This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2834341, IEEE Transactions on Visualization and Computer Graphics

Bridging Text Visualization and Mining: A Task-Driven Survey

Shixia Liu, Xiting Wang, Christopher Collins, Wenwen Dou,
Fangxin Ouyang, Mennatallah El-Assady, Liu Jiang, and Daniel A. Keim

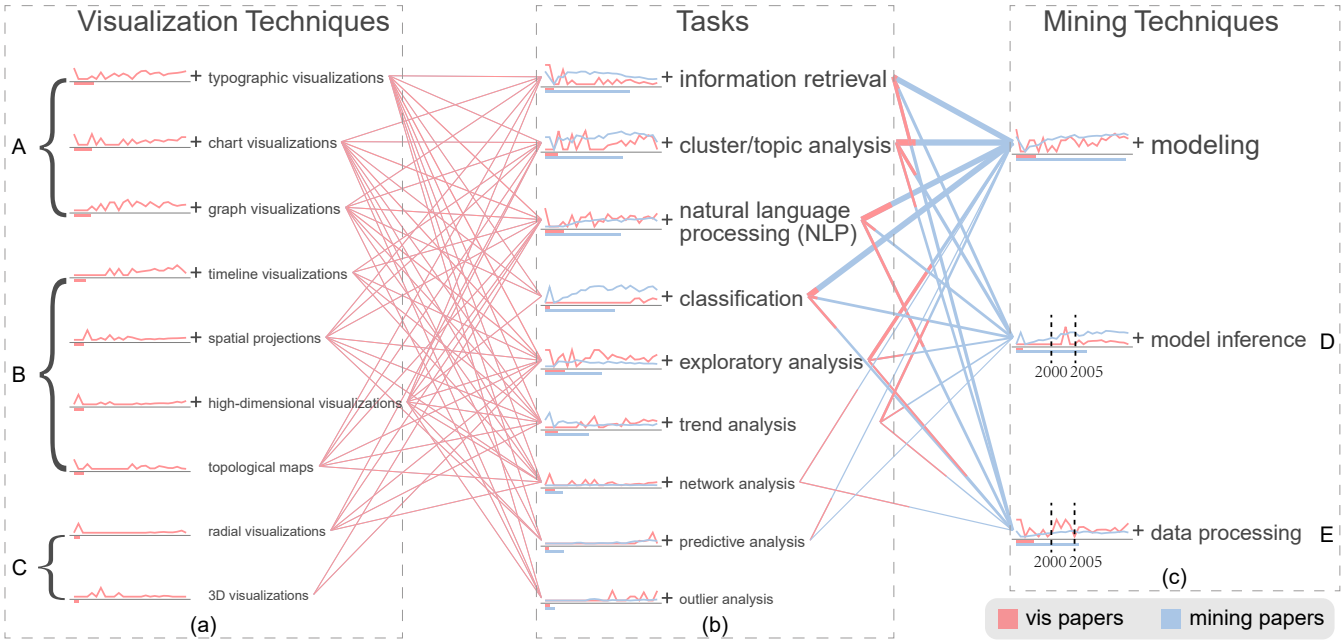


Fig. 1: Task-oriented analysis: visualization and mining techniques are connected through their shared analysis tasks (here, only 80% of the edges are shown), to aid researchers and practitioners in understanding current practices in visual text analytics, identifying research gaps, and seeking potential research opportunities. Concepts are organized into three taxonomies which can be navigated interactively.

Abstract—Visual text analytics has recently emerged as one of the most prominent topics in both academic research and the commercial world. To provide an overview of the relevant techniques and analysis tasks, as well as the relationships between them, we comprehensively analyzed 263 visualization papers and 4,346 mining papers published between 1992-2017 in two fields: visualization and text mining. From the analysis, we derived around 300 concepts (visualization techniques, mining techniques, and analysis tasks) and built a taxonomy for each type of concept. The co-occurrence relationships between the concepts were also extracted. Our research can be used as a stepping-stone for other researchers to 1) understand a common set of concepts used in this research topic; 2) facilitate the exploration of the relationships between visualization techniques, mining techniques, and analysis tasks; 3) understand the current practice in developing visual text analytics tools; 4) seek potential research opportunities by narrowing the gulf between visualization and mining techniques based on the analysis tasks; and 5) analyze other interdisciplinary research areas in a similar way. We have also contributed a web-based visualization tool for analyzing and understanding research trends and opportunities in visual text analytics.

Index Terms—Visualization, Visual Text Analytics, Text Mining.

1 Introduction

The significant growth of textual data and the rapid advancement of text mining have led to the emergence and prevalence of visual text analytics. This research combines the advantages of interactive visualization and text mining techniques to facilitate the exploration and analysis of large-scale textual data from both a structured and unstructured perspective. Visual text analytics has recently emerged as one of the most prominent topics in both academic research and the commercial world. For example, a leading business intelligence system, *Power BI*, announced features that enable the exploration and analysis of textual collections in 2016 [366]. The ultimate goal of visual text analytics is to enable human understanding and reasoning about large amounts of textual information in order to derive insights and knowledge [95].

- *S. Liu, F. Ouyang and L. Jiang are with Tsinghua University.*
- *X. Wang is with Microsoft Research.*
- *C. Collins is with University of Ontario Institute of Technology (UOIT).*
- *W. Dou is with University of North Carolina at Charlotte.*
- *M. El-Assady is with UOIT and University of Konstanz.*
- *D.A. Keim is with University of Konstanz.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2834341, IEEE Transactions on Visualization and Computer Graphics

Due to the rapid expansion of research in the area of visual text analytics [10], [95], there is a growing need for a meta-analysis of this area to support understanding how approaches have been developed and evolved over time, and their potential to be integrated into real-world applications. There are several initial efforts to summarize the existing text visualization techniques by different aspects, such as data sources, tasks, and visual representations [10], [120], [168]. For example, Alencar and Oliveira summarized around 30 text visualization techniques published before 2013 [10]. Later, Kucher and Kerren extended this survey work by creating a comprehensive taxonomy with multiple categories and items in order to classify the techniques with fine granularity [189]. They also developed a useful Web-based survey browser to facilitate the exploration of the created taxonomy. The most recent effort focuses on analyzing scientific literature [110]. While these efforts provide an overview of text visualization techniques, they do not investigate the underlying text mining techniques. On the other hand, several comprehensive surveys of the work on text mining have been published summarizing the relevant research progress [6], [7], [44], [130], [146], [225]. These surveys provide much valuable complementary information to existing literature reviews on text visualization. However, they do not establish a link between text mining and interactive visualization. The most powerful visual text analytics systems make use of advanced data mining algorithms and techniques, and we are still in need of an overview, accounting for both the user-facing visualization and the back-end data mining approaches. Our survey will provide practical knowledge that relates to building visual text analytics tools, and it will support researchers in discovering opportunities to narrow the gulf between the visualization and text mining fields. In particular, the gaps between these two fields based on analysis tasks have not been explored yet. Many available text mining techniques that may be useful for visual text analytics have not been connected to visualizations. This may hinder the further development of this research area.

