that specifies seven tasks at a high level of abstraction[4]:

- *Overview*. Gain an overview of the entire collection.
- *Zoom*. Zoom in on items of interest.
- *Filter*. Filter out uninteresting items.
- *Details-on-demand*. Select an item or group and get details when needed.
- *Relate*. View relationship among items.
- *History*. Keep a history of actions to support undo, replay, and progressive refinement.
- *Extract*. Allow extraction of sub-collections and of the query parameters.

When designing document visualization, we should consider the above principles. And based on those selected, after introducing the representative document visualization methods, we will evaluate those methods in *Discussion* section.

OVERVIEW OF DOCUMENT VISUALIZATION METHODS

In recent decades, a variety of document visualization methods have been proposed for different visual targets. Considering the classification of document visualization methods propounded by Daniel Keim[3] and Haidong Chen,[5] and taking into account that the visualization is task dependent, we will classify document visualization methods according to the visualization tasks and objects, in order to give a better understanding of document visualization.

We firstly divide document visualization methods into three main categories: (1) single document

visualization that has more emphasis on individual words and actual single document contents; (2) document collection visualization that has more emphasis on large document collections, themes and concepts across collection, and how documents are relate to others; (3) extended document visualization which often deals with comprehensive tasks, involves other attributes beyond the content of documents, and is always applied in specific field, such as *social media* and *search*.

Furthermore, we introduce each main category from more detailed classifications. We then present several representative methods for each category, together with extensions of these methods.

Single Document Visualization

In single document visualization, the goal is to quickly understand and absorb core content and text features. The visualization focuses on words, phrases, semantic relations, and contents. Depending on the objects and tasks of attention, we introduce single document visualization from the following aspects: vocabulary-based visualization, visualization based on semantic structure, and visualization based on document content.

Vocabulary-Based Visualization

Vocabulary is the basic unit of a document. The visualization assists people in understanding words through visual representation of the document vocabulary features, such as word frequency, word distribution, and lexical structure, thereby providing a general idea of contents and features in a document.

Tag Clouds[6,7] and Wordle[8,9] are representative methods mainly visualizing word frequency. They are widely used in the news media and personal home pages. They provide layouts of raw tokens, colored, and sized by the corresponding word frequency within a single document. We may know the main research areas/content discussed in the text by the compact visual form of words. Figure 1 presents Tag Clouds, taken from a single document introducing *operating systems*. Words are displayed alphabetically here. Alternatively, the most frequent words can be placed in the middle of the cloud. Wordle adds a layout algorithm to allow users to modify the font, color, or configuration. Figure 2 presents a Wordle of the same document. Recently, some other methods have been proposed, extending the tag/word cloud, such as parallel tag clouds (PTC),[10] ManiWordle,[11] context preserving dynamic word cloud,[12] visualization of internet discussion with extruded word clouds.[13]

TextArc[14] presents a view of a text that concisely reveals the frequency and distribution of words of an entire book. It has been developed as a way to get an overview of the document. A TextArc is a structure built with all words that are placed in the same order how they appear in the document. Each word is placed in order around an ellipse with a slight offset at its start and in a tiny unreadable font. The text is then repeated word by word around an inner ellipse and in a visible font, and the word is drawn only once if it appears more than once. Words with higher frequencies are drawn in the center, and pulled by its occurrences on the ellipse. Users can highlight the underlying text and understand the content by visualizing the flow of the text through connected words.[14] The visualization also provides links to the original document. Figure 3 shows a TextArc that visualizes *Alice's Adventures in Wonderland*. Although TextArc was designed as a tool, it has been exhibited in the Museum of Modern Art in New York.[15]

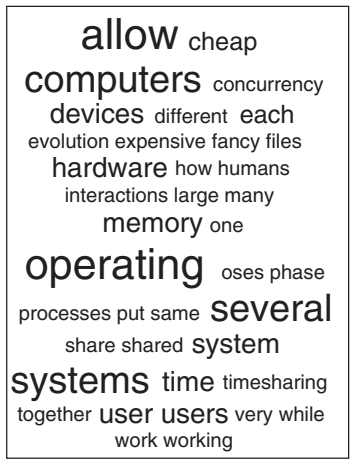


FIGURE 1 | Tag clouds.

DocuBurst combines word frequency with the human-created structure in lexical databases to create a visualization that reflects both document content and semantic content. It intends to bring visual exploratory in digital libraries, and to provide (comparative at a glance) interactive summaries of documents, which can serve as decision support when selecting documents of interest.[16] It combines the human-annotated IS-A noun and verb hierarchies of WordNet,[17] and uses a radial, space-filling layout to visualize document lexical content. Figure 4 shows a DocuBurst of a science textbook rooted at {idea}. The root node's immediate children are the synsets (i.e., sets of words and collocations) containing that word. Users can switch the root node by clicking a displayed word or searching for a word of interest. The synset is sized and colored by word frequency in the document. The visualization also provides link access to the source document.

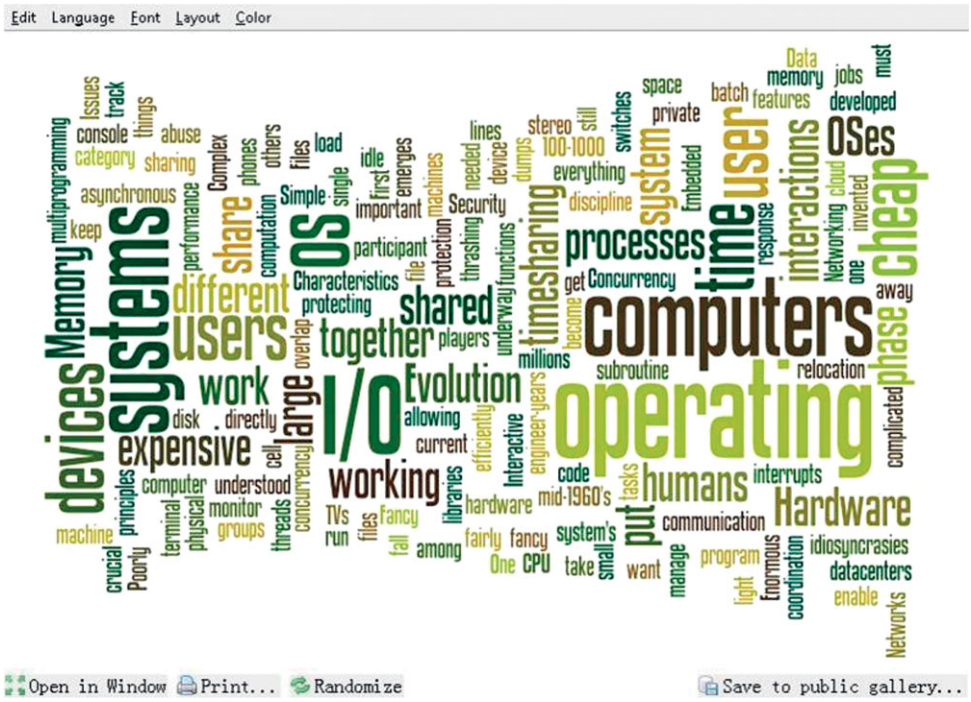


FIGURE 2 | Wordle.

Visualization Based on Semantic Structure

Semantic structure[18] is the semantic representation associated with language. Understanding of semantic structure in a document helps users get an overview of the key content of the document without entirely reading it. Visualization based on semantic structure usually use entities and their relationships to reveal the semantic content. For example, extracted subject–verb–object triplets are mapped to the visual view.

Semantic Graphs[19] is a visualization based on the semantic representation of a document in the form of a semantic graph. Firstly, it extracts subject–verb–object for each sentence by the Penn Treebank parse tree. Then, it links the triplets to their corresponding entity, which needs to resolve pronominal anaphors as well as to attach the associate WordNet synset. Thus, the document is summarized with the semantic graph and the list of extracted triplets. Figure 5 shows a semantic graph obtained from a news article. It describes a new record for powered personal flight created by Yves Rossy, who flew above the Alps with four jet engines strapped to his back. There also exist some similar visualization methods[20–22] based on semantic structure.

Visualization Based on Document Content

Visualization based on document content is not only to search for specific words but also to obtain the characteristics and relations of the contents in the document.

