J Vis (2018) 21:321–335
https://doi.org/10.1007/s12650-017-0462-2

REGULAR PAPER

Changhong Zhang · Zeyu Li · Jiawan Zhang

A survey on visualization for scientific literature topics

Received: 31 August 2017 / Accepted: 5 November 2017 / Published online: 7 December 2017

Abstract The topics in scientific literature illustrate the contents of science domain, and the evolution of topics help in recognizing the research trend and front. Since the number of scientific works is growing exponentially, it is a great challenge for people to discover new research topics and topic changes. Fortunately, aided by text mining, visualization technologies are being widely used for topic analysis. Visualization is an effective tool for revealing the current status and topic evolution trend in a research field. Owing to the importance of topic analysis as well as the lack of a comprehensive description of this theme, we present a survey on the visualization methods for scientific literature topics. This paper introduces the basic concepts of bibliometrics and the pipeline of topic visualization. Based on the topic analysis tasks, we classify these papers into three categories: topic contents, topic relation, and topic evolution. Furthermore, each part is divided into smaller categories on the basis of the visual patterns. Some existing free software that integrates multiple functions are also introduced. Finally, we discuss the challenges and opportunities in the field of topic visualization.

Keywords Visualization · Scientific literature · Survey · Topic evolution · Topic

1 Introduction

The topics in scientific literature illustrate the contents of science domain, and its evolution helps in recognizing the research trend and front. Different users have different requirements; for example, a student who is a freshman in one field needs to find the key authors, pivotal papers, and open questions for starting the study, whereas an experienced researcher may be interested in the recent leading work and potential research direction in his or her field. Managers in universities or in the government expect to know the research trends and emerging fields for the purpose of development and investigation. In short, it is important to discover the main research topics, emergent topics, research frontiers, and topic evolution in a particular knowledge domain.

C. Zhang · Z. Li · J. Zhang (✉)
School of Computer Software, Tianjin University, Tianjin, China
E-mail: jwzhang@tju.edu.cn

C. Zhang
E-mail: ch_zhang@tju.edu.cn

Z. Li
E-mail: lzytianda@tju.edu.cn

C. Zhang
School of Computer Science and Technology, Qinghai University for Nationalities, Xining, China

Due to the exponential growth of scientific literature and the rapid emergence of interdisciplinary studies, it is a great challenge for people to master the current topics and discover the new research topics in one field. To gain insights into scientific literature efficiently, researchers have designed and exploited many techniques to collect, analyze, and manage this kind of data. These efforts include text mining, social network technology, and visualization. By combining the human intelligence and data analysis techniques, data visualization has been proved to be an effective tool in scientific literature topic analysis: for instance, in providing intuitive overviews on publication and in discovering the patterns of topic evolution through interaction.

The scientific literature referred to in this paper is limited to academic works including papers, Ph.D. dissertations published in journals, and proceedings that are accepted after a strict peer review. Scientific literature data comprise full text and bibliographic records. Bibliographic records mostly come from large commercial literature databases [such as the Web of Science (WoS), Scopus, ScienceDirect, etc.] and open network databases (such as Google Scholar, CiteSeerX, etc.). Bibliographic records include elements such as title, abstract, keywords, author information, references, publication name, and address. The full text data is stored in pdf, XML, or HTML format. PlosOne, which is a well-known open access journal, provides full-text download service in PDF and XML formats. The world-renowned academic databases (such as Springer, Elsevier, and Wiley) also provide full or partial text download in XML and HTML formats. The textual content of a publication is its central component, as it encodes its main scientific contribution. Therefore, topics are typically extracted from the textual content of papers. Citations are also an essential part of scientific papers and can be considered as a measure of the influence of papers. It can be used for tracking the dissemination of new concepts and identifying knowledge flows (Federico et al. 2017).

There have been several related studies that introduce the visual approaches for analyzing scientific literature, focusing on various aspects. Chen (2013) provides the basic visual theory for mapping scientific fronts and proposes visual methods for citation analysis and co-words analysis. It focuses on the collaborations among authors and scientific communities, which were detected by making use of citation data. On the contrary, we focus more on textual contents. Brner (2010) focuses on a science map, creating a spatial representation of scientific communities based on citation and text data. It depicts the structure or landscape of science. We focus on comparing various visualization methods for revealing the topic and the topic evolution of some domains. Alencar et al. (2012a) provides an overview of visual text analysis, focusing on all kind of texts, whereas we consider only scientific literature. Another comprehensive collection of text visual approaches (Kucher and Kerren 2015) is available online.[1] Federico et al. (2016) provides an overview of the visualization approaches for scientific documents and patents; these approaches are classified into two levels according to the data type and abstract task. Instead of using an abstract task, our work utilizes task categories that are more specific.

Our work aims to summarize the state-of-the-art visual methods for the topic analysis of scientific literature. In this way, we are making up for the lack of a comprehensive study in this research field. We collected 38 papers from the field of visual community and bibliometrics, with a focus on papers published after 2010. There are common analytical requirements in both the areas; therefore, we derive the analysis task that is used as the classification scheme in this paper. The methods are classified according to the analysis task, including topic content, topic relation, and topic evolution. Based on the visual pattern, we further divide each part into more detailed categories as the second level. However, some of the softwares that are widely used in bibliometrics for integrating many functions cannot be simply classified as one of the above; therefore, we introduce them in a single section.

The rest of this paper is organized as follows. A background of topic extraction and visualization pipeline is provided in Sect. 2. The involved state-of-the-art topic visual approaches based on our two-level classification scheme are presented in Sect. 3. Section 4 describes several integrated free tools, which are widely used in bibliometrics. In Sect. 5, we discuss the challenges and opportunities in the field of topic analysis and visualization. Finally, we conclude this paper in Sect. 6.

2 Background

Since the current works reviewed in this paper come from both the bibliometrics and visualization community, it would be necessary to introduce some basic concepts and general techniques of both the domains.

[1] http://textvis.lnu.se/.

2.1 Basic concepts

The methods for obtaining topics can be roughly divided into two categories: citation based and content based. Using the citation-based method, we can extract the topic and quickly find the key literature. Through the content-based method, we can obtain the topic more directly; however, we cannot find the key literature.

In citation-based methods, the documents are usually first aggregated into a network, where vertexes represent documents and edges represent the relation between them. Document clustering is executed after constructing the network, and then the topics are extracted from each cluster, for instance, by simply counting the word frequency. There are three commonly applied citation analysis models for constructing the network: direct citation, co-citation, and bibliography coupling (Yan and Ding 2012). As the latter two are indirect relations, they can be derived from the direct citation. In direct citation, the relationship represents the direct reference between two documents. Bibliographic coupling is defined as the relation between two papers that cite one or more common papers in their bibliographies. Co-citation is defined as the number of common documents that cite two given documents. The more number of times they are cited by the same documents, the more likely they are semantically related (Small 1973).