Hence, our approach is to highlight associations between analysis tasks, visualization techniques, and text mining techniques through a set of taxonomies. As shown in Fig. 1, these taxonomies contribute to 1) **understanding** current practices in developing visual text analytics tools, the on-going research, and the principal research trends; and 2) **seeking** potential research opportunities by relating *text visualization research* with *text mining research*. Ultimately, these taxonomies will pave the way for identifying the relationships and gaps between analysis tasks and techniques (visualization and text mining) to explore future research directions and develop novel applications.

To this end, we analyzed over 4,600 research papers published between 1992–2017 in 4 journals (IEEE TVCG, ACM TOCHI, IEEE TKDE, and JMLR) and 16 conference proceedings (InfoVis, VAST, SciVis, EuroVis, PacificVis, AVI, CHI, IUI, KDD, WWW, AAAI, IJCAI, ICML, NIPS, ACL and SIGIR) in the fields of text mining and visualization. As shown in Fig. 2, a semi-automatic analysis process was designed to analyze the text visualization and mining literature. We first extracted and summarized the concepts (described in phrases) that capture visualization techniques, mining techniques, and analysis tasks related to visual text analytics. Our extraction method is based on a pattern-based concept extraction algorithm [129]. Accordingly, three concept taxonomies—visualization techniques, analysis tasks, and mining techniques—were derived. The relationships between the concepts in the taxonomies were then extracted by using the co-occurrence statistics between different types of concepts in papers (e.g., the co-occurrence of the visualization techniques and analysis tasks). Finally, a graph-based interactive visualization was developed to help understand and analyze the three taxonomies and the relationships between them.

By semi-automatically refining and analyzing these concepts in an interactive and progressive process, we identified the following key features of visual text analytics. First, the most used visualization techniques are traditional ones such as chart visualizations, typographic visualizations, and graph visualizations. The popularity of these techniques is probably due to their simplicity and intuitiveness, as well as the employment of more advanced mining techniques in recent years. Second, a set of less frequently studied tasks in visual text analytics were identified, which include "information retrieval," "network analysis," "classification," "outlier analysis," and "predictive analysis." Third, different analysis tasks are supported by different techniques, including text mining and visualization techniques, as well as their combination.

Through a task-oriented and data-driven analysis, we thoroughly investigated each of the three concept taxonomies and their connections to identify overarching research trends and under-investigated research topics within visual text analytics. Consequently, we map out future directions and research challenges to enhance the development of new visual text analytics techniques and systems. The major contributions of this work are:

- **A semi-automatic analysis approach** that focuses on extracting, understanding, and analyzing the major concepts in the area of visual text analytics. This approach can be easily extended to analyze other research areas.
- **Three concept taxonomies and a data-driven method** to extract the relationships between them, better revealing overarching research trends and missing research topics within visual text analytics.
- **A web-based visualization tool** that enables the analysis of the major research trends and the potential research directions in visual text analytics. This visualization tool is published at: http://visgroup.thss.tsinghua.edu.cn/textvis.
- **A comprehensive survey** of the literature in visual text analytics, classifying thousands of papers along our technique- and task-taxonomies.

2 Survey Landscape

To obtain an overview of visual text analytics and the relationship between the relevant visualization and mining fields, we systematically reviewed research articles from both fields. The approach we took for each type of article was different. For visualization papers, we followed an exhaustive manual review of relevant venues. For text mining papers, we followed the semi-automated approach of Sacha et al. [276], which is a combination of a manual selection and an automatic keyword-based extraction method.

2.1 Paper Selection: Visualization

There are fewer visualization papers than text mining papers. Therefore, it was possible to do an in-depth manual selection. For this group of papers we preferred *precision* in that we only wanted papers that deal with visualizing text data. Depending on how closely related the venue (e.g., a conference) was to visualization research, we followed two main approaches: full coverage or search-driven selection. For venues identified as being primarily about visualization, we reviewed every title from all the conference

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2834341, IEEE Transactions on Visualization and Computer Graphics

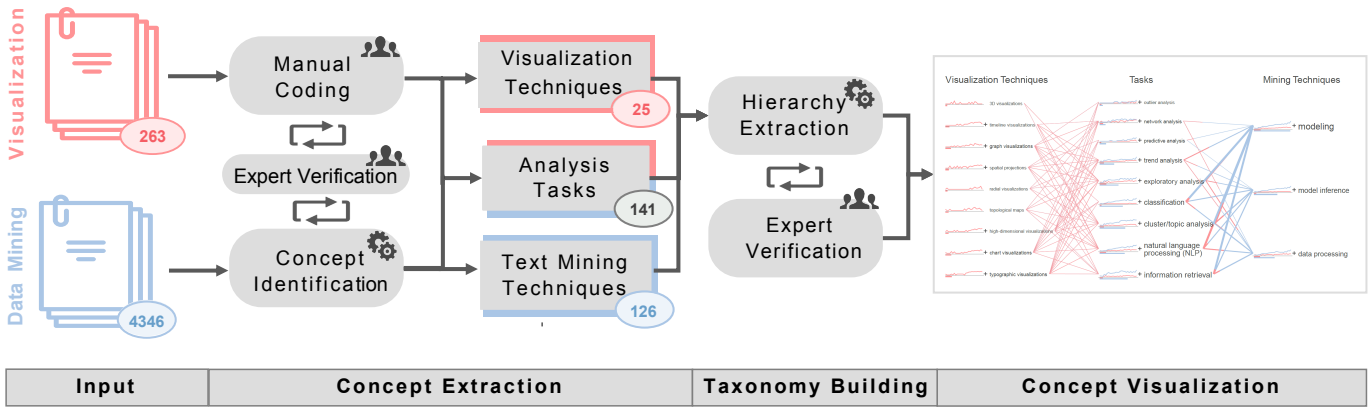


Fig. 2: The analysis pipeline aims at extracting, correlating, organizing, and presenting three types of concepts: mining techniques, analysis tasks, and visualization techniques. Three concepts comprise our description of the space of visual text analytics.

proceedings to identify candidates. We then reviewed the abstracts of candidates, and finally the full text of any papers when it was not clear from the title and abstract whether the paper contained any text visualizations. The venues for this approach were: InfoVis, VAST, Vis (later SciVis), EuroVis, AVI, PacificVis, and IUI. For higher volume venues with a larger proportion of irrelevant papers, we used two search queries ("text" `AND` "visualization"; "text" `AND` "analytics"), then reviewed titles, abstracts, and full texts to finalize the selections. This approach was applied to all available years of: CHI, KDD, and WWW. Finally, we used this same search-driven selection approach on all available years of two relevant journals: IEEE TKDE and IEEE TVCG. This resulted in a total of 263 included papers that deal with text visualization and visual text analytics. Our method is extensible, so it would be possible to add further venues to the analysis and accompanying website in the future.