The WordTree visualization provides the representation of both word frequency and context. Size is used to represent frequency of the term or phrase. The root of the tree is a user-selected word or phrase, and the branches represent the contexts in which the word or phrase is used in the document. Users can click on a branch, choose a different search term or re-center the tree.[23] Figure 6 shows the WordTree visualization generated by the free online service ManyEyes.[24]

Martin Wattenberg's Arc Diagrams[25] is a visualization method that focuses on showing complex patterns of repetition. It is suited to the analysis of highly structured data like musical compositions and less well-structured data like a web page. Repeated subsequences are identified and connected by semicircular arcs. Height of the arcs represents the distance between the subsequences; and thickness of the arcs represents the length of the subsequences.[25] Inspired by Wattenberg's Arc Diagrams, Jeff Clark has made some improvements to make it suitable for visualizing arbitrary text documents by connecting segments that contain similar words, rather than using arcs to connect identical patterns. And it adds a text label. Figure 7 shows an Arc Diagram generated from a document for the State of the Union Address for 2012. The figure shows that some same terms exist in

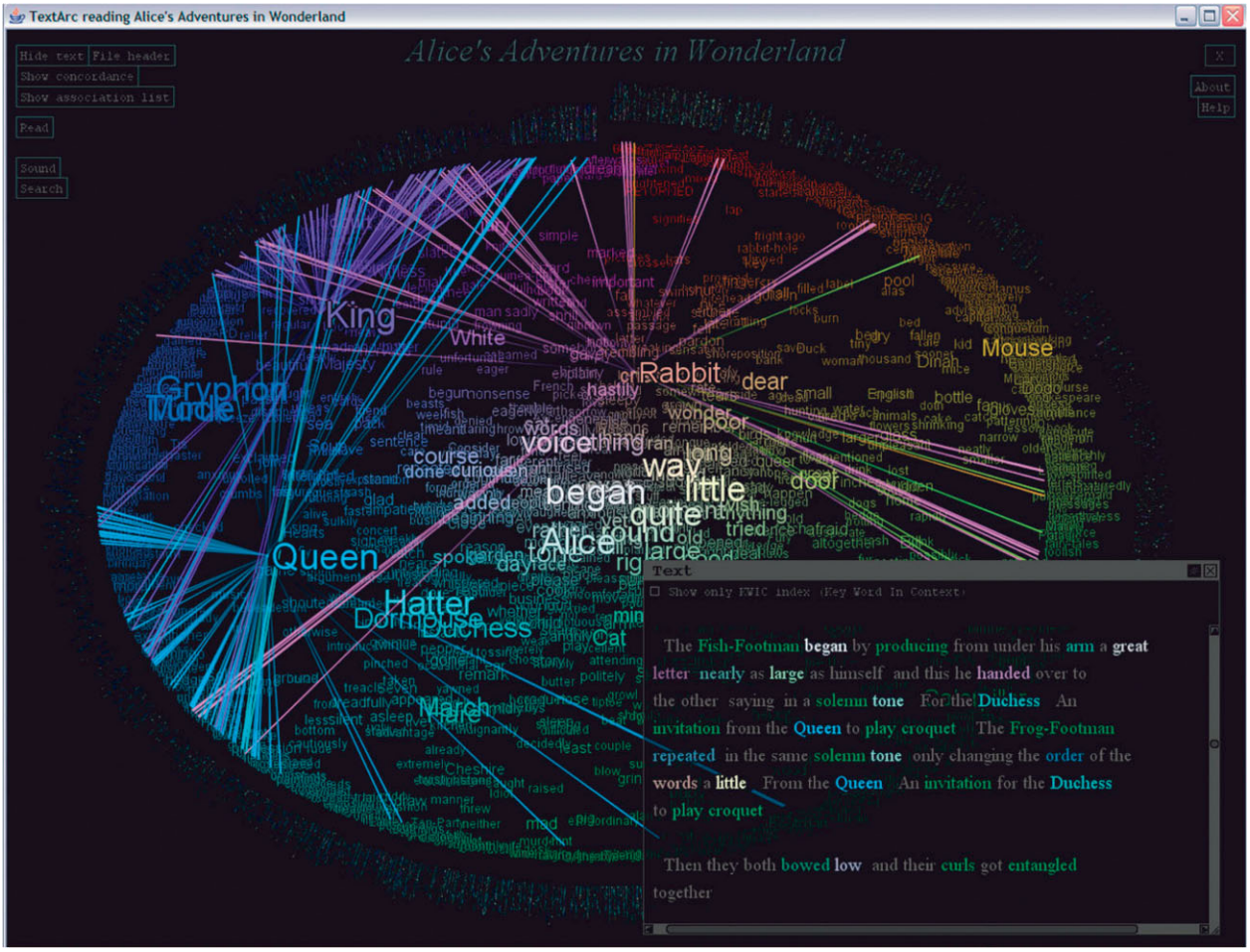


FIGURE 3 | Alice's adventures in Wonderland in TextArc. http://www.textarc.arc.org/.

the beginning and the end, which focus on *America*, *war*, and *together*. The middle mainly consists of three parts that separately focus on *taxes*, *work*, and *jobs*; *energy* and *technologies*; *right* and *tax*.

Document Collection Visualization

When faced with large quantities of document collections, searches for information in traditional ways are inefficient and the results may not be satisfactory. Document collection visualization usually intends to reveal the topic or topics that define document clusters, the similarities and differences among documents, and how contents change over time. According to the different visualization tasks, we introduce document collection visualization from the following aspects: visualization of document themes, visualization of document core content, visualization of changes over different versions, visualization of document relationships, and visualization of document similarity.

Visualization of Document Themes

Visualization of document themes is a common pattern for large-scale documents. The main goal is to discover one or more specific topics and to reflect the relationships among various topics. It may be used to find hot disciplines, evolutions, and trends.

The methods, such as ThemeScapes,[26] IN-SPIRE's ThemeView, and The Galaxy,[27] all developed by the Pacific Northwest National Laboratory, having less emphasis on the time factor, focus more on characteristics of the document themes at some specific points. ThemeScapes and ThemeView have something similar. ThemeView uses a 3D terrain map display to represent different themes. The height of a mountain represents the theme's strength, and the distance between two mountains represents the similarity between the two themes. Keywords are used to distinguish each mountain.[27] Figure 8(a) shows a ThemeView generated from a collection of news reports. The Galaxy visualization uses a similar approach that themes are visualized as 2D

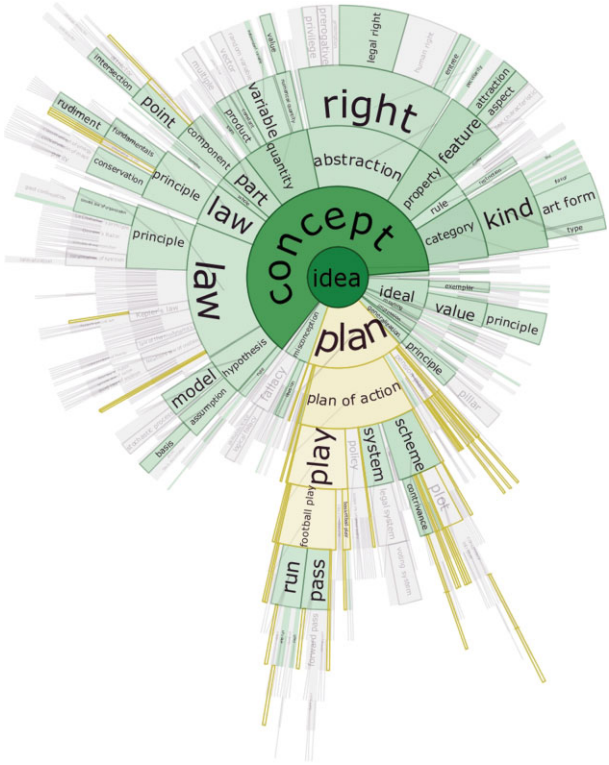


FIGURE 4 | DocuBust of a science textbook rooted at {idea}. A search query for words starting with 'pl' has been performed. Nodes matching the query are highlighted in gold. (Reprinted with permission from Ref 16. Copyright 2009 The Eurographics Association and Blackwell Publishing Ltd)

clouds of document points-stars in a theme galaxy (Figure 8(b)).[27]

There are other representations for visualizing documents and topics as nodes in a node-link graph. TopicNets is a web-based system for visual and interactive analysis of large sets of documents using statistical topic models.[28] The main view is a document topic graph which can allow aggregate nodes. The time dimension is represented as a separate visualization, with documents placed chronologically around a broken circle, and connected to related topic nodes which are placed inside the circle.