The content-based methods can be further classified into co-word analysis and topic model analysis. The word occurring in the same document share a co-word relation, and a co-word network is created, followed by clustering, similar to that in the case of citation network. Topic model analysis has several ways to obtain topics, such as vector space model, latent semantic analysis (LSA), and probability topic model. However, co-word analysis and vector space model have a limitation: they cannot handle synonymy and polysemy problems. LSA can find synonymy; however, it cannot capture polysemy, and as a result the interpretation of results is difficult. The probability topic model (Blei 2012) is capable of discovering and annotating a large number of documents with thematic information. With the rapid development of the probability topic model, its extension and variation have attracted considerable attention from researchers.

The characteristics of the two aforementioned kinds of topic extraction techniques are obvious. Firstly, the reason and treatment of noise is different. In citation analysis, the motivations for citing a prior work may be different. However, negative citations and the difference in motivation, to the degree in which they appear, do not materially affect our analysis. Hence, noise is usually overlooked because of its rarity in citation analysis. In contrast, the noise in content analysis is mainly caused by synonymy, polysemy, and the morphology problem, which is much more complicated. These noises can be eliminated by preprocessing through natural language processing techniques; however, for now, it is still a difficult problem to overcome perfectly. The second difference is that citation-based methods can intuitively identify the key and influential paper or author according to their cited times; however, further steps are required to obtain the topic words, which are directly given by the content-based methods. Ding and Chen (2014) compared the performance of three methods for topic detection and tracking: co-citation, co-word, and hierarchical Dirichlet process (HDP). The result shows that HDP is more sensitive and reliable than the other two methods for detecting and tracking emerging topics.

2.2 Topic visualization pipeline

The pipeline for creating a topic visualization system usually contains three steps: data model construction, visual design, and visualization implementation and verification. To construct the data model, the designer first specifies the requirements to consider what kind of information he or she wants to reveal. Based on the requirements, the designer filters the raw data and retains the data that may be useful for gaining insights. The data model is designed for transforming the raw data into cleaner and structural data, which can be associated with each other and can be easily used for visual encoding. Text processing and mining techniques like tagging, text clustering, dimension reduction, and topic modeling are applied in this stage. The next stage is visual design, which is important as it determines the effectiveness of the data representation. Basic graphics like dot, curve, and circle are used for visual mapping, and layouts like ring, flow, matrix, map, and animation should be elaborately designed. Interaction plays an increasingly important role in meeting the need for exploration. Therefore, it is essential to verify whether the developed visualization is effective and satisfies the initial requirements. User study or case study is widely used for this verification.

3 Visualization of topic

Obtaining the correct topic, understanding the topics, building a relationship between topics, and discovering the temporal evolution are crucial tasks for the topic analysis of scientific literature. The visualization of a topic helps in providing an overview of the content of the papers and identifies the hidden knowledge and pattern of the papers. According to the topic analysis task of scientific literature, we divide these papers into three categories, namely, topic content, topic relation, and topic evolution. Each category is divided into several groups according to mainly the visual pattern. The classification results can be seen in Fig. 1.

3.1 Topic content

Topic model selection, topic operation, and topic representation are the main tasks of topic content analysis, which help in understanding the text collection. Topic model selection is used for obtaining the appropriate topic model or the best parameter of the model for obtaining the correct clusters. The aim of topic operation is to obtain the correct clusters by removing unimportant clusters and merging similar clusters. Topic representation displays the content of each cluster based on key words list, word cloud, or key word label. There are several main visual patterns for this task. The visual techniques are organized into four categories: node-link diagram, scatter plot, matrix view, and other methods.

3.1.1 Node-link diagram

To remove noisy features and outliers in the automatic clustering process, Lee et al. (2012) proposed the clustering method, iVisClustering, which combines the latent Dirichlet allocation (LDA) model with visual interaction. The main view uses a force-directed layout of a node-link graph of the documents (see Fig. 2a). Each node represents a paper,the link between the nodes represents the similarity of the papers, and each color represents a topic. The users can create the tree structure according to their requirement. The system supports cluster-level interactions such as sub-clustering, removing unimportant clusters, merging the clusters that have similar meanings, and moving certain clusters to any other node in the tree structure. In addition, users can move mis-clustered documents to another cluster or remove useless documents.

Visual pattern	Topic content	Topic relation	Topic evolution
Node link diagram	[Lee et al(2012), Maiya and Rolfe(2014)]	[Gretarsson et al(2012), Cao et al(2010)]	[Morris et al(2003), Cobo et al(2011), Chuang et al(2012b), Janssens and Moor(2007), Mane and Brner(2004), Alencar et al(2012b), Hascot and Dragicevic(2011)]
Scatter plot	[Wise(1999), Choo et al(2013)]		
Matrix view	[Alexander and Gleicher(2015), Chuang et al(2012a)]		
Tree		[Dou et al(2013), Wang et al(2016), Jiang and Zhang(2016)]	
Map		[Davidson et al(1998), Kohonen et al(1999), Skupin(2002), Skupin(2004), Oesterling et al(2010), Fried and Kobourov(2013)]	
Parallel coordinate		[Dou et al(2011), Collins et al(2009)]	
Stream graph			[Havre et al(2000), Wei et al(2010), Dou et al(2013), Dou et al(2011), Jiang and zhang(2016), Heimerl et al(2016), Wang et al(2014), Alexander and Gleicher(2015), Gad et al(2015)]
Animation			[Alsakran et al(2012)]
Other	[Murdock and Allen(2015), Wu et al(2011),Chaney and Blei(2012)]	[Oelke et al(2014), Keena et al(2016),Chen and Paul(2001)]	[Lee et al(2005)]