2.2 Paper Selection: Text Mining

For the text mining papers, we optimize for *recall* as we are open to including text mining techniques that may be unknown or underutilized in the visualization field. To provide good coverage of visual text analytics and related text mining methods, we followed the approach of Sacha et al. [276] to extract relevant research papers in a semi-automated fashion. The paper collection was done in three steps.

In the first step, we performed seed paper selection manually by reviewing titles and abstracts from the most recent year of papers from leading data mining conferences and journals (AAAI, KDD, WWW, SIGIR, ICML, NIPS, IJCAI, ACL, and JMLR). 607 papers were thus identified. These were combined with the papers from our visualization paper collection to create the seed collection of papers used in the second step. A detailed description of the statistics is shown in Fig. 3.

In the second step, we performed keyword extraction based on the seed paper collection. Specifically, we manually checked the top keywords of the seed paper collection and selected 10 keywords that denote the data source. The keywords then serve as query terms to retrieve more mining papers. The 10 extracted keywords were: "*text, document, blog, news, tweet, twitter, wikipedia, book, microblog, textual*."

In the third step, we retrieved the full text of all papers from the aforementioned top data mining venues. The retrieved papers were indexed by using Lucene [1]. We then searched the Lucene index with the 10 keywords extracted from the second step, and ranked the papers based on the relevance score provided by Lucene. Papers with relevance scores larger than 0.1 were selected as the text mining papers. This cut-off threshold of 0.1 was used because we found that it can balance precision and recall. After this step, we obtained 4,346 mining papers.

Since the number of text mining papers are an order of magnitude greater than the number of visual text analytics related articles, analyzing the two corpora jointly may lead to the results being dominated by patterns from the text mining articles. Therefore, we developed an approach for analyzing each corpus separately and connecting the two corpora by tasks and techniques.

2.3 Analysis Process

Recently, Isenberg et al. [163] developed a data-driven approach to determine major visualization topics and point out missing

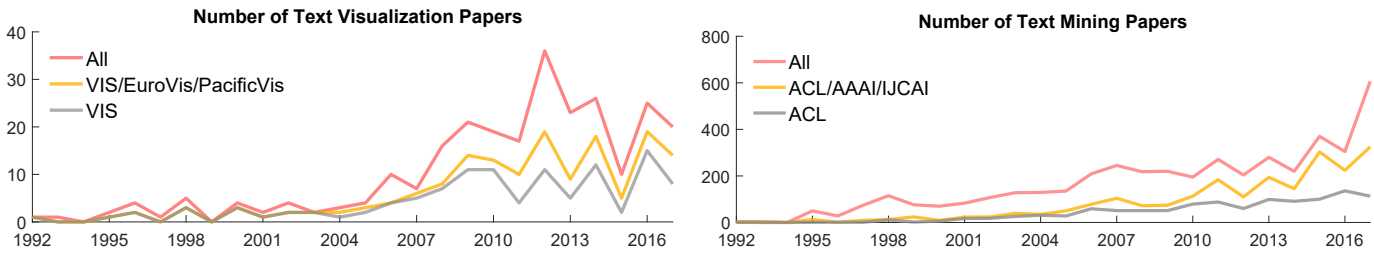


Fig. 3: Corpus size by data source over time. The volume of mining papers is much larger than that of visualization papers.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2834341, IEEE Transactions on Visualization and Computer Graphics

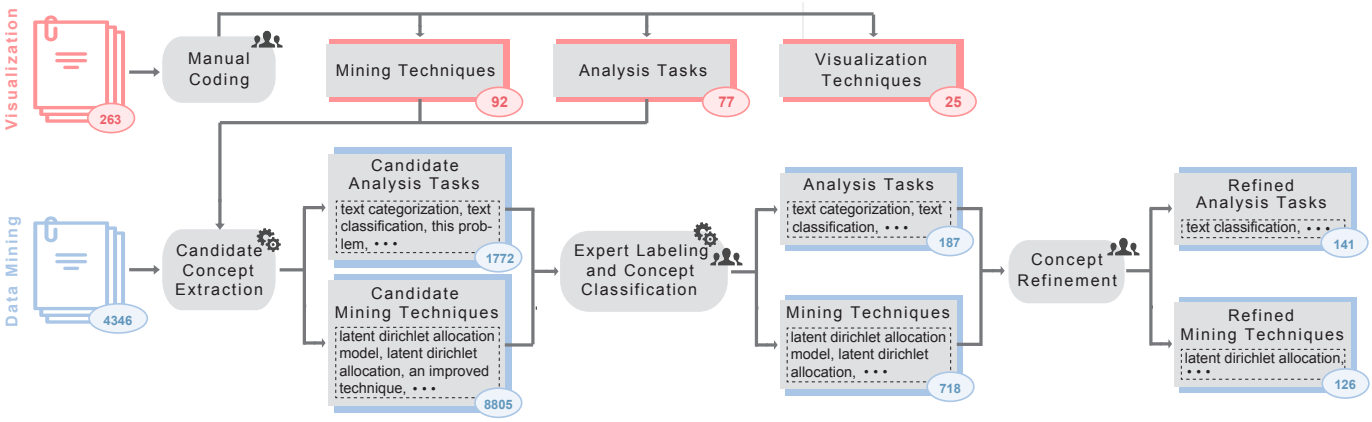


Fig. 4: The pipeline of concept extraction.

and overemphasized topics. To this end, they mainly focused on examining and analyzing two sets of keywords: author-assigned keywords and PCS taxonomy keywords. Inspired by their method, we first examined the 263 collected text visualization papers and performed a bottom-up analysis, mainly focusing on checking the employed techniques and the tasks supported by the techniques.