The methods, such as ThemeRiver[29] and Topic Island,[30] have greater emphasis on the time factor, focusing more on visualizing thematic variations over time within a collection of documents. ThemeRiver is in the form of axes, with the *X*-axis representing time and the *Y*-axis representing different themes. The 'river' flows from left to right through time, changing width to portray changes in theme strength of corresponding documents. Rivers of different colors represent different themes, and the width of river (i.e., narrow or wide) indicates the strength (decreasing or increasing) of an individual topic in the associated documents.[29] Figure 9 shows a portion of the ThemeRiver generated by the box office receipts from 1986 to 2007. Each 'river' corresponds to a single film, the height shows the weekly box office revenue, and the width shows its longevity. The color also reflects the box office. We can easily notice some successful films, like 'Harry Potter and the Goblet of Fire'. ThemeRiver visually shows topic summary in an esthetically pleasing manner simultaneously brings a better user experience. There are many researches based on it, as Topic-based Visual Text Summarization and Analysis,[31] Understanding Text Corpora with Multiple Facets,[32] TextFlow,[33] and TIARA.[34]

Visualization of Document Core Content

Visualization of document core content mainly intends to give an overview of a collection of documents without reading them entirely.

Document Cards[35] visualizes large document collections, such as paper collections and news reports, which contain both texts and images to describe facts, methods, or stories. It represents the document's key content as a mixture of images and important terms, similar to cards in a top trumps game.[35] The pipeline for creating Document Cards is as follows: firstly, extract the text from the original document, and use a text mining approach to extract the key terms; then go to the phases of image extraction, including image processing and image packing; finally layout the extracted key terms and images to generate the corresponding document cards. As shown in Figure 10, the IEEE InfoVis 2008 proceeding corpus is represented by a matrix of Document Cards. Users can click an image to translate it to the corresponding page or a term to highlight the lines containing the term. It is convenient to switch pages directly based on the numbers on the right side of each card.

Visualization of Changes over Different Versions

Visualization of changes over different versions is used to visualize differences among multiple document versions that are generated over time. For instance, it can be used for visualizing Wikipedia documents, code documents and corresponding documents in the software development.

History Flow[36] is designed to show changes between multiple document versions on Wikipedia. It can visualize the process of content changes and the corresponding authors who make the amendments. It also reveals some complex patterns of cooperation and confliction, such as vandalism and repair, anonymity versus named authorship, negotiation, and content

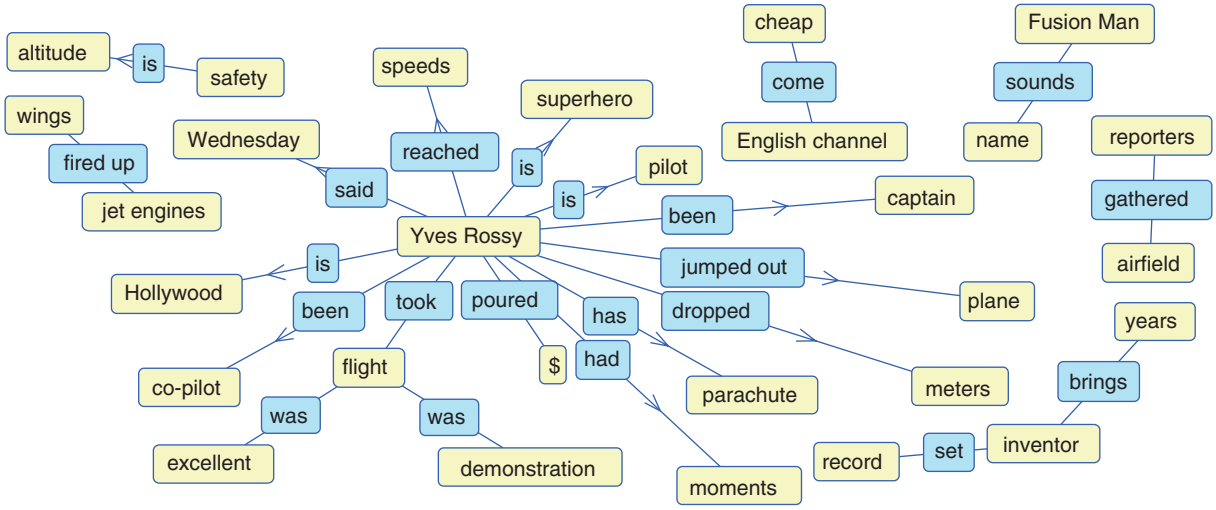


FIGURE 5 | The semantic graph obtained for a news article. http://saatviga.wordpress.com/2010/10/13/constructing-semantic-graphs-from-text-documents/.

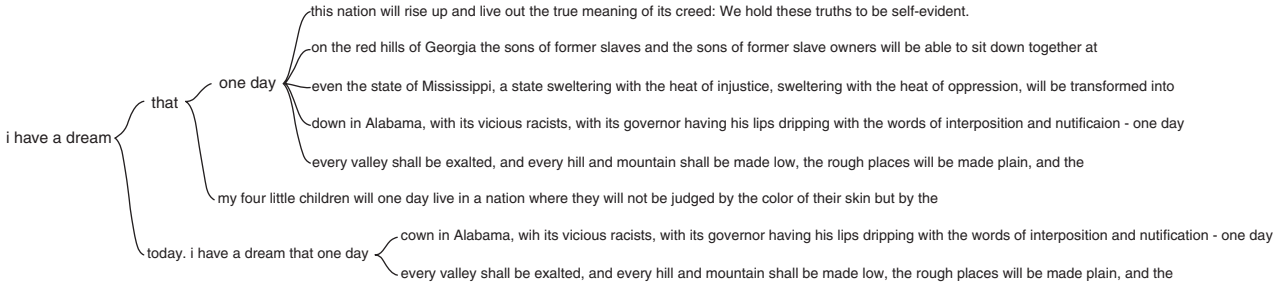


FIGURE 6 | WordTree shows all occurrences of 'I have a dream' in Martin Luther King's historical speech. http://www-958.ibm.com/software/data/cognos/manyeyes/.

stability.[36] As shown in Figure 11, each version of the document is represented by a line with length which is related to the length of its text, and the lines are ordered by date. Sections of each line are colored differently according to the different authors. And the key step is to visually link sections of text that keep the same between adjacent versions by drawing shaded connections. Figure 12 shows a view of the version history for the Wikipedia entry *Treaty of Trianon*. The left side reveals different authors that are colored differently. The center part shows the visualization; and in the right side, a text view closely linked with the visualization shows the detailed content. Users can locate on the visualization by moving a set of crosshairs, and the text view shows the corresponding version and position. The buttons shown at the top of view allow users to change the color scheme, such as highlighting one author's contributions, showing multiauthor simultaneously, highlighting the text changes, etc.[36]

Software visualization[37–39] focuses on visualizing the software development. SeeSoft,[40] Augur,[41] and Advizor[42] are visualizations for code documents. Xia gives visual insight into version control activities, like architectural and coding differences between two software versions.[43] Beagle visualizes changes among different released versions.[44] Spectrograph shows the time and location where changes happen in the system.[45]

Visualization of Document Relationships

With gradual increases in document quantity, the concepts and entities within documents become larger and larger, making the analyst's task of evaluation and sense-making more difficult. Thus, it is quite meaningful to visualize connections among documents. The visualization focuses on the correlation among documents, like the connections among entities across different documents.