Task

Fig. 1 Categories of visualization methods used for scientific literature topics

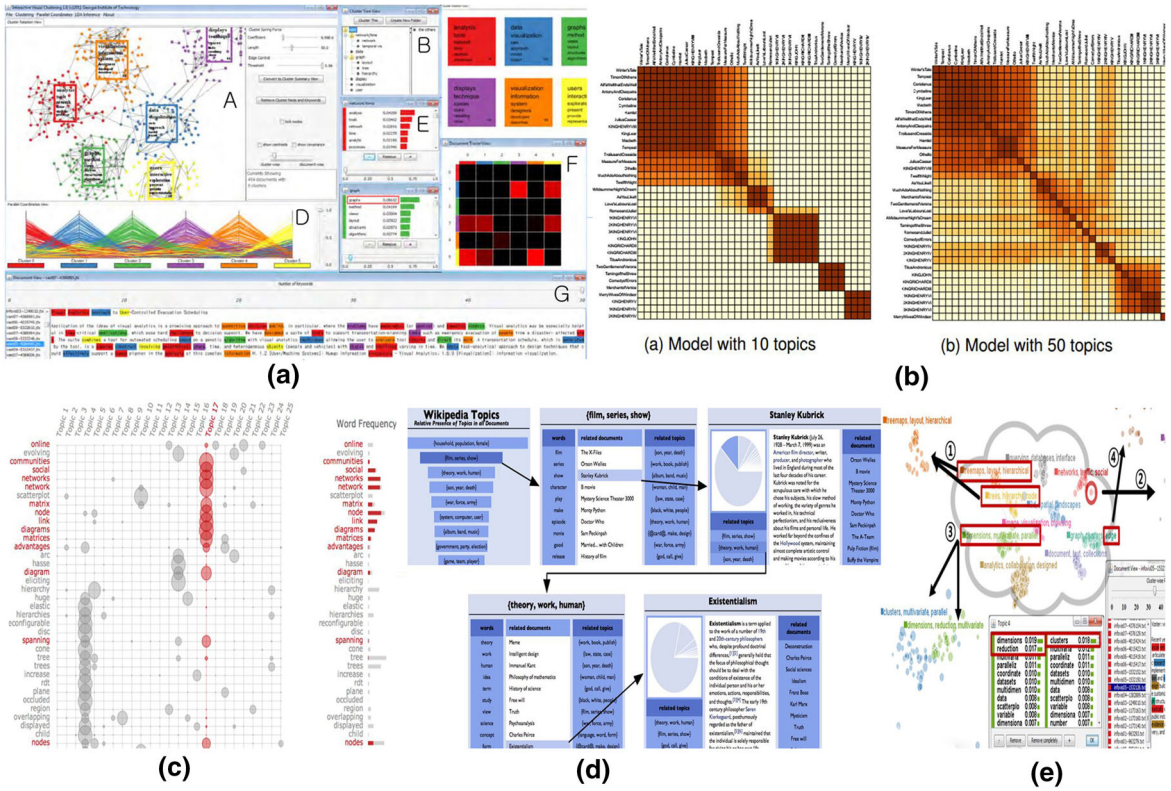

Fig. 2 **a** Node-link diagram (in Lee et al. 2012); **b** matrix view (in Alexander and Gleicher 2016); **c** term–topic matrix (in Chuang et al. 2012a); **d** navigator of documents (in Chaney and Blei 2012); **e** UTOPIAN (in Choo et al. 2013)

To improve the interpretability of LDA topic models in large document sets, Maiya and Rolfe (2014) constructed a topic similarity network with labeling. Each node represents a topic, and the color and size of the node indicate the affiliation and strength of the topic, respectively.

In summary, the above two papers aim to understand the topic model: LDA. They use node-link diagrams to reflect the relationships of documents and topics. Maiya and Rolfe (2014) shows the relationship between topics and the same color represents topic affiliation. Moreover, the size of the data is very different; the latter can handle large document sets.

3.1.2 Scatter plot

The galaxy view projects the articles according to their similarity in the 2-D plane (Wise 1999). A dense point cluster represents documents with the same topic, and a more dense point indicates that there are more documents on this topic.

To solve the uncertainty in topic representation, Choo et al. (2013) presented a system called UTOPIAN, which maps the clustering documents on 2D scatter plots. In this plot (see Fig. 2e), the clusters are labeled with keywords, and the users can interact with the topic model algorithm. Possible interactions comprise topic refinement by modifying the weight of keywords in a topic, merging of topics that are similar, splitting of topics, and constructing topics based on documents or keywords selected by the users. They verified their technique using papers from InfoVis and VAST.

To sum up, these papers are different in the manner of topic representation and how they choose the topic number. The number of topics is automatically determined in Wise (1999), and the number of topics grows as the size of paper collection increases. In the method proposed by Choo et al. (2013), the users need to determine the number of topics and, subsequently, can merge or split topics by interaction. In addition, the latter uses color coding for different topics.

3.1.3 Matrix view

To select a wise topic model, Alexander and Gleicher (2016) used the visualization method to compare the differences between topic models based on three aspects: topic content, topic similarity, and topic evolution. To compare the topic content, they employed matrix and parallel coordinates to find the differences between topic models for the same articles (see Fig. 2b). In the figure, it can be seen that the clustering effect of the document set becomes different when the number of themes changes.The buddy plot is used to explore the difference in similarity between articles in different topic models.

Chuang et al. (2012a) presented the Termite system, which uses a tabular layout term–topic matrix, as shown in Fig. 2c. The radius of the circle represents the term weights in a topic. It provides a saliency measure for selecting the relevant terms, as well as a series of algorithms that can be used to reveal the clustering structure and improve the legibility of related terms. As shown in the figure, the user can quickly find the topic of the keywords and the overlapping relationship between the topics.

To sum up, the methods proposed in these two papers perform different tasks using a matrix view of the topic. Alexander and Gleicher (2016) estimated the number of clusters by using the document matrix in different topic models. Chuang et al. (2012a) used the term–topic matrix to represent the relationship between the topics and discover the key terms of a topic.

3.1.4 Other methods

In the study by Murdock and Allen (2015), the topic explorer is used to recognize document–document and topic–document relations of the LDA topic models. The users compare the distinctions of document–document and topic–document relationships for different topic numbers. The users are able to visualize the topic terms and find the closest document according to the rank list.

Word cloud is often used to represent the topic content. Wu et al. (2011) presented a new algorithm for creating semantic-preserving and compact word clouds using an adapted seam carving technique.

Chaney and Blei (2012) proposed a method in which a navigator is created for the topics. This allows users to explore the hidden structure discovered by the topic model. An example is depicted in Fig. 2d. These browsing interfaces can help end users explore and understand the topics and their distribution in one document collection.

These papers reveal the content of the topic and the relationship between the topic and document in their own way.

3.2 Topic relationship

Topic relationship mainly includes the relationship between topics, and the relationship between the topic and paper. The relationship between the topics includes the similarity of the topic, the hierarchical relationship, and the common topic of different research domains.The relationship between the document and topic reveals the topic distribution of the document. The visual techniques use to represent and understand large scientific literature are organized into five categories: tree, map, parallel coordinate, node-link diagram, and other methods.

3.2.1 Tree

Tree is a popular visualization method for hierarchical structure data. It enables users to quickly understand the topic relationship. In a traditional distance-based agglomerative clustering tree, many topics in the same level are impertinently laid out at different levels. This is not consistent with people's understanding. In the hierarchical LDA model, the topics in the top layer are mainly composed of some meaningless words, such as the stop word. Blundell et al. (2010) proposed the Bayesian rose tree (BRT) algorithm, which produces a tree with a multi-branching structure at each node. The BRT model can better explain the data and their relationship. Liu et al. (2015) presented the evolutionary Bayesian rose tree (EBRT), which is an evolutionary multi-branch tree that models hierarchical topics and their evolutionary patterns for all time.

To analyze the text corpora, which includes a large number of topics, a hierarchical structure was proposed, and it is shown in Fig. 3a. Dou et al. (2013) introduced a topic rose tree, hierarchical topics. In this method, the users can iteratively modify the initial topic tree based on their interest. This leads to a