Our preliminary analysis revealed that *visualization techniques*, *analysis tasks*, and *mining techniques* are the key types of concepts in visual text analytics. We distilled the three types of concepts by 1) studying the definition, scope, and pipelines of visual (text) analytics; and 2) learning the key aspects of academic papers. Previous study in the field of natural language processing has shown that the application domain and technique are two key aspects for a scientific paper [129]. In the field of visual text analytics, information regarding application domains is usually captured by analysis tasks [178]. According to the pipelines of visual (text) analytics, mining techniques and visualization techniques are two essential types of techniques [95], [178]. Hence, our analysis is based on analysis tasks, mining techniques, and visualization techniques. In particular, we focus on identifying the key concepts in each type, as well as, the co-occurrence relationships between different types of concepts (Fig. 2).

To identify tasks and techniques from the surveyed research articles, we performed both manual coding and automated analysis that leverages the manual coding results to bootstrap concept identification on a much larger scale. In particular, visualization papers were gathered first; the manually labeled tasks and mining techniques from these papers were later used as seeds for extracting further concepts from the mining papers.

We developed a semi-automatic method to extract the three taxonomies from our source papers and identified the relationships between them. Fig. 2 shows the corresponding workflow to generate and analyze the three types of concepts in visual text analytics. The workflow consists of three levels: **concept extraction** to extract the three types of concepts by using a computational linguistics method [129]; **taxonomy building** to create a concept taxonomy for each type of concept manually or by the K-means clustering algorithm; and **concept visualization** to facilitate the understanding and analysis of the concepts and the relationships among them. In the following sections, each of the three levels will be described in more detail.

3 Concept Extraction

Our approach extracts concepts from both text visualization papers (total number: 263) and text mining papers (total number: 4,346). As shown in Fig. 4, concepts from the text visualization papers were annotated manually. In total, we got 92 mining techniques, 77 analysis tasks, and 25 visualization techniques. Since the number of text mining papers was large, we used a semi-automatic method to extract the corresponding analysis tasks and mining techniques. This method combines automatic concept extraction with two types of expert knowledge: 1) the mining techniques and analysis tasks extracted manually from the text visualization papers; and 2) expert labeling and refinement of the concepts. Specifically, our method consists of three steps: candidate concept extraction, expert labeling and classification, and concept refinement.

Candidate concept extraction. In the first step, we extract the candidate concepts using a computational linguistics method proposed by Gupta et al. [129]. Given a collection of research papers, this method extracts analysis tasks (e.g., speech recognition) and techniques (e.g., latent Dirichlet allocation) by matching dependency patterns between words. For example, given a sentence "we addressed this problem by using latent Dirichlet allocation," we can extract the technique "latent Dirichlet allocation" by using pattern "$using \rightarrow (direct - object)$." A key here is to define the patterns used to extract the techniques and tasks (application domains). In this paper, we combine the seed patterns provided by Gupta et al. [129] with the patterns we extracted by using the mining techniques and tasks extracted from the visualization papers. In total, we extracted 1,772 candidate tasks and 8,805 candidate mining techniques.

Expert labeling and concept classification. While the computational linguistics method is able to identify many candidate techniques and tasks, its ability to differentiate meaningless phrases such as "this problem" from meaningful phrases such as "text categorization" is limited. To reduce the noise, we sampled 2,000 candidate concepts, asked experts to label whether each concept was noise or not, and then used classification to identify noise in the rest of the candidate concepts. Specifically, five experts, all having more than five years of research experience in text mining and/or text visualization, were asked to label the candidate concepts. We assigned the concepts to the experts so that each concept was labeled by two experts. The labeling agreement was 81.4%. 1,628 concepts that were given the same label by two experts were used in the classification.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2834341, IEEE Transactions on Visualization and Computer Graphics

TABLE 1: Taxonomy of visualization techniques and the papers demonstrating each technique. The number in the bracket is the number of papers that use each visualization technique.

First-level	Second-level	Examples
typographic visualizations (125)	text highlighting (97)	[22], [25], [27], [35], [39], [42], [65], [73], [78], [84], [94], [114], [123], [125], [140], [142], [152], [161], [175], [181], [185], [205], [207], [214], [227], [236], [247], [252], [258], [354]
	word cloud (56)	[5], [22], [26], [37], [52], [53], [55], [62], [79], [85], [94], [126], [145], [155], [159], [194], [171], [184], [187], [190], [201], [211], [212], [219], [243], [253], [301], [334], [337], [324]
	hybrid (2)	[77], [194]
chart visualizations (102)	other charts (e.g. bar chart) (52)	[22], [42], [50], [53], [56], [58], [61], [65], [79], [91], [106], [109], [111], [121], [126], [140], [142], [161], [170], [184], [190], [195], [211], [212], [222], [248], [249], [257], [265], [320]
	scatterplot (32)	[25], [36], [39], [65], [69], [76], [96], [111], [122], [140], [142], [181], [193], [200], [208], [223], [233], [230], [235], [251], [252], [254], [265], [275], [326], [332], [338], [342], [358], [361]
	line chart (26)	[23], [34], [37], [113], [122], [150], [157], [162], [194], [211], [212], [223], [224], [230], [232], [250], [268], [279], [282], [313], [316], [320], [342], [344], [355], [362]
	table (15)	[12], [20], [59], [94], [122], [139], [151], [169], [177], [201], [233], [235], [282], [303], [312]
graph visualizations (95)	node-link (48)	[27], [28], [38], [47], [57], [58], [66], [71], [81], [101], [105], [107], [115], [152], [165], [184], [186], [195], [202], [216], [220], [232], [265], [285], [286], [288], [298], [310], [315], [333]
	tree (34)	[3], [19], [35], [39], [46], [76], [82], [86], [98], [122], [155], [179], [205], [208], [212], [220], [230], [254], [270], [272], [278], [293], [320], [325], [329], [338], [336], [346], [355], [359]
	matrix (26)	[16], [43], [49], [58], [68], [75], [81], [103], [115], [125], [143], [162], [180], [195], [235], [239], [240], [242], [244], [267], [268], [282], [299], [300], [360]
timeline visualizations (50)	stream graph (35)	[57], [63], [84], [86], [96], [97], [98], [115], [119], [134], [135], [141], [188], [204], [205], [207], [214], [216], [247], [286], [290], [302], [324], [323], [325], [330], [336], [344], [354], [361]
	flow (28)	[9], [20], [26], [54], [59], [70], [84], [86], [98], [115], [118], [119], [167], [169], [171], [197], [204], [212], [274], [285], [286], [288], [302], [324], [325], [333], [336], [344]
spatial projections (35)	galaxies (33)	[12], [15], [46], [47], [55], [59], [60], [94], [108], [114], [115], [117], [124], [125], [138], [143], [144], [151], [162], [190], [198], [202], [219], [232], [237], [267], [303], [311], [326], [336]
	voronoi (6)	[15], [59], [202], [238], [284], [325]
high-dimensional visualizations (34)	glyph (24)	[12], [22], [23], [27], [34], [68], [94], [101], [105], [106], [115], [142], [143], [159], [202], [239], [241], [269], [297], [342], [344], [346], [355], [358]
	PCP (11)	[11], [77], [96], [194], [195], [211], [269], [271], [288], [316], [333]
topological maps (33)		[35], [47], [46], [53], [54], [63], [82], [93], [97], [108], [114], [117], [121], [144], [152], [167], [179], [184], [200], [212], [219], [224], [226], [232], [307], [308], [309], [320], [323], [334]
radial visualizations (15)		[3], [46], [59], [66], [69], [75], [104], [149], [150], [208], [212], [272], [288], [320], [351]
3D visualizations (10)		[56], [88], [132], [143], [165], [230], [233], [245], [273], [308]