Jigsaw[46] is an interactive visualization for document exploration and sense-making, and it supports the analysis of relationships among documents. It visually shows connections between entities in the documents; where entities could be

FIGURE 7 | A Document Arc Diagram generated from the text for the State of the Union Address for 2012. http://www.neoformix.com/Projects/DocumentArcDiagrams/index.html.

people, places, dates, organizations, and so on. It is suitable to documents describing a set of observations or facts, like news stories and case reports. It provides multiple views and each view provides a different perspective. As shown in Figure 13, (a) *list view* obtains multiple lists of entities colored differently in which connections between entities are shown by drawing links among them, also multiple sorting options are accessible; (b) *graph view* shows connections between documents and entities in a form of node-link diagram, documents are represented by white rectangles and entities are represented by circles colored differently, it allows analysts to explore the documents by showing and hiding links or nodes; (c) *scatter plot view* highlights pairwise relationships between any two entity types and supports concentrate on a specific subset; (d) *document view* shows the original document content with entities highlighted, displays the frequency of a document already has been viewed, and allows an entity to be modified; (e) *calendar view* provides an overview of documents and entities in a calendar way, according to the creation date; (f) *document cluster view* can group the documents into meaningful clusters in accordance with users' requirements.[46]

There are other methods for visualizing relations among multiple facets. ContexTour[47] presents the relations among conferences, authors, and topics in paper collections. FacetAtlas[48] shows relations among causes, symptoms, treatments, and diagnoses in Google Health documents. PivotPaths[49] visually explores relations of authors, keywords, and citations in academic publications.

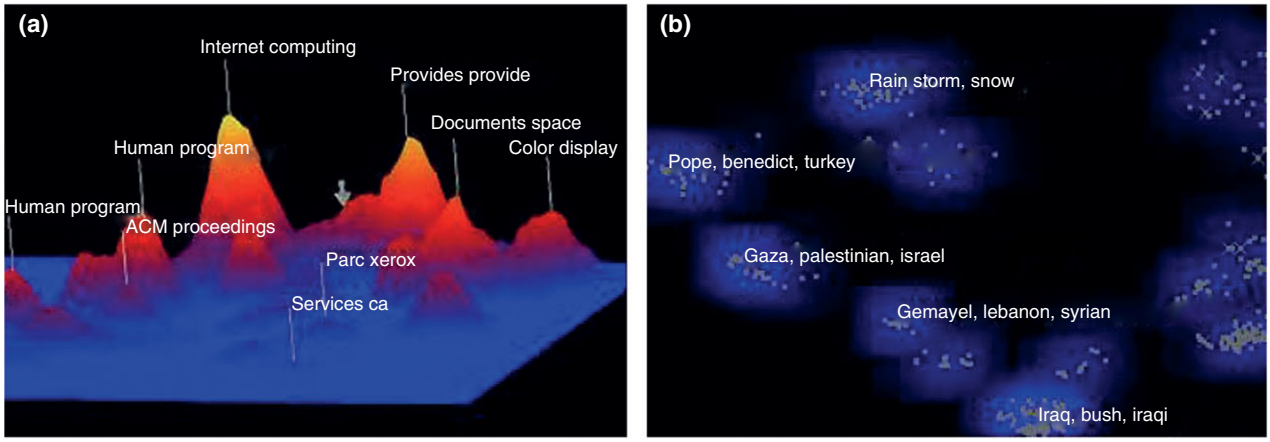


FIGURE 8 | IN-SPIRE (a) ThemeView: mountain height represents a theme's strength. The distance between mountains represents their similarity. (b) Galaxy: a 2D view of clouds of document points. http://in-spire.pnnl.gov/.

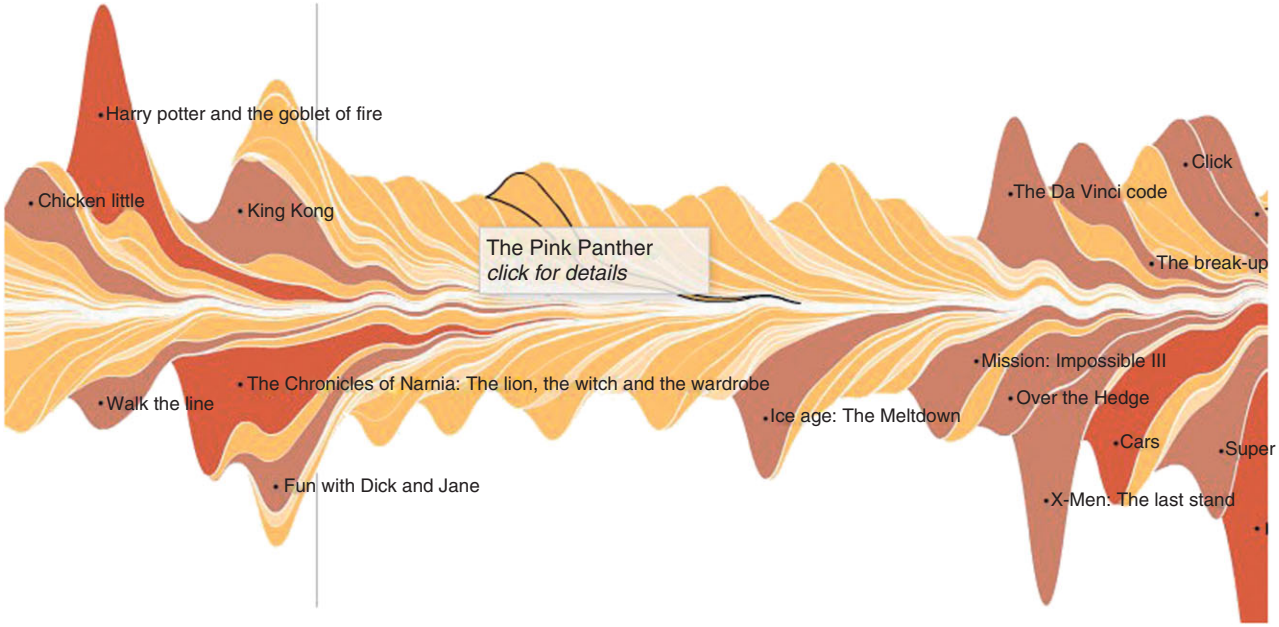


FIGURE 9 | A portion of the ThemeRiver generated by the box office receipts from 1986 to 2007. http://infosthetics.com/archives/2008/02/ebb_flow_of_box_office_movies.html.

Visualization of Document Similarity

In many cases of document collection visualizations, the goal is to place similar documents close to each other and dissimilar ones far apart.

The self-organizing map (SOM)[50] is a nonlinear projection method. It expresses complex, nonlinear relationships between high dimensional data items into simple geometric relationships on a 2D display. When applying to information retrieval, it usually uses map displays.[51] Different colored areas represent different concepts in documents. Size of area indicates its relative importance in collection. Neighboring regions show commonalities in concepts. Dots in regions can represent documents. Additional information can be referred in Xia Lin's map display[52] and WEBSOM.[53]

Extended Document Visualization

In this section, we will discuss the existence and examples of domain-centric document visualization. Relatively speaking, the methods in the previous two main categories are more versatile; while methods for domain-centric document visualization are always designed for specific domains. Here we briefly introduce some examples, as streaming text visualization and search visualization.