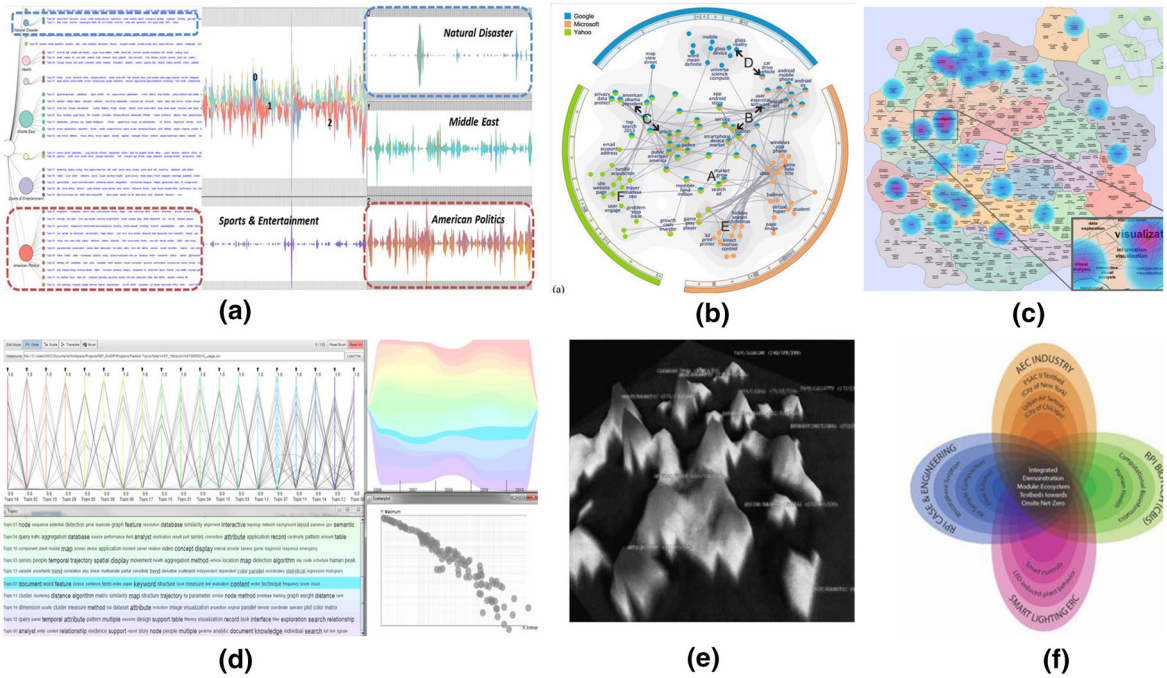

Fig. 3 **a** HierarchicalTopics (in Dou et al. 2013); **b** TopicPanorama (in Wang et al. 2016); **c** computer science map (in Fried and Kobourov 2013); **d** parallel coordinate (in Dou et al. 2011); **e** ThemeScape (in Davidson et al. 1998); **f** multi disciplinary (in Keena et al. 2016)

better explanation of the topics groups. The initial topic tree is created through the topic rose tree (TRT) algorithm. The core idea of TRT comes from BRT.

Wang et al. (2016) introduced the TopicPanorama. It displays the common and distinctive topics of each data source in a hierarchical fashion. This system is shown in Fig. 3b. The topic tree is built by applying the constraint-based BRT model in which each non-leaf node represents a topic cluster. A radical icicle plot is employed to display the topic hierarchies. Its visualization graph combines a node-link diagram with a density map.

Jiang and Zhang (2016) employed a hierarchical topic model (CATHY model) to construct a topic tree with a hierarchical structure. The treemap simultaneously displays the second-level topic relationship among the three related research fields.

According to the above, the topic relationship was visualized using a hierarchical structure for various analysis tasks. The users can interactively merge the topics to form a hierarchical structure of topic according to the interest of the users in Dou et al. (2013). Simultaneously, it uses ThemeRiver to show the intensity change of the topic. In Wang et al. (2016), the common topics in various levels and the portion of three data sources are shown; however, it does not show the topic evolution. In Jiang and Zhang (2016), the common topic of three cross domains and the topic evolution over time are shown.

3.2.2 Map

Map is a popular visualization metaphor for scientific disciplines or publications. The map utilizes people's cognitive skills in geography to help them understand the field of science. It can be a good expression of the relationship between topics. However, the development of the field of science is not balanced, and topics do not have a clear boundary, which poses a considerable challenge for mapping nongeographic information.

Davidson et al. (1998) proposed ThemeScape shown in Fig. 3e, which is an improvement of the galaxy view. It is based on the document projection position and uses contours to represent the distribution of similar documents. The density of document distribution is mapped to the height of the mountain, and the contours and encoding colors are used to depict the density of text distribution. If the documents are more similar, then the contours appear closer to each other and the colors are more significant.

To organize large document collections according to textual similarity, Kohonen et al. (2000) presents a scalable map to map the large number text collections. This approach uses the specifying keyword to feature the text set.

To visualize information in a low-dimensional display space using map metaphor, Skupin (2002) created a map based on the contents of conference abstracts via automation. Skupin (2004) presented an entire pipeline into which he plugged various other clustering methods and discussed some principles of visualization methodology for visualizing a knowledge domain using cartographic means.

To overcome the drawback of having to measure cluster distances in very high-dimensional spaces, Oesterling et al. (2010) combined the topic model with density estimation to extract the cluster structure from a document set. It provides an improved two-stage method for topology-based projections from the original high-dimensional information space into both two-dimensional (2-D) and three-dimensional (3-D) visualizations. The 3D visualization represents document clusters as hills or islands on a plane, depending on the users' preference.

Fried and Kobourov (2013) created a map for computer science based on titles from the DBLP database, as shown in Fig. 3c. Multi-word terms are selected based on their importance in the titles and are then positioned and clustered according to the co-occurrence similarities between terms. A specific journal or conference was used to generate the base map, and then the heat map overlays was used to show the evolution of research topics in the field over the years. The profiles of researchers, research institutions, or conferences can be visualized using the heat map overlay of a journal base map.

According to the above, ThemeScape and its improvements can reflect the distribution and intensity of topics, as well as display the topic relationship in Davidson et al. (1998), Kohonen et al. (2000), Skupin (2002, 2004) and Oesterling et al. (2010). In Fried and Kobourov (2013), the base map overlay of the heat map can better reflect the main research topics and research frontier.

3.2.3 Parallel coordinate

To help users understand large document collections, Dou et al. (2011) designed ParallelTopics, which integrates the probabilistic topic model LDA with interactive visualization. An example of this is depicted in Fig. 3d. The parallel coordinate view gives a detailed account of the topic distribution for selected documents. The word clouds represent the content of each topic and the scatter plot displays the distribution of the number of topics in each document.

Collins et al. (2009) designed the parallel tag clouds, which provides overviews of the different facets of a document collection. The system visualization technique combines parallel coordinates with traditional tag clouds. Each column of parallel coordinates represents the text information of a certain facet of interest to the user. Each column of the tag cloud visualizes the content of the facet, and the link shows the distribution of the keywords that are of interest to the user for different facets.

To sum up, the methods proposed in these two papers are capable of discovering the topic relation and topic evolution. However, it is more intuitive in Dou et al. (2011), since the users can find the distribution of topics for selected literature and similar topics. However, users can comprehensively explore the document collection from multiple facets in Collins et al. (2009). However, they do not reveal the relationship between large document sets and topics.

3.2.4 Node-link diagram

Gretarsson et al. (2012) presented TopicNets, a Web-based topic exploration system, which combines fast topic modeling with graph-based interactive visualization for large sets of documents. It uses a node-link diagram for visualizing the document and topic for various understanding tasks, such as interdisciplinarity of the domain, the relation between the publication, and university department.