Next, we employed the support vector machine (SVM) model [80] to classify the remaining concepts. The SVM inputs were feature vectors and labels of the concepts. To calculate the feature vector of each concept, we used KNET [83], which is a deep learning framework for learning word embeddings. By using this framework, concepts with similar syntactic and semantic relationships were assigned to similar feature vectors. The SVM model was trained using a five-fold cross validation with average accuracy of 89.4% for analysis tasks and 91.4% for mining techniques. After applying the model to the remaining 8,949 candidate concepts, we found 187 analysis tasks and 718 mining techniques.

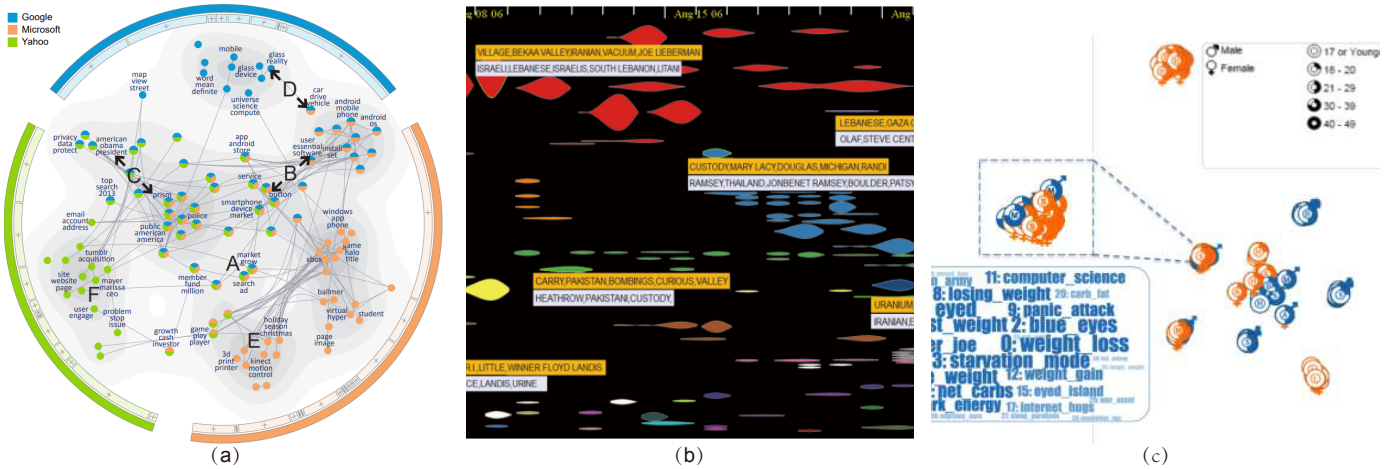
Concept refinement. When examining the concepts extracted, we observed that some of them were quite similar (e.g., "text categorization" and "text classification"). Also, the classification result of the second step was not perfect. To solve these problems, we asked two experts to manually check the results, merge similar concepts, and reduce noise. To facilitate the labeling process, we organized similar concepts into clusters by applying K-means on the word embedding feature vectors. By checking the clusters, the experts were able to detect concepts that needed to be merged or removed. Following this step, we obtained 141 refined analysis tasks and 126 refined mining techniques.

4 Taxonomy Building

Based on the three-level workflow (Fig. 2) and the analysis of a large number of text visualization papers and text mining papers, we have constructed three taxonomies: analysis tasks, visualization techniques, and mining techniques.

To build the visualization technique taxonomy, two co-authors, who are the experts in the visualization field, manually constructed the taxonomy based on the 25 visualization technique concepts that were manually extracted from text visualization papers. The other co-authors then examined and refined the visualization taxonomy. The refinement was done in an iterative fashion; all co-authors participated in multiple rounds of discussions to finalize

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2834341, IEEE Transactions on Visualization and Computer Graphics



(a) (b) (c)

Fig. 5: Example text visualizations: (a) TopicPanorama [326] leverages graph visualization to encode topic graphs from multiple sources; (b) EventRiver [214] helps users browse, search, track, associate, and investigate events by using a timeline-based visualization; (c) In DemographicVis [94], spatial projection is employed to show user groups based on topic interests.

the visualization taxonomy. With this method, we generated a two-level visualization technique hierarchy with 6 internal nodes.

As more concepts were extracted for analysis tasks and mining techniques, a semi-automatic method was employed to build the corresponding taxonomies. Previous research has indicated that good taxonomies should not be too deep or too wide [90]. As a result, we strived to establish a compromise between the tree depth and width while building these two taxonomies. In particular, we first employed the popular K-means clustering algorithm to create the task taxonomy and mining taxonomy. We iteratively divided the embedding feature vectors into K clusters by using K-means and generated a four-level task hierarchy with 24 internal nodes as well as a five-level mining hierarchy with 25 internal nodes. Then two authors worked iteratively to refine and improve these two taxonomies. Next, all other authors examined and refined these two taxonomies iteratively. Finally, we consulted with 4 text mining or machine learning experts to derive the final taxonomies.

The main objective of the resulting taxonomies is to provide a framework that is useful from the researcher's and practitioner's standpoint. They help to match techniques to real-world problems (represented by tasks), and more importantly serve as a foundation to develop new techniques and new applications.

4.1 Taxonomy of Visualization Techniques

We manually labeled all visualization papers with the concepts in the visualization technique taxonomy. In Table 1, we present the two-level taxonomy of visualization techniques, along with a **list of representative papers** that use each individual visualization technique in the hierarchy (at most 30 papers for each second level visualization technique). For each visualization technique, we reported the number of papers that use this technique.