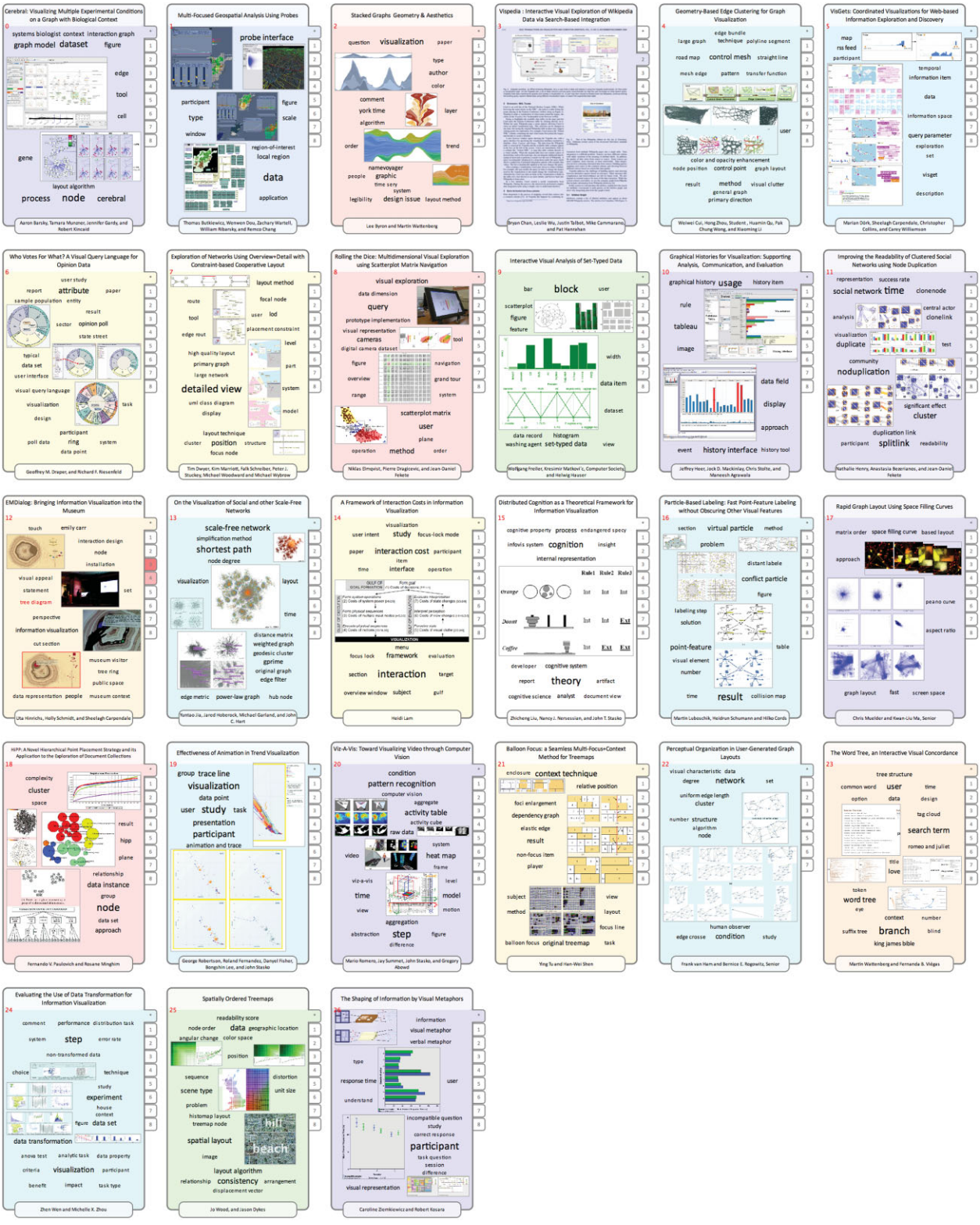


FIGURE 10 | The IEEE InfoVis 2008 proceeding corpus represented by Document Cards (DC). DC3 is switched to the page view on page2. In DC12 the term 'tree diagram' is clicked. The frequency of the term on each page is shown on the right side of the DC (the more red, the higher the frequency). (Reprinted with permission from Ref 35. Copyright 2009 IEEE)

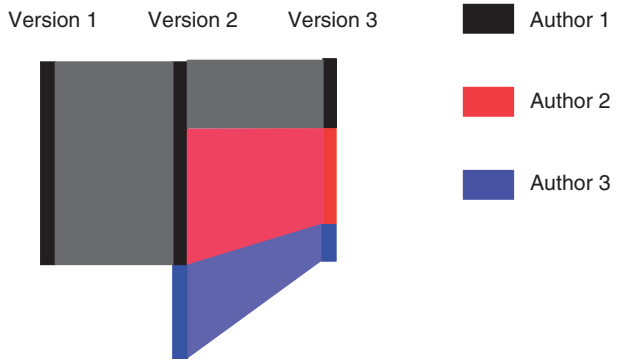


FIGURE 11 | Explanation of the visualization design for history flow.

With the rise of social media (a textual medium), text streams, such as Twitter posts, are being generated in volumes that grow every day. A large body of research has appeared in recent years. Those works have different focuses and always involve multiple targets, such as dealing with the statistical analysis and presentation of topics or terms, focusing on the emergence of topic events,[33] and visualizing the text messages themselves.[54] Christian Rohrdantz[55] provides an overview of real-time visualization of streaming text data. STREAMIT[56] presents a similar visual representation of text streams which applies to news documents. Whisper[57] fulfills the requirement for tracing information diffusion processes in social media, in a real-time manner.

Search Visualization visualizes the results of search operations. The relatively early approach is TileBars[58] that intends to minimize time and effort for deciding which documents to view in detail. Susan Havre[59] introduces a graphical method called Sparkler for visually presenting and exploring the results of multiple queries simultaneously. RankSpiral addresses the problem of how to enable users to visually explore and compare large sets of documents that have been retrieved by different search engines or queries.[60]

DISCUSSION

In the previous section, we have introduced several representative methods on each category for visualizing documents. In this section, we will discuss the limitations of our classification. And then those techniques will be summarized in Table 1 in order to allow readers quickly get an overview of the methods. Finally, we simply summarize the characteristics of document visualization, in particular, the commonness among these methods.

We have mainly considered the visualization objects and tasks when classifying document visualization methods. Our classification is considered more acceptable than other classifications (e.g., representations: pixel-based, map-based, tree-based graphs, node-link diagrams, circle graphs, etc.[1]), since visualization is usually task dependent, and users commonly begin with data and tasks. Actually, each method may belong to different category even under the same classification criteria; and we classify each method according to its key visualization focus (the visualization objects and tasks). Due to the space limitations, we only present the representative methods for each category. Nevertheless, some other methods for each category are listed in order to refer the interested readers to the original publications for detailed explanations.

In Table 1, we summarize and compare those methods mainly from four aspects to give readers a brief view.

- *Characteristics visualized*. The characteristics of a document visualized by the method, as word frequency, semantic relations, content, changes, or connections among documents.
- *Principles satisfied*. The design principles satisfied, as noted in *Document and Document Design* section, the seven tasks: ①Overview; ②Zoom; ③Filter; ④Details-on-demand; ⑤Relate; ⑥History; ⑦Extract.
- *Requirements for a document*. Document types suitable to the visualization method, i.e., whether the visualization method has special requirements for a document, like document content, structure, etc.
- *Main features*. Discuss the visualization method's features, especially the versatility and interactivity.

Generally, the design principles introduced in section *Document and Document visualization Design* are considered as one of the methods for evaluating visualizations. The visualization method might be considered better if it meets more design principles, although some classical visualization methods which do not concern all the principles, like ThemeRiver, are widely recognized. Actually from Table 1, it is obvious that the visualization methods have paid more and more attention to interactivity, and functions have become more and more powerful with the development of document visualization.

On the whole, text is nominal data; it does not seem to be mapped to graphical representation as

TABLE 1 | A Summary of the Reviewed Document Visualization Methods

Classification		Representative Methods	Characteristics Visualized	Principles Satisfied	Requirements for a Document	Main Features
Single document	Vocabulary-based visualization	Tag Clouds /Wordle	Word frequency	①③⑥	Arbitrary text documents	Strong versatility; interactive
		Textarc	Word frequency and distribution	①③④⑤	Arbitrary text documents	
		Docuburst	Word frequency and human-created structure in lexical databases	①②③④⑤⑦	Arbitrary text documents	
	Visualization based on semantic structure	Semantic graphs	Semantic content	①③⑤	Arbitrary text documents	Limited interactivity; relatively strong versatility
	Visualization based on document content	Wordtree	Word frequency and their context	①②③④⑤⑥⑦	Arbitrary text documents	Weak interactivity; versatile
		Arc diagrams	Patterns of repetition	①	Sequence based (Music, Code, Web Page, DNA)	Static
Document collections	Visualization of document themes	Themescapes/ Themeview/ Galaxy view	Relationships between various topics	①⑤	Certain theme needed in the document	Versatile within a certain range; weak interactivity
		TopicNets	Topics and their relations	①②③④⑤⑥⑦	Topic based	Versatile; strong interactivity
		Themeriver	Thematic variations over time	①⑤	Theme based	Versatile; weak interactivity
	Visualization of document core content	Document cards	Core content of a collection of documents	①②③④	Both text and image contained in the document (Paper, News Feeds)	Relatively weak interactivity; high utilization of space
	Visualization of changes over versions	History flow	Content changes between multiple document versions	①②④⑤⑥	Different versions generated over time, (Wikipedia documents)	Limited interactivity; versatile
	Visualization of document relationships	Jigsaw	Connections among document collections (especially connections among entities)	①②③④⑤⑥⑦	Facts or observations described in document (news stories, case reports)	Multiple view; strong interactivity
	Visualization of document similarity	WEBSOM	Similarity among document collections	①②④⑤⑦	Arbitrary ext documents	Versatile