To display the different relations based on the different facets of the document, Cao et al. (2010) introduced a visualization technique, FacetAtlas, which combines search technology with advanced visual analytical tools to convey both global and local patterns simultaneously. Global relations are displayed using a density map, and local relations are conveyed using compound nodes and edge bundling techniques. In the multifaceted graph, entities are represented by circles, and their facets are represented by different colors. The internal relations are encoded using links between corresponding facet nodes of two different compound nodes. The external relations are encoded implicitly through the construction of compound nodes.

To sum up, different relationships were displayed using the node-link diagram. For visualizing the document and the topic used for various understanding tasks in Gretarsson et al. (2012), different colors are used to represent different scientific domains. Multifaceted visualization of the entity relation of documents displays different facets of a specific disease in Cao et al. (2010); the different colors represent different facets.

3.2.5 Other methods

To detect and explore discriminative and common topics coming from several sources, Oelke et al. (2014) presented DiTop-View, which combines automatic extraction of topics with visualization. They split up a rectangular area into three subareas that represent three sources. Circular document coins that contain the word clouds of topics are placed within or at the border of the areas, to specify the affinity of a topic to one or multiple topics. The authors compared the topics of papers from InfoVis, SciVis, and Siggraph to verify their approach.

To understand the possible relationships between multidisciplinary topics, Keena et al. (2016) presented a trial encompassing an interdisciplinary research center collaborators, experiments, and results, and represented them simultaneously using a high-resolution visualization. Figure 3f shows the visualization of the relationship between multidisciplinary topics. The best method to visualize the multivalent parameters of interdisciplinary topics are discussed in this paper.

Chen and Paul (2001) proposed an approach to visualize the intellectual structure of a domain based on co-citation. It was represented in a 3D knowledge landscape. Pathfinder and factor analysis were integrated to visualize the specialties in the underlying knowledge domain.

In summary, some of the visualization methods for understanding multidisciplinary topics were discussed. A topic belonging to three domains was visualized for finding the common topic and topic distribution in Oelke et al. (2014). For understanding the possible relationship of multidisciplinary topics, the authors discussed data visualization in Keena et al. (2016). In Chen and Paul (2001), the underlying knowledge domain was represented using 3D knowledge landscape.

3.3 Topic evolution

Topic evolution is an important task for analyzing the topic trend, discovering emergent topics, and identifying the research front. The topic trend indicates the change in topic content and topic strength.The discovery of emergent topics includes the determination of the time and the analysis of the possible causes for the emergence of the topic. The research front needs to be found to find out the current important research directions in each topic. We organized the visual techniques into four categories: streamgraph, node-link graph, and other methods.

3.3.1 Streamgraph

Stream is a familiar visual metaphor for topic evolution, which is easily understood and is widely used in many fields. When the number of topics is more than a dozen, the stream chart becomes more chaotic. Therefore, the visualization of large-scale topic evolution is a great challenge.

ThemeRiver (Havre et al. 2000) depicts the thematic variation of large-scale text collections over time. At any point in time, the width of the river indicates the collective strength of the selected themes. Color encoding is used to represent individual themes. The flow from left to right is interpreted as the change in topic through time. TIARA (Wei et al. 2010) combined ThemeRiver with the word cloud to describe the change in topics more specifically, as shown in Fig. 4b.

Dou et al. (2011) utilized ThemeRiver to depict the topic evolution over time. Similar topics appear next to each other in the visualization. Dou et al. (2013) introduced a hierarchical Themeriver, which views the temporal pattern of topics in a hierarchical manner over time. The users can modify the hierarchy of the topic based on their analysis objectives.

To help users obtain an overview and find the cross-domain relationship, Jiang and Zhang (2016) used Sankey diagram to depict the topic evolution of the related science fields and discover the common topics, as shown in Fig. 4a. Meanwhile, scatter plot depicts the topic relationship among related science fields, and word clouds are employed to represent the topic content.

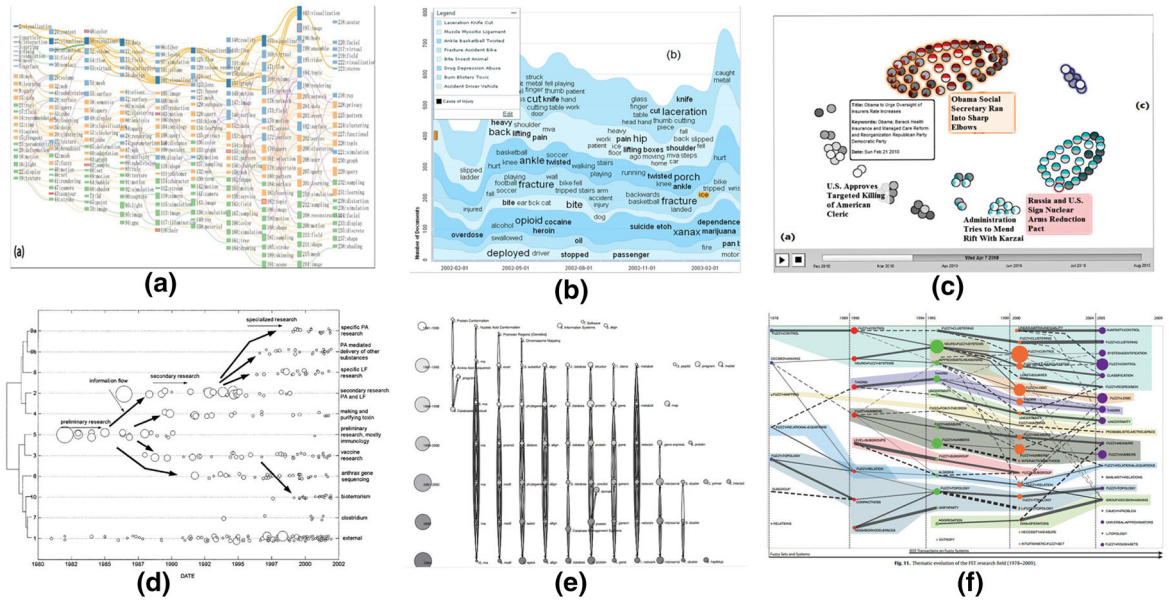

Fig. 4 **a** Sankey diagram (in Jiang and Zhang 2016); **b** Themestream (in Wei et al. 2010); **c** STREAMIT (in Alsakran et al. 2012); **d** time line (in Morris et al. 2003); **e** bioinformatics evolution (in Janssens and Moor 2007); **f** topic evolution of the Fuzzy Sets Theory (in Cobo et al. 2011)

For analyzing and exploring scientific literature, Heimerl et al. (2016) designed a new visual analysis system (CiteRivers), which shows the topic content of the collection, the reference patterns, and the key authors. The authors combined streamgraph with word clouds for depicting each topic evolution. The topic number can be adjusted using the level slider. The users can explore the cited conferences or journals from the topic block and citation pattern.

To reflect the topic evolution, Gad et al. (2015) designed the ThemeDelta system, using a time division algorithm to extract the characteristics of the trend changes. The system uses keywords to represent the topic, color coding to represent the topic of the category, the width of the line to depict the weight of the keyword, and curves with sharp angles to display the beginning and end of the topic. Similar topics are laid out together.