The first level visualization techniques include 9 concepts, such as "graph visualizations" (Fig. 5(a)), "timeline visualizations" (Fig. 5(b)), and "spatial projections" (Fig. 5(c)). Some first-level concepts are further organized into 2 to 4 second-level concepts. Taking "timeline visualizations" as an example, it contains two second-level concepts: "stream graph" and "flow." As shown in Table 1, all papers that belong to a second-level visualization concept are included in the papers column.

The three first-level visualization concepts that have the largest number of related papers are "typographic visualizations," "chart visualizations," and "graph visualizations." The first-level concept that captures the largest number of publications is "typographic visualizations," which contains 125 papers. Since all of the visualization papers included in our survey are related to text data, it makes sense that many of these visualizations include a view that presents the texts in a typographic form.

The second-level techniques under "typographic visualizations" include "word cloud," "text highlighting," and "hybrid." For example, 56 papers incorporate a "word cloud" visualization; note that some papers designate word clouds as a primary view to summarize texts, while other papers use word clouds as facilitative visualization in addition to other visualization techniques. Another large first-level concept, "graph visualizations," is mentioned by 95 text visualization papers. This concept includes three second-level concepts, namely "tree," "matrix," and "node-link." The "node-link" technique captures the largest number of papers (48) under this concept. The "tree" and "matrix" techniques capture 34 and 26 papers, respectively. As many papers employ tree-based visualizations as the main view, we separate this concept to emphasize its role in visual text analytics.

Overall, the taxonomy provides a categorization of the visualization techniques presented in all the papers from the visualization publication venues. The taxonomy provides both an overview of the available visualization techniques and a way for researchers and practitioners to quickly identify papers related to a particular visualization technique.

4.2 Taxonomy of Analysis Tasks

When iteratively refining and improving the automatically generated task taxonomy, we first referred to the call for papers (CFPs) and section organization of several top-tier text-mining-related conferences and journals, including SIGIR, ACL, WWW, KDD, AAAI, ICML, NIPS, IJCAI and JMLR. In particular, we used the topics in the CFPs to organize and refine the concept hierarchy. We further used the session name of each paper to validate the concept(s) that are contained in the paper. For example, initially, "fragment detection" and "duplicate detection" were put at the first level by the K-means-based taxonomy building method. After checking the CFPs and section names of several SIGIR conferences, we found a topic and similar section names in SIGIR 2009 and

TABLE 2: Taxonomy of analysis tasks (top five largest clusters ranked by the number of papers in each cluster) and example papers. The number in the bracket represents the number of papers that aim at tackling each analysis task.

First-level	Second-level	Examples
information retrieval (1884)	entity ranking (6)	[8], [21], [166]
	XML retrieval (5)	[67], [72], [116]
	evaluation (35)	[51], [160], [284]
	user activity tracking (2)	[18], [27]
	search (76)	[17], [27], [124]
	recommendation (27)	[2], [41], [217]
	structure analysis (22)	[240], [260], [263]
	query analysis (302)	[19], [133], [322]
	filtering (15)	[35], [100], [357]
	interactive retrieval (12)	[136], [140], [278]
	unstructured information retrieval (107)	[88], [202], [348]
	efficiency and scalability (361)	[291], [343], [352]
cluster/topic analysis (1624)	community discovery (6)	[209], [305], [363]
	text segmentation (26)	[64], [172], [318]
	topic analysis (556)	[11], [26], [86]
	contextual text mining (2)	[228], [229]
	clustering (699)	[158], [192], [345]
natural language processing (NLP) (1623)	geotagging (8)	[13], [182], [199]
	data/information extraction (535)	[63], [182], [212]
	domain adaption (54)	[317], [335], [364]
	data enrichment (7)	[22], [121], [227]
	alignment (3)	[74], [9], [267]
	event analysis (113)	[53], [113], [214]
	discourse analysis (28)	[16], [190], [359]
	content analysis (78)	[155], [184], [314]
	sentiment analysis (467)	[244], [338], [325]
	lexical/syntactical analysis (73)	[244], [325], [338]
	question answering (156)	[151], [259], [289]
	text summarization (526)	[49], [62], [185]
classification (1299)	cross language text classification (2)	[128], [340]
	image classification (37)	[213], [221], [341]
	sub-document classification (4)	[109], [140], [170]
	binary classification (64)	[231], [264], [277]
	taxonomy integration (2)	[288], [353]
	hierarchical classification (23)	[102], [255], [266]
	query classification (23)	[92], [154], [176]
	web page classification (14)	[36], [287], [321]
	uncertainty tackling (4)	[42], [202], [288]
	tandem learning (1)	[261]
exploratory analysis (1028)	monitoring (113)	[12], [35], [296]
	comparison (541)	[77], [94] , [240]
	navigation/exploration (348)	[25], [27], [220]
	region of interest (6)	[125], [234], [360]

2010, for some of the papers related to the two concepts. The topic was "structure analysis," so we organized these two concepts as sub-concepts under the concept "structure analysis" in "information retrieval." Second, two senior PhD students, who majored in interactive machine learning and are not the co-authors of this paper, worked closely with two of the co-authors to iteratively refine the taxonomy through face-to-face discussions. Finally, we also worked with two researchers who majored in text mining or machine learning, a senior researcher from Microsoft Research, and a professor from the Hong Kong University of Science and Technology, to further verify and refine our taxonomy.

The final taxonomy of analysis tasks consists of three levels. The first level includes 9 concepts ranging from "information retrieval," "cluster/topic analysis," to several mining-related concepts such as "outlier analysis" and "network analysis." The selection and refinement of the first level concepts was inspired by previous studies on data mining tasks [7], [131]. In particular, we roughly divide the first-level concepts into three categories: **tasks on model building** (e.g., "classification" and "cluster/topic analysis"), **tasks on pattern detection** (e.g., "outlier analysis" and "trend analysis"), and **tasks on applications** (e.g., "natural language processing (NLP)" and "information retrieval").

The largest first-level cluster in terms of number of papers and child nodes is "information retrieval," which contains 1,884 papers (33 visualization papers and 1,851 mining papers) and 12 children, such as "entity ranking," "XML retrieval," and "efficiency and scalability" (Table 2). Another first-level cluster which also has the largest number of child nodes (12 children) is "natural language processing." The second largest cluster in terms of number of papers is "cluster/topic analysis," which contains 1,624 papers (54 visualization papers and 1,570 mining papers). "Exploratory analysis" and "natural language processing" are the largest and second largest clusters in terms of number of visualization papers, containing 132 and 120 papers, respectively. An interesting fact is that "exploratory analysis" only ranks 5th in terms of paper number. This indicates that the well-studied visual text analytics topic is not the most popular in the text mining field. The second level contains 62 concepts. For example, the cluster "topic analysis" contains sub-clusters "flat topic analysis", "hierarchical topic detection," and "topic evolution." The cluster "clustering" contains sub-clusters "flat clustering" and "hierarchical clustering." The third level consists of 62 fine-granularity concepts, such as "part-of-speech tagging," "co-reference analysis," "query processing," "search log analysis," and "indexing." The number of the second-level concepts is the same as that of the third-level concepts. After examining the taxonomy, we found that only 14 of the 62 second-level concepts had third-level child nodes.