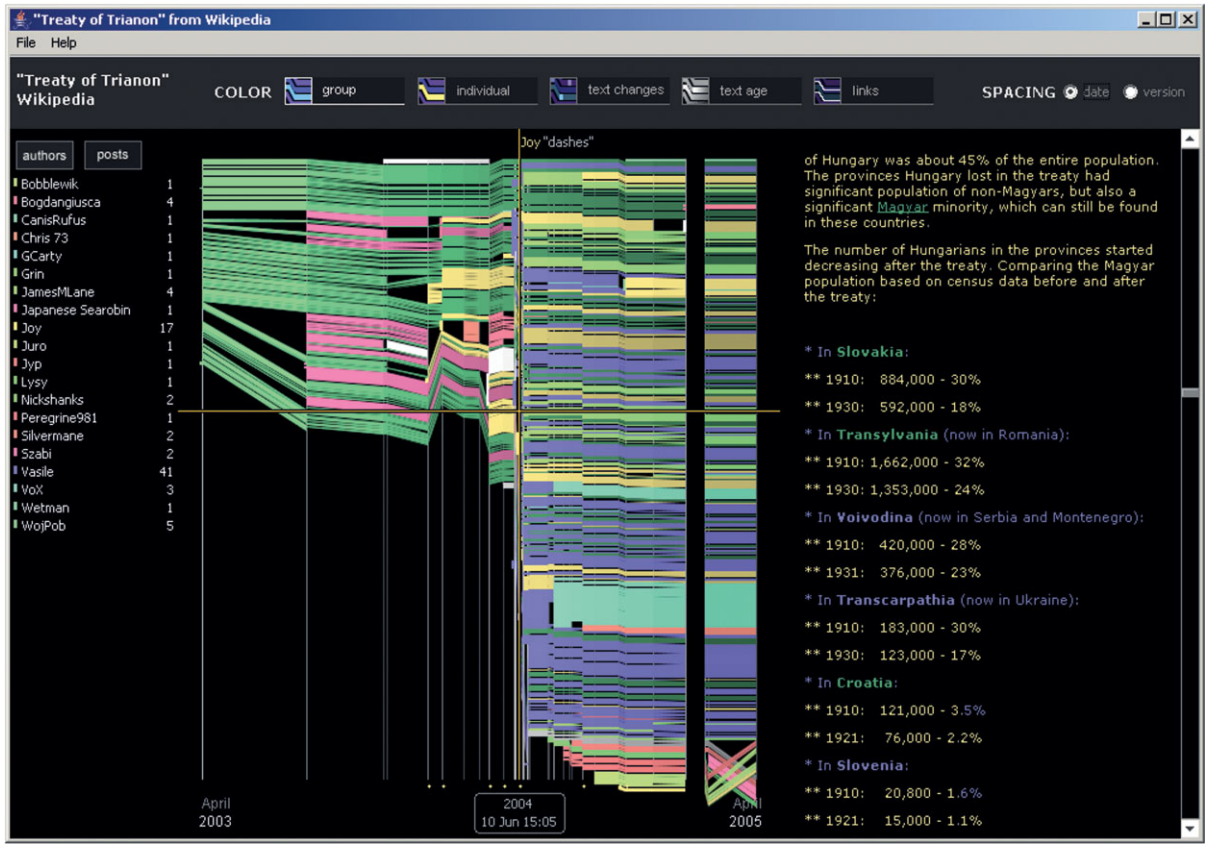


FIGURE 12 | History flow shows the chocolate page on Wikipedia. The zigzag pattern which turns out that this is an argument over whether a certain type of surrealist sculpture exists or not. http://commons.wikimedia.org/wiki/File:English_Wikipedia_Treaty_of_Trianon_History_Flow.png.

easily as ordinal and quantitative data. Moreover, the visualization representation forms for documents are various and flexible. There is no versatile pattern. The visualization is always designed according to specific tasks. Despite this, document visualization shares the same pipeline: get the data (a document or documents), transform it into vectors, then run algorithms based on the tasks of interest (i.e., similarity, search, clustering) and generate the visualizations.

CONCLUSION

Document visualization techniques combine human wisdom and computer graphics, allowing users to efficiently and intuitively browse, explore, and understand the increasing quantity of documents. In this article, we introduced document visualization, including the definition, the difference between it and other information visualization techniques, and the design principles. We then presented related research and representative methods on each category of document visualization, focusing on the design, strength and weakness, and applicability of each method. Also, we discussed the limitations of our proposal and presented the summary of those methods.

We conclude this overview with some recommendations based on emerging challenges and the most promising directions for future research. In the following, we are going to outline some issues that have the potential to be interesting areas for future research.

Improvement of Current Methods

1. Extension Existing methods can be extended to suit for large-scale document collections. Especially with the concept of *Big Data* proposed, recording and storing massive information has been bringing greater opportunities and challenges.

2. Versatility It is significant to design relatively general visualization models for different tasks within this field, since existing methods always

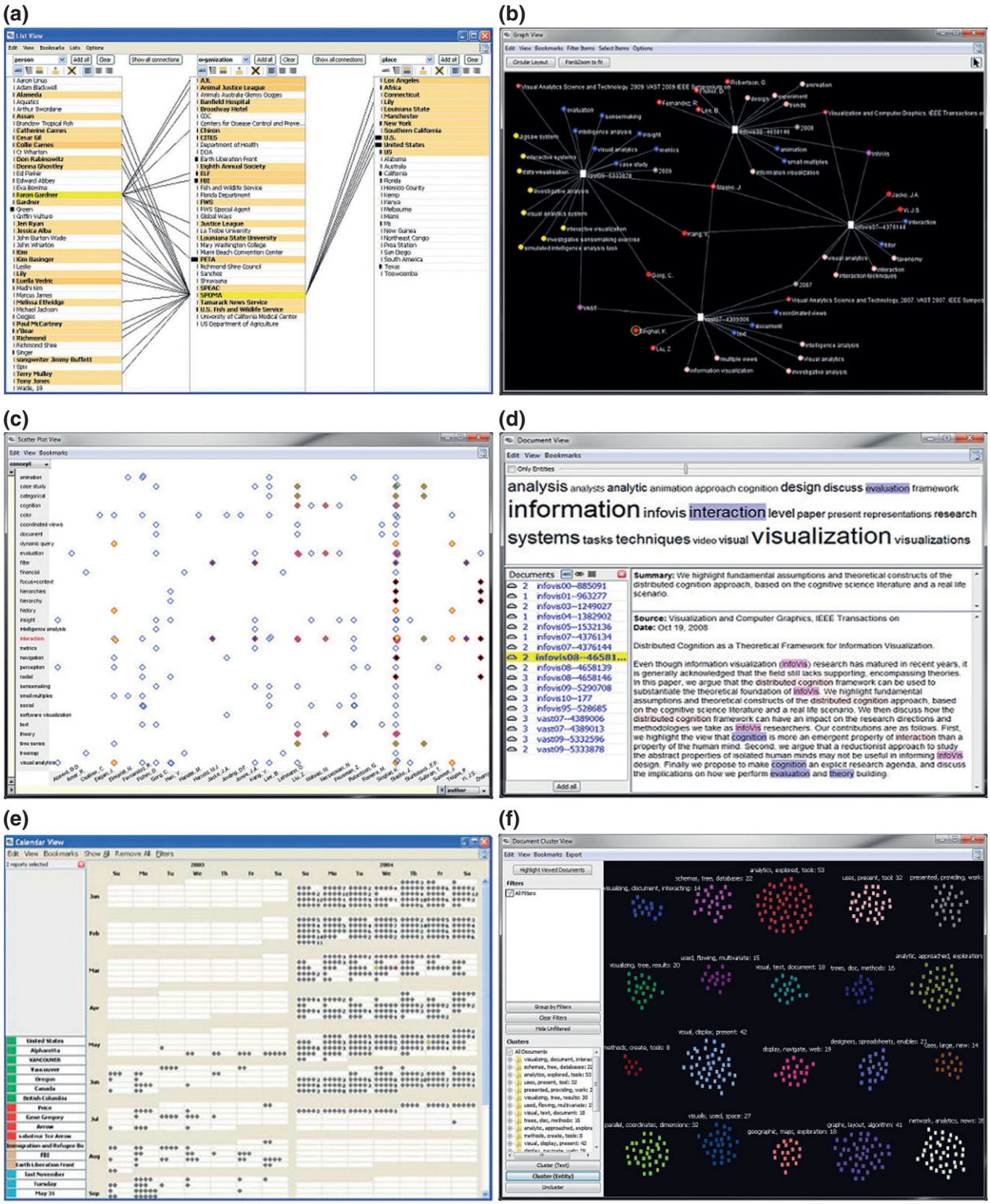


FIGURE 13 | Multiple views of Jigsaw. (Reprinted with permission for Ref 46. Copyright 2008 Palgrave Macmillan)