To better understand topic evolution, Wang et al. (2014) designed the Netviewer software, which discovers topic evolution based on dynamic co-word networks and community discover techniques. It visualizes the evolution of topics at the macro and micro level, respectively, using alluvial diagrams and coloring networks. Alluvial diagrams, which originated from geography, are used to map the evolution of the network. The colored rectangle areas represent the communities and their sizes; the colored curve areas between two time stamps depict the evolution process. There are two types of coloring algorithms: the forward coloring algorithm and the backward coloring algorithm. The forward coloring helps reveal the trend of each node in a community, and the backward coloring helps reveal the source of each node in a community.

To select a reasonable topic model, Alexander and Gleicher (2016) used the visualization method to compare the differences between the topic models. The topic evolution comparison uses Themeriver plot to find out the effect of different models on topic changes.

According to the above, these papers visualize the topic evolution mainly based on the ThemeRiver in Havre et al. (2000), Wei et al. (2010) and Dou et al. (2013, 2011). The change in topic can be discovered in Jiang and Zhang (2016); however, the change in intensity of the topic cannot be discovered. In Heimerl et al. (2016), the topic evolution combines the ThemeRiver with the word clouds, and the users can discover the cited pattern of the cited journal. In Wang et al. (2014), alluvial diagram and coloring network are used to show topic merge and split, and the users can quickly trace the topic evolution.

3.3.2 Node-link diagram

Mane and Brner (2004) visualized the papers published in PNAS in the years 1982–2001. This method extracts important terms from the title and keywords utilizing the Kleinberg's burst detection algorithm, and uses co-word analysis and graph layout technique to generate maps. The users can quickly identify the major research topic and trend. Six domain experts examined and commented on the resulting map. In the co-word space, the size of the node circle represents the maximum burst level of a word, and color coding is used to indicate the year with the maximum burst. The edge thickness is proportional to the number of word co-occurrences.

From the bibliographic coupling filtering data, Morris et al. (2003) computed the similarity for each pair of documents, laid out the topic using agglomerative hierarchical clustering, drew clusters of the dendrogram in the *y*-axis and documents in the *x*-axis for tracking the publication data, and extracted cluster topics from the document titles of each cluster. Figure 4d shows the topic evolution. Each node represents a document, and its radius represents the total cited number of a document. The users can quickly find the time of emergence of a topic and identify the seminal papers and possible emerging research fronts.

Cobo et al. (2011) presented an approach that quantifies and visualizes the topic evolution of a research field (see Fig. 4f). Co-word analysis is used to detect the different topics of the research field over time. Each column represents the topics of a time slice. In topic evolution visualization, the solid lines indicate linked themes that share the same name, and the dotted lines indicate linked themes that do not share the same name. The thickness of the edge is proportional to the inclusion index, and the volume of the spheres is proportional to the number of published documents. The different color shadows group the topics belonging to the same area. There are topics with more than one shadow, which implies that the topic belongs to more than one topic area.

Alencar et al. (2012b) introduced a technique to create a sequence of maps that indicate content-based similarity over time. This map represents the evolution of scientific collection across successive time stamps. In the maps, each point represents an article, the point color denotes the publication year, the point size denotes the value of the global citation count, and the edges represent the bibliographic coupling strength between articles. It is only capable of analyzing small scientific collections.

Hascot and Dragicevic (2011) introduced a tool, Donatien, which supports the comparison of document collections. The set of documents are represented by a graph in which the nodes denote the documents,and the edges denote the relationship among the documents. The users can quickly look through the evolution of topics, author, and paper through crossing-based interaction.

Chuang et al. (2012b) researched the interdisciplinary and scientific innovation by measuring the impact of interdisciplinary collaboration, quantifying the emergence and converge of subfields, and identifying the patterns of cross-discipline collaboration. The final conclusion is that successful model-driven visualizations depend on appropriate alignment of the model, visualization, analysis task, and users' expertise. He used three methods to extract topics from Ph.D. dissertations published at Stanford University. TopicFlow visualization uses node-link graph in a chronological sequence to display computational linguistics data of 45 years, by combining network analysis with topic model. They used circle view to show the topical overlap between research areas.

To improve the performance of unsupervised clustering, Janssens and Moor (2007) merged textual contents with the structure of the citation graph, and proceeded with a hybrid clustering method based on Fisher's inverse Chi-square. The same clustering methodology is applied to each time span, and in a subsequent phase chains are formed by matching and tracking the clusters through time (see Fig. 4e). Finally, the author provides an evolution view of the different subfields in the bioinformatics community over time, which is shown in Fig. 4e. The HITS and PageRank algorithms are used to confirm the representative publications in each cluster. Term networks of each cluster chain presents the cognitive structure of the field.

According to the above, combining topic with the time line can intuitively show the topic evolution in Morris et al. (2003), Cobo et al. (2011), Chuang et al. (2012b) and Janssens and Moor (2007). The agglomerative hierarchal clustering lays out the similarity of topics together in the study by Morris et al. (2003). Color shadows groups and different line types directly reflect the relationship between topics and topic evolution in Cobo et al. (2011). In Mane and Brner (2004), the color coding displays the time change; however, it cannot intuitively discover topic evolution. In Alencar et al. (2012b), a sequence node-link diagram shows the topic evolution; however, it is only used in small document collections.

3.3.3 Animation

Animation rapidly displays the temporal evolution of topics and monitors the emerging topics. As it is based on the promotion of dimension, the relationship between topics is expressed better. However, due to the effect of short-term memory, it is hard to remember the temporal evolution of each topic and show the detailed information of a topic.

Alsakran et al. (2012) designed an interactive visual system, STREAMIT, which displays the topic evolution of a text stream. It is based on a dynamic force-directed simulation into which documents are continuously inserted. First, the topics are extracted by the LDA model. Then the document location is computed using the similarity of documents. In the visualization system (see Fig. 4c), a particle represents a document that is displayed as a pie chart, each cluster is encoded using a different color, and the documents of the cluster are mapped to the spiral according to their time stamps. The users can learn the temporal trends of the cluster by observing the distribution of the pie charts on the spiral. When there is a new document to be added, the new particles will automatically move to the cluster with the highest similarity. When the distance between the two clusters is less than the threshold, they are merged into a new cluster.

For the study of topic evolution, people well prefer to compare static maps for analyzing the topic evolution. At present, researches on methods based on animation are relatively scarce.

3.3.4 Other methods

PaperLens (Lee et al. 2005) visualized InfoVis conference proceedings and CHI conference proceedings, and revealed topic trends, author connections, and activities throughout the community. The topics are depicted in the form of a bar chart per year, and the top ten cited papers/authors per year are listed. The users can find connections between authors in the author degrees of separation link view.

4 Visualization tools

There are some softwares that use citation analysis and co-word analysis for scientific literature. We cannot simply classify these tools into one single category for the reason that they perform multiple functions and become too complex and comprehensive. Moreover, there is great distinction in terms of the final visualization result, and compared to the aforementioned methods the results obtained using these softwares need to be explained by domain experts. Therefore, we introduce these tools as a separate section. We only list the free visualization tools that use bibliographic records, such as NWB, SCI2, BibExcel, and CiteSpace. The data process, analysis method, and applications of these tools are listed in Table 1.