The top five first-level clusters (including their children) ranked by the number of papers in each cluster are shown in Table 2. For each second-level node in the hierarchy, we select a few exemplar papers to show a range of analysis tasks within the corresponding visualization and mining papers. As there may be dozens of papers related to each second-level concept, we selected three visualization papers for those concepts whose relevant paper number is greater than three, if the number of visualization papers is less than three, mining papers are additionally selected.

4.3 Taxonomy of Mining Techniques

We categorized the mining techniques based on the three major stages of the machine learning life cycle: "data processing," "modeling," and "model inference" [29]. In the "data processing" stage, data is gathered and preprocessed for training and testing. In the "modeling" stage, we gather knowledge about the problem domain, make assumptions based on our knowledge, and express these assumptions in a precise mathematical form. In the "model inference" stage, the model variables are computed based on the data. Techniques related to other stages (e.g., evaluation and diagnosis) were merged with the three major stages to keep the taxonomy concise and clear. After building the taxonomy, we asked three researchers who majored in data mining to help refine the taxonomy. One researcher is a professor from the Hong Kong University of Science and Technology, the other two researchers are senior PhD students with more than four years of research experience in data mining.

The first two levels of the mining technique taxonomy are presented in Table 3. As shown in the table, we divided data processing techniques based on the data types. Accordingly, we got four second-level concepts: "document-level" data processing (e.g., document segmentation), "sentence/paragraph-level" data processing (e.g., sentence embedding), "word/phrase/entity-level" data processing (e.g., tokenization), and "hybrid" processing (e.g., relevance calculation). The modeling techniques were summarized based on textbooks on text mining and machine learning [7], [30]. Specifically, we have models for "classification" (e.g., Support Vector Machine), "clustering" (e.g., K-means), "dimension reduction" (e.g., multidimensional scaling), "topic" (e.g., latent Dirichlet allocation), and "regression" (e.g., convex regression). We also have "language model," "graphical models" (e.g., hidden Markov model), "neural networks" (e.g., convolutional neural networks), and "mixture models." The data inference techniques were categorized based on whether the method was probabilistic or not. For "probabilistic inference," the taxonomy contains parametric methods such as expectation maximization, as well as non-parametric methods such as Gibbs sampling. For "non-probabilistic inference," we have optimization techniques such as dynamic programming and convex optimization.

5 Visualization of Concept Relations

To understand the relationship between the visualization techniques, analysis tasks, and mining techniques, we developed an interactive visualization which connects the three concept hierarchies and provides access to the underlying research papers (see Fig. 1). Our visualization is task-oriented, placing analysis tasks at the center as the bridge between mining and visualization techniques. The goal of the visualization is to reveal common connections between visualization and text mining, as well as show research topics that are not well connected to identify potential opportunities for future work.

Our visualization is a tripartite graph, framed around the three concept hierarchies extracted from the research papers. Each column initially contains the first-level concepts, while connections are shown between the columns (see Fig. 6 left). Level-of-detail filters are provided to reduce clutter by decreasing the number of edges (co-occurrences) based on frequency thresholds. Connections at lower levels of the hierarchy are propagated to the parent concept, so that the initial overview shows a summary of all connections in the dataset, and drill-down can be used to see the details.

We separated source concepts into ones coming from visualization papers, shown in red, and others coming from data mining papers, shown in blue (bi-colored connections between (b) and (c) in Fig. 1). The total number of occurrences of a concept across all

TABLE 3: Taxonomy of mining techniques and example papers. The number in the bracket is the number of papers that leverage each mining technique.

First-level	Second-level	Examples
data processing (1175)	sentence-level (sentence embedding) (25)	[319], [327], [331]
	word/phrase/entity-level (679)	[4], [280], [283]
	document-level (288)	[14], [246], [365]
	hybrid (387)	[40], [153], [339]
model inference (1335)	non-probabilistic inference (267)	[89], [99], [328]
	probabilistic inference (1160)	[183], [196], [349]
modeling (3085)	models for classification (1636)	[45], [174], [210]
	models for clustering (908)	[294], [295], [350]
	models for dimension reduction (247)	[127], [215], [281]
	topic models (1089)	[32], [33], [137]
	models for regression (256)	[48], [147], [292]
	language model (271)	[24], [87], [156]
	graphical models (187)	[256], [262], [356]
	neural networks (412)	[173], [191], [218]
	mixture models (128)	[31], [164], [306]

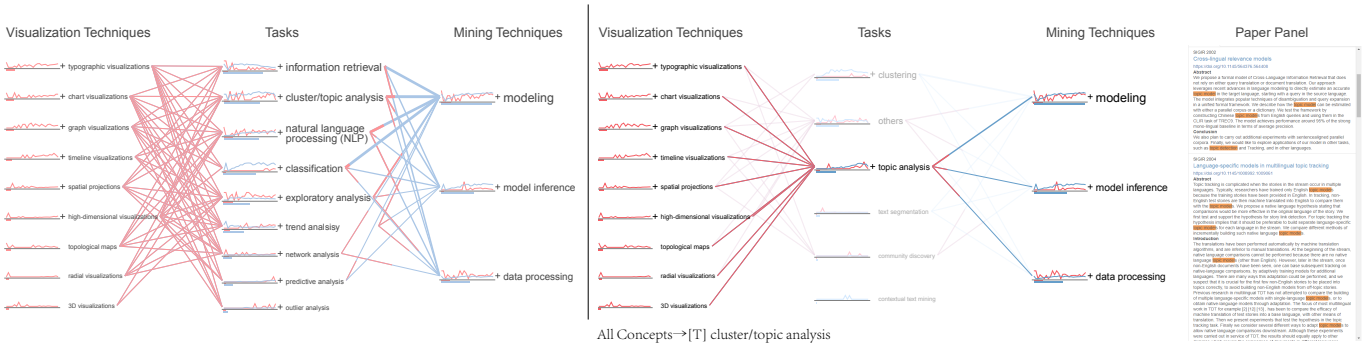


Fig. 6: Our hierarchical visualization of concept relationships. On the left, the initial overview shows relationships extracted from visualization papers in red and data mining papers in blue. On the right, a drill down operation has been applied to investigate tasks under the high level concept of *cluster/topic analysis*, and the timeline for task *topic analysis* is hovered, isolating connections to this task and revealing detailed statistics in a tooltip. The paper panel is populated with papers related to this task.

papers is encoded in the size of the concept label. Trends over time between visualization and data mining are revealed through spark lines appearing beside the concept label. Since the total number varies widely across concepts, we normalized the spark lines for each concept so that they reveal the relative number of papers at each year. The absolute number of papers in each concept is encoded in the horizontal bar charts under the spark lines.