have narrow scope of application due to its pointed direction.

3. Interactivity It is important to design a more intuitive man–machine interface to improve user's experience of interaction. Also it is crucial to find some interstices to allow users to participate in researching and developing process, especially the testing period. Besides, as more and more attention has been paid in the field of *visual analytics*, visual text analytics is also significant and will become a trend. In *visual analytics*, the quite important factors are the involvement of human and interactive visual interfaces.

4. Techniques

- Algorithms. Develop and adopt scalable, high-performance algorithms for text processing, such as text summary and clustering.
- Parallel processing technology. With the adoption and popularity of Coordinated and Multiple Views (CMV), a visualization system usually includes multi-views. And with the rapid increase of document quantities, the main challenge is that we need to concurrently process data to guarantee each view updated synchronously. Apply and optimize parallel techniques to improve processing efficiency.
- Real-time processing technology. Real-time visualization of documents is crucial for different analysis scenarios and can be expected to become one of the important future research topics in the document visualization domain. Especially the complex requirements of real-time text analysis tasks lead to new visualization challenges which pose a great pressure on real-time processing techniques.[55]

Combination of Current Methods and Others

1. Various document visualization methods can be combined and optimized to get a more effective and comprehensive performance of document information. For example, TopicNets[28] associates topic modeling with graph visualization to represent topics and their relations at the same time. TIARA[34] combines Tag Clouds and ThemeRiver.
2. Integrating document visualization methods with other information visualization methods or even scientific visualization methods to visualize more information simultaneously. In practical projects, there are usually including various forms of data, such as multidimensional data, spatial data, text, etc. Utilizing various visual representations can balance the advantages and disadvantages of different methods, also help users solve problems from different perspectives. For example, PTC[10] utilizes the advantages of both Parallel Coordinates and Tag Clouds.
3. Visual data mining. Applying document visualization to the process of text data mining, the result could be visually represented and the analysis process could be monitored, thus to improve the shortcomings of low user interactivity in traditional data mining methods, and to improve the accuracy of results.

Application

1. Social Media With the increasing use of social media (Facebook, Twitter, and non-English equivalents such as Weibo), the age of *Big Data* is upon us. Such data is better modeled as a stream of small packets or messages, e.g., Facebook posts, cell phone messages, or Twitter tweets. Taking Twitter as an example, one estimate gave, on average, 3000 tweets/second in February 2012, a sixfold increase over the same month 2 years earlier.[61] Creating new visualizations or optimizing current document visualizations to make sense of massive volume of streaming text data is an active area of research.

2. Mobile Internet In recent years, with the popularity of smart phones and the increase of mobile Internet users, large scale of textual documents, such as Internet logs, are generated. Those data contains a great deal of information and its analysis has significant research value both in theory and practice. Existing researches have mainly focused on disaster recovery, storage management, access control, and data mining; while the typical cases of visualization systems are relatively few.

Evaluation and Theory

1. Evaluation Many document visualization methods or even information visualization methods lack a quantitative measurement which can indicate the overall quality, novelty, uncertainly, and other evaluative metrics. More recently, there exist more and more publications that reflect upon

current practices in visualization evaluation. In fact, the BELIV workshop was created as a venue for researchers and practitioners to 'explore novel evaluation methods, and to structure the knowledge on evaluation in information visualization around a schema'.[62]

2. Theoretical Foundations The 2007 Dagstuhl Workshop identified collaborative information visualization with theory building as major directions for future development.[63]

ACKNOWLEDGMENT

This work is a part of the joint research project with EMC Labs China (the Office of CTO, EMC Corporate), which is funded by EMC China Centre of Excellence.

REFERENCES

1. Ku C-H, Nguyen JH, Leroy G. TASC-Crime report visualization for investigative analysis: a case study. In: *IEEE 13th International Conference on Information Reuse and Integration*. Las Vegas, NV: IEEE; 2012, 466–473.
2. Buckland MK. What is a "document"? *J Am Soc Inf Sci* 1997, 48:804–809.
3. Ward M, Grinstein G, Keim D. *Interactive Data Visualization: Foundations, Techniques, and Applications*. Natick, MA: AK Peters, Ltd.; 2010, 291–292.
4. Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. In: *IEEE Symposium on Visual Languages*. Boulder, CO: IEEE; 1996, 336–343.
5. Chen H, Wang G, Peng D, Zuo W, Chen W. Sequential document visualization based on hierarchical parametric histogram curves. *Tsinghua Sci Technol* 2012, 17:409–418.
6. Viegas FB, Wattenberg M. Timelines tag clouds and the case for vernacular visualization. *Interactions* 2008, 15:49–52.
7. Tag Cloud home page. Available at: http://tagclouds.com/ (Accessed March 23, 2012)
8. Viegas FB, Wattenberg M, Feinberg J. Participatory visualization with wordle. *IEEE Trans Vis Comput Graph* 2009, 15:1137–1144.
9. Wordle home page. Available at: http://www.wordle.net/ (Accessed March 23, 2012)
10. Collins C, Viegas FB, Wattenberg M. Parallel tag clouds to explore and analyze faceted text corpora. In: *IEEE Symposium on Visual Analytics Science and Technology*. Atlantic City, NJ: IEEE; 2009, 91–98.
11. Koh K, Lee B, Kim B, Seo J. Maniwordle: providing flexible control over Wordle. *IEEE Trans Vis Comput Graph* 2010, 16:1190–1197.
12. Cui W, Wu Y, Liu S, Wei F, Zhou MX, Qu H. Context preserving dynamic word cloud visualization. In: *IEEE Pacific Visualization Symposium (PacificVis)*. Taipei: IEEE; 2010, 121–128.
13. Fabo P, Novotny M. Three-level visualization of internet discussion with extruded word clouds. In: *16th International Conference on Information Visualisation*. Montpellier: IEEE; 2012, 13–17.
14. Paley WB. *TextArc: Showing word frequency and distribution in text*. Boston, MA: IEEE Symposium on Information Visualization; 2002.
15. Chen C. Information visualization. *WIREs Comput Stat* 2010, 2:387–403.
16. Collins C, Carpendale S, Penn G. Docuburst: Visualizing document content using language structure. *Comput Graph Forum* 2009, 28:1039–1046.
17. Miller G, Fellbaum C. *Wordnet: An Electronic Lexical Database*. Cambridge, MA: MIT Press; 1998.
18. Evans V. Semantic structure versus conceptual structure: The nature of lexical concepts in a simulation-based account of language understanding. Unpublished technical report; 2009, 1–6.
19. Rusu D, Fortuna B, Mladenic D, Grobelnik M, Sipos R. Document visualization based on semantic graphs. In: *13th International Conference on Information Visualisation*. Barcelona: IEEE; 2009, 292–297.
20. Leskovec J, Grobelnik M, Milic-Frayling N. Learning sub-structures of document semantic graphs for document summarization. In: *Proceedings of the KDD 2004 Workshop on Link Analysis and Group Detection LinkKDD*, Seattle, WA, 2004.
21. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. *J Am Soc Inf Sci* 1990, 41:391–407.
22. Iwata T, Yamada T, Ueda N. Probabilistic latent semantic visualization: topic model for visualizing documents. In: *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM; 2008, 363–371.