The Network Workbench (NWB) Tool[2] is a free software developed by the University of Indiana. It is a network analysis, modeling, and visualization toolkit for scientific collections research. It provides special algorithms to process document data and can read almost all common data formats to build and analyze knowledge networks and maps. It can perform data preprocessing, different types of network construction, knowledge network analysis, and knowledge visualization of the entire process.

Science of Science (SCI2) Tool[3] is also a free software developed by the University of Indiana. It supports the temporal, geospatial, topical, and network analysis and visualization of datasets at the micro (individual), meso (local), and macro (global) levels. It accepts data in common formats as the input, provides a variety of methods to deal with the data, and is capable of building a common knowledge unit network. The users can utilize different visualizations to interactively explore and understand specific datasets.

BibExcel[4] is a free software developed by the Swedish scientific economist Persson. It is designed to assist a user in analysing bibliographic data, or any data of a textual nature formatted in a similar manner. The idea is to generate data files that can be imported to Excel, or any other program that accepts tabbed data records, for further processing. The processed data can then be used for further visualization using Pajek, NetDraw, and SPSS.

[2] http://nwb.cns.iu.edu/.

[3] http://sci2.wiki.cns.iu.edu/.

[4] http://homepage.univie.ac.at/juan.gorraiz/bibexcel/.

Table 1 Visual software of scientific collection

Software	Data processing	Networking building	Application
NWB	Dereplication, time span, data and network reduction	DBCA, ACAA, DCA, CWA,DL	Burst detection, social network, temporal analyses
SCI2	Dereplication, time span, data and network reduction	DBCA, ACAA, CCAA, ICAA, ACA, DCA, JCA, CWA, DL	Burst detection, geospatial patterns, social network, temporal analyses
BibExcel	Data and network reduction	DBCA, ACAA, CCAA, ICAA, ACA, DCA, JCA, CWA	Social network
CiteSpace	Time span, data and network reduction	DBCA, ACAA, CCAA, ICAA, ACA, DCA, JCA, CWA	Burst detection, geospatial patterns, social network, temporal analyses

ABCA author bibliographic coupling analysis, *DBCA* document bibliographic coupling analysis, *JBCA* journal bibliographic coupling analysis, *ACAA* author collaboration analysis, *CCAA* country collaboration analysis, *ICAA* institution collaboration analysis, *ACA* author cocitation analysis, *DCA* document cocitation analysis, *JCA* journal cocitation analysis, *CWA* co-word analysis, *DL* direct citation network

CiteSpace[5] has been developed by Dr. Chen Chaomei, professor of information science and technology at Drexel University in the USA. It is a freely available Java application for visualizing and analyzing the trends and patterns in scientific literature. There are three kinds of visualization patterns: clustering view, time line, and time region model, which can be clustered to show the evolution of knowledge in different time periods. It uses mutation detection to show the trend of knowledge. It focuses on finding critical points in the development of a field, especially intellectual turning points and pivotal points (Chen 2006).

5 Challenge and opportunity

In this section, we illustrate the existing challenges to inspire future research work.

Scalability With the rapid growth of scientific literature, the topics of research also change rapidly. It is important for discovering topic trends, identifying the emerging topics and extracting valuable insights from a large text flow. The effective representation and mining of information from larger input data are a great challenge. This involves efficient and reliable text mining techniques, scalable visual expression, and real-time interactive technology.

Data fusion In scientific collections, the citation data and text data have different characteristics, and both of them can be used for creating links between documents. The combination of multiple aspects of data in a reasonable and effective manner is a challenge. In Janssens and Moor (2007), a method combining the content similarity matrix with the bibliography similarity matrix is proposed. They obtain more accurate and reasonable topics compared to that obtained when using a single matrix. This is a method for integrating two types of data. However, the automatic and efficient determination of the mixture ratio of two matrices is a difficult problem. How to effectively integrate multiple attributes of data using a visual design is still an open question.

Representation To understand the topic content and the topic relationship, we use word list, word clouds, or labels representing the topic content simultaneously. Various visualization patterns represent the relationship between topics. With respect to representing a large number of topics, the method of parallel coordinate is chaotic, tree can well represent the hierarchical structure of topics; however, it cannot uncover adjacent relations, map and node-link diagram can well represent the topic relation but have a difficulty in the layout of keywords. How to represent the topic group is a challenge for interacting with topics on an understanding basis.

Evolution Topic evolution reveals the research trends and hot topics. However, owing to the increasing number of topics, we need to rapidly discover the burst topics, turning point, and front. It is a great challenge to discover the reasons for the occurrence of burst topics and important turning points. When the number of topics is more than ten, the evolution graph is chaotic. The representation of topic evaluation for a large number of topics is a challenge.

Interdisciplinary With the rapid development of various disciplines, the growth of interdisciplinary topics is becoming more and more prominent. New research directions and research hot spots are most likely to appear in interdisciplinary topics. The accurate and clear expression of the relationship of interdisciplinary topics is still a difficult problem. Due to the complexity of the overlap between disciplines, it is

[5] http://cluster.cis.drexel.edu/~cchen/citespace/.

a great challenge to make users understand the cross relationship and discover the interactions among them. When the number of domains is more than three, we cannot effectively show each and every overlap between domains of topics in a 2D plane. The representation of topics and their relation is a big challenge when there are many domains.

Combination The wise combination of data model, analysis task, visualization, and expertise is a difficult problem. In Chuang et al. (2012b), the author demonstrates that successful model-driven visualization depends on appropriate data model, visualization, analysis task, and users' expertise. Visual design is a great challenge for assembling the analysis task, the characteristics of the data, the data model, the domain of knowledge, and proper visualization.

6 Conclusion

In this survey, we presented a comprehensive overview of the visual approaches of scientific literature topics. Depending on the topic analysis task, we classify the existing visual approaches into three categories: topic content, topic relation, and topic evolution. Each category is further classified according to visualization pattern. In the light of topic analysis of scientific literature, some challenges and opportunities are illustrated.

Acknowledgements This work was supported by Qinghai Science and technology Projects (No. 2016-ZJ-Y04).

References

Alencar AB, Oliveira MCFD, Paulovich FV (2012a) Seeing beyond reading: a survey on visual text analytics. Wiley Interdiscip Rev Data Min Knowl Discov 2(6):476–492
Alencar AB, Paulovich FV, Oliveira MCFD (2012b) Time-aware visualization of document collections. In: ACM symposium on applied computing, pp 997–1004
Alexander E, Gleicher M (2016) Task-driven comparison of topic models. IEEE Trans Vis Comput Graph 22(1):320–329
Alsakran J, Chen Y, Luo D, Zhao Y, Yang J, Dou W, Liu S (2012) Real-time visualization of streaming text with a force-based dynamic system. IEEE Comput Graph Appl 32(1):34
Blei DM (2012) Probabilistic topic models. Commun ACM 55(4):77–84
Blundell C, Teh YW, Heller KA (2010) Bayesian rose trees. In: UAI 2010, Proceedings of the twenty-sixth conference on uncertainty in artificial intelligence, Catalina Island, CA, USA, July, pp 65–72
Brner K (2010) Atlas of science: visualizing what we know. The MIT Press, Cambridge
Cao N, Sun J, Lin YR, Gotz D, Liu S, Qu H (2010) Facetatlas: multifaceted visualization for rich text corpora. IEEE Trans Vis Comput Graph 16(6):1172–1181
Chaney AJB, Blei DM (2012) Visualizing topic models. ICWSM 2012
Chen C (2006) CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J Am Soc Inf Sci Technol 57(3):359–377
Chen C (2013) Mapping scientific frontiers: the quest for knowledge visualization, 2nd edn
Chen C, Paul RJ (2001) Visualizing a knowledge domain's intellectual structure. Computer 34(3):65–71
Choo J, Lee C, Reddy CK, Park H (2013) UTOPIAN: user-driven topic modeling based on interactive nonnegative matrix factorization. IEEE Trans Vis Comput Graph 19(12):1992
Chuang J, Manning CD, Heer J (2012a) Termite: visualization techniques for assessing textual topic models. In: International working conference on advanced visual interfaces, pp 74–77
Chuang J, Ramage D, Mcfarl DA, Manning CD, Heer J (2012b) Large-scale examination of academic publications using statistical models
Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F (2011) An approach for detecting, quantifying, and visualizing the evolution of a research field: a practical application to the fuzzy sets theory field. J Informetr 5(1):146–166
Collins C, Viegas FB, Wattenberg M (2009) Parallel tag clouds to explore and analyze faceted text corpora. In: IEEE symposium on visual analytics science and technology, 2009. VAST 2009, pp 91–98
Davidson GS, Hendrickson B, Johnson DK, Meyers CE, Wylie BN (1998) Knowledge mining with vxinsight: discovery through interaction. J Intell Inf Syst 11(3):259–285
Ding W, Chen C (2014) Dynamic topic detection and tracking: a comparison of HDP, C-word, and cocitation methods. J Assoc Inf Sci Technol 65(10):20842097
Dou W, Wang X, Chang R, Ribarsky W (2011) Paralleltopics: a probabilistic approach to exploring document collections. In: Visual analytics science and technology, pp 231–240
Dou W, Yu L, Wang X, Ma Z, Ribarsky W (2013) Hierarchicaltopics: visually exploring large text collections using topic hierarchies. IEEE Trans Vis Comput Graph 19(12):2002–2011
Federico P, Heimerl F, Koch S, Miksch S (2017) A survey on visual approaches for analyzing scientific literature and patents. IEEE Trans Vis Comput Graph 23(9):2179–2198
Fried D, Kobourov SG (2013) Maps of computer science. In: Visualization symposium, pp 113–120

Gad S, Javed W, Ghani S, Elmqvist N, Ewing T, Hampton KN, Ramakrishnan N (2015) Themedelta: dynamic segmentations over temporal topic models. IEEE Trans Vis Comput Graph 21(5):672–85
Gretarsson B, Bostandjiev S, Asuncion A, Newman D, Smyth P (2012) TopicNets: visual analysis of large text corpora with topic modeling. ACM Trans Intell Syst Technol 3(2):23
Hascot M, Dragicevic P (2011) Visual comparison of document collections using multi-layered graphs. RR-11020, 2011, pp 1–10
Havre S, Hetzler B, Nowell L (2000) Themeriver: visualizing theme changes over time. In: Proceedings of the IEEE symposium on information visualization InfoVis, pp 115–115
Heimerl F, Han Q, Koch S, Ertl T (2016) CiteRivers: visual analytics of citation patterns. IEEE Trans Vis Comput Graph 22(1):190
Janssens FAL, Glänzel W, Moor BD (2007) Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. In: ACM SIGKDD international conference on knowledge discovery and data mining, pp 360–369
Jiang X, Zhang J (2016) A text visualization method for cross-domain research topic mining. J Vis 19(3):561–576
Keena N, Etman MA, Draper J, Pinheiro P, Dyson A (2016) Interactive visualization for interdisciplinary research. Vis Data Anal 2016:1–7
Kohonen T, Kaski S, Lagus K, Salojärvi J, Honkela J, Paatero V, Saarela A (2000) Self organization of a massive document collection. IEEE Trans Neural Netw Learning Syst 11(3):574–585
Kucher K, Kerren A (2015) Text visualization techniques: taxonomy, visual survey, and community insights, In: PacificVis, pp 117–121
Lee B, Czerwinski M, Robertson G, Bederson BB (2005) Understanding research trends in conferences using PaperLens. In: Extended abstracts proceedings of the 2005 conference on human factors in computing systems, CHI 2005, Portland, Oregon, USA, April, pp 1969–1972
Lee H, Kihm J, Choo J, Stasko J, Park H (2012) iVisClustering: an interactive visual document clustering via topic modeling. Comput Graph Forum 31(3pt3):1155–1164
Liu S, Wang X, Song Y, Guo B (2015) Evolutionary Bayesian rose trees. IEEE Trans Knowl Data Eng 27(6):1533–1546
Maiya AS, Rolfe RM (2014) Topic similarity networks: visual analytics for large document sets. In: IEEE international conference on big data, pp 364–372
Mane KK, Brner K (2004) Mapping topics and topic bursts in PNAS. Proc Natl Acad Sci USA 101(Suppl 1):5287
Morris SA, Yen G, Wu Z, Asnake B (2003) Time line visualization of research fronts. J Am Soc Inf Sci Technol 54(5):413–422
Murdock J, Allen C (2015) Visualization techniques for topic model checking. In: AAAI conference on artificial intelligence, pp 4284–4285
Oelke D, Strobelt H, Rohrdantz C, Gurevych I, Deussen O (2014) Comparative exploration of document collections: a visual analytics approach. Comput Graph Forum 33(3):201–210
Oesterling P, Scheuermann G, Teresniak S, Heyer G, Koch S, Ertl T, Weber GH (2010) Two-stage framework for a topology-based projection and visualization of classified document collections. In: Visual analytics science and technology, pp 91–98
Skupin A (2002) A cartographic approach to visualizing conference abstracts. IEEE Comput Graph Appl 22(1):50–58
Skupin A (2004) The world of geography: visualizing a knowledge domain with cartographic means. Proc Natl Acad Sci USA 101(Supplement 1):5274
Small H (1973) Co citation in the scientific literature: a new measure of the relationship between two documents. J Am Soc Inf Sci 24(4):265–269
Wang X, Cheng Q, Lu W (2014) Analyzing evolution of research topics with NEViewer: a new method based on dynamic co-word networks. Scientometrics 101(2):1253–1271
Wang X, Liu S, Liu J, Chen J, Zhu J, Guo B (2016) TopicPanorama: a full picture of relevant topics. IEEE Trans Vis Comput Graph 22(12):2508
Wei F, Liu S, Song Y, Pan S, Zhou MX, Qian W, Shi L, Tan L, Zhang Q (2010) TIARA: a visual exploratory text analytic system. In: ACM SIGKDD international conference on knowledge discovery and data mining, Washington, DC, USA, July, pp 153–162
Wise JA (1999) The ecological approach to text visualization. Wiley, New York
Wu Y, Thomas P, Wei F, Liu S, Ma K (2011) Semantic-preserving word clouds by seam carving. Comput Graph Forum 30(3):741–750
Yan E, Ding Y (2012) Scholarly network similarities: how bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. Wiley, New York