23. Wattenberg M, Viegas FB. The word tree, an interactive visual concordance. *IEEE Trans Vis Comput Graph* 2008, 14:1221–1228.
24. IBM Many Eyes. Available at: http://www-958.ibm.com/software/data/cognos/manyeyes/ (Accessed March 31, 2012)
25. Wattenberg M. Arc diagrams: visualizing structure in strings. In: *IEEE Symposium on Information Visualization*. Washington, DC: IEEE; 2002, 110–116.
26. Thomas JJ, Cowley PJ, Kuchar O, Nowell LT, Thompson J, Wong PC. Discovering knowledge through visual analysis. *J Univ Comput Sci* 2001, 7:517–529.
27. Wong PC, Hetzler B, Posse C, Whiting M, Havre S, Cramer N, Shah A, Singhal M, Turner A, Thomas J. IN-SPIRE InfoVis 2004 contest entry. In: *IEEE Symposium on Information Visualization, 2004*. Austin, TX: IEEE; 2004, r2.
28. Gretarsson B, O'donovan J, Bostandjiev S, Höllerer T, Asuncion A, Newman D, Smyth P. Topicnets: visual analysis of large text corpora with topic modeling. *ACM Trans Intell Syst Technol* 2012, 3:1–26.
29. Havre S, Hetzler B, Nowell L. ThemeRiver: visualizing theme changes over time. In: *IEEE Symposium on Information Visualization*. Salt Lake City, UT: IEEE; 2000, 115–123.
30. Miller NE, Wong PC, Brewster M, Foote H. TOPIC ISLANDS-A wavelet-based text visualization system. In: *Proceedings of the 9th IEEE Visualization Conference*. Research Triangle Park, NC: IEEE; 1998, 189–196.
31. Liu S, Zhou MX, Pan S, et al. Interactive, topic-based visual text summarization and analysis. In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. Hong Kong: ACM; 2009, 543–552.
32. Shi L, Wei F, Liu S, Tan L, Lian X, Zhou MX. Understanding text corpora with multiple facets. In: *IEEE Symposium on Visual Analytics Science and Technology (VAST)*. Salt Lake City, UT: IEEE; 2010, 99–106.
33. Cui W, Liu S, Tan L, Shi C, Song Y, Gao Z, Tong X. TextFlow: towards better understanding of evolving topics in text. *IEEE Trans Vis Comput Graph* 2011, 17:2412–2421.
34. Liu S, Zhou MX, Pan S, Song Y, Qian W, Cai W, Lian X. TIARA: interactive, topic-based visual text summarization and analysis. *ACM Trans Intell Syst Technol* 2012, 3:1–28.
35. 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.
36. Viegas FB, Wattenberg M, Dave K. Studying cooperation and conflict between authors with history flow visualizations. In: *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*. Vienna: ACM; 2004, 575–582.
37. Diehl S. *Software Visualization*. Berlin, Heidelberg: Springer; 2007.
38. Storey M-AD, Čubranić D, German DM. On the use of visualization to support awareness of human activities in software development: a survey and a framework. In: *Proceedings of the 2005 ACM Symposium on Software Visualization*. New York: ACM; 2005, 193–202.
39. Diehl S. *Software Visualization: Visualizing the Structure, Behaviour, and Evolution of Software*. New York: Springer; 2007.
40. Eick S, Steffen JL, Sumner EE Jr. Seesoft-a tool for visualizing line oriented software statistics. *IEEE Trans Softw Eng* 1992, 18:957–968.
41. Froehlich J, Dourish P. Unifying artifacts and activities in a visual tool for distributed software development teams. In: *Proceedings of the 26th International Conference on Software Engineering*. Washington, DC: IEEE Computer Society; 2004, 387–396.
42. Eick SG, Graves TL, Karr AF, Mockus A, Schuster P. Visualizing software changes. *IEEE Trans Softw Eng* 2002, 28:396–412.
43. Wu X, Murray A, Storey M-A, Lintern R. A reverse engineering approach to support software maintenance: version control knowledge extraction. In: *11th Working Conference on Reverse Engineering*. Delft, The Netherlands: IEEE; 2004, 90–99.
44. Tu Q, Godfrey MW. An integrated approach for studying architectural evolution. In: *10th International Workshop on Program Comprehension*. Waterloo: IEEE; 2002, 127–136.
45. Wu J, Holt RC, Hassan AE. Exploring software evolution using spectrographs. In: *11th Working Conference on Reverse Engineering*. Waterloo: IEEE, Waterloo University; 2004, 80–89.
46. Stasko J, Görg C, Liu Z. Jigsaw: supporting investigative analysis through interactive visualization. *Inf Vis* 2008, 7:118–132.
47. Lin Y-R, Sun J, Cao N, Liu S. Contextour: Contextual contour visual analysis on dynamic multi-relational clustering. In: *SIAM Data Mining Conference*, Columbus, Ohio, 2010, 418–429.
48. 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.
49. Dörk M, Riche NH, Ramos G, Dumais S. PivotPaths: strolling through faceted information spaces. *IEEE Trans Vis Comput Graph* 2012, 18:2709–2718.
50. Kohonen T. The self-organizing map. *Proc IEEE* 1990, 78:1464–1480.
51. Lin X. Visualization for the document space. In: *IEEE Conference on Visualization*. Boston, MA: IEEE; 1992 274–281.

52. Lin X. Map displays for information retrieval. *JASIS* 1997, 48:40–54.

53. WEBSOM Research Group. WEBSOM Map-Million Documents. Available at: http://websom.hut.fi/websom/milliondemo/html/root.html (Accessed March 31, 2012)

54. Gansner ER, Hu Y, North S. Visualizing streaming text data with dynamic graphs and maps. In: *Graph Drawing*. Redmond, WA: Springer; 2013, 439–450.

55. Rohrdantz C, Oelke D, Krstajic M, Fischer F. Real-time visualization of streaming text data: tasks and challenges. In: *IEEE VisWeek Workshop on Interactive Visual Text Analytics for Decision-Making*, Providence, RI, 2011.

56. Alsakran J, Chen Y, Zhao Y, Yang J, Luo D. STREAMIT: dynamic visualization and interactive exploration of text streams. In: *IEEE Pacific Visualization Symposium (PacificVis)*. Hong Kong: IEEE; 2011, 131–138.

57. Cao N, Lin Y-R, Sun X, Lazer D, Liu S, Qu H. Whisper: tracing the spatiotemporal process of information diffusion in real time. *IEEE Trans Vis Comput Graph* 2012, 18:2649–2658.

58. Hearst MA. TileBars: visualization of term distribution information in full text information access. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM Press/Addison-Wesley Publishing Co.; 1995, 59–66.

59. Havre S, Hetzler E, Perrine K, Jurrus E, Miller N. Interactive visualization of multiple query results. In: *Proceedings of the IEEE Symposium on Information Visualization*, Citeseer, 2001, 105–112.

60. Spoerri A. RankSpiral: toward enhancing search results visualizations. In: *IEEE Symposium on Information Visualization*. Austin, TX: IEEE; 2004, 18–19.

61. Gansner E, Hu Y, North S. Interactive Visualization of streaming text data with dynamic maps. *Journal of Graph Algorithms and Applications* 2013, 17:515–540.

62. Huang W, Eades P, Hong S-H. Beyond time and error: a cognitive approach to the evaluation of graph drawings. In: *Proceedings of the 2008 Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*. Florence, Italy: ACM; 2008, 3.

63. Kerren A, Stasko J, Fekete J. Workshop report: information visualization-human-centred issues in visual representation, interaction, and evaluation. *Inf Vis* 2007, 6:189–196.

FURTHER READING

Ware C. *Visual Thinking for Design*. Burlington: Morgan Kaufmann; 2008.

Thomas J, Cook K. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Los Alamitos: IEEE Computer Society; 2005.

North C. Towards measuring visualization insight. *IEEE Comput Graph Appl* 2006, 26:6–9.

Chen C. *Information Visualization: Beyond the Horizon*. Heidelberg: Springer; 2004.