Connections in the visualization are based on concept co-occurrences. The number of papers containing both concepts at the endpoints of an edge is encoded in the thickness of the edge. The color of the edge is split along the length of the edge to show the proportion of contributing papers from each research field. For example, in Fig. 6 on the left, the connection between the task "cluster/topic analysis" and the mining technique "modeling" shows that most co-occurrences of these concepts came from mining papers (edge is mostly blue). We found a similar proportional pattern between "natural language processing" and "model inference," but overall a lower number of co-occurrences (thinner edge).

Hovering on a concept label highlights all reachable edges and concepts while fading the others, thus revealing the co-occurring concepts across the dataset (see Fig. 6 right). Hovering on a timepoint in a spark line graph reveals a rich tooltip with precise numerical data. Since each spark line is independently scaled to maximize the visibility of trends, the precise values can be used to compare across spark lines. Selecting a concept populates the paper panel at the right to show titles, abstracts, and metadata for papers labelled with that concept. Selecting an edge populates the paper panel with papers containing both of the associated concepts. Target concepts appearing in the abstract text are highlighted in the paper panel for quick identification. Finally, the full text of any paper can be accessed by clicking its DOI link in the paper panel.

6 Results

In this section, we examine the **current practices** in visual text analytics by analyzing the three concept taxonomies, the connections between them, and the temporal trends revealed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2834341, IEEE Transactions on Visualization and Computer Graphics

by our visualization tool. We also discuss **potential research opportunities** by comparing the trends observed in the literature of visual text analytics to the trends in text mining.

6.1 Current Practices of Visual Text Analytics

This study demonstrates that the visualization tool based on our literature analysis can help researchers and practitioners to better understand the current practices in visual text analytics. In particular, we discuss different trends in visualization techniques, the major analysis tasks, frequently used mining techniques, and the connections between them.

6.1.1 Current practices of visualization techniques

Fig. 1(a) summarizes nine first-level visualization techniques, their temporal trends, and how frequently they are used. One can see that the usage patterns vary among different types of techniques. To better understand the current practices, we divided these techniques into three groups based on how frequently they were used.

The first group (Fig. 1A) contains the three most frequently used techniques: "typographic visualizations" (125 papers), "chart visualizations" (102 papers), and "graph visualizations" (95 papers). We observed that these popular techniques were the *traditional* ones: they were also frequently used in text visualization papers before 2000. Moreover, the proportion of papers that use these techniques has tended to increase the past few years.

To study this phenomenon in more detail, we drilled into the next level of "chart visualization." As shown in Fig. 7, "chart visualization" has four children: "line chart," "scatterplot," "table," and "other charts." After we switched to the temporal spark lines that display the absolute number of papers at each year, we observed an interesting "revival" of these techniques (Fig. 7). While these techniques were popular in early years, there was a time period (2000 to 2005) when these techniques were not used frequently. After this time period, researchers started to use these techniques again. The trend to use these simple yet effective techniques has been even stronger the past five years. This phenomenon is interesting because of two reasons. First, all four types of chart visualizations share similar patterns. Second, the percentage of visualization papers involved (39%) is large. We then checked the latest papers that used the techniques with the paper panel in the visualization tool. The paper panel reveals that in some cases, these techniques served a supportive role as part of a detail view or dashboard (e.g., in [212]), but in other cases, they were the main visualization components described in the papers (e.g., in [25]). One hypothesis regarding the revival of chart visualizations is due to the preference for simpler and more intuitive visualizations in order to reduce the learning curve of users. Another possible reason is that researchers tend to rely on more advanced learning methods instead of more complex visualizations to discover interesting patterns. For example, Berger et al. presented cite2vec [25], a

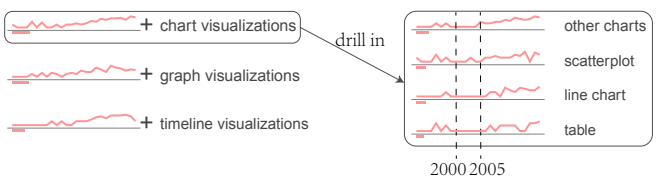


Fig. 7: All types of chart visualizations began to "revive" after 2005.

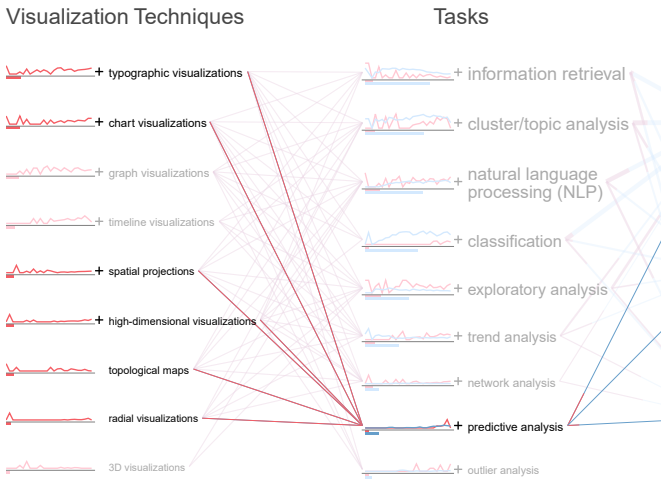


Fig. 8: Visualization techniques supporting the "predictive analysis" task.

visualization scheme that allows users to dynamically browse documents via how other documents use them. This usage-oriented exploration is enabled by projecting the words and documents into a common embedding space via word embedding. Here, a simple scatterplot was leveraged to visualize the embedding space. While the visualization was relatively simple, a variety of useful patterns could be found due to special properties of the embedded space.

The second group consists of four first-level techniques: "timeline visualizations" (50 papers), "spatial projections" (35 papers), "high-dimensional visualizations" (34 papers), and "topological maps" (33 papers) (Fig. 1B). These techniques are effective for specific types of data. For example, "timeline visualizations" are suitable for analyzing textual data with time stamps. "Topological maps" are intuitive choices for joint analysis of geographical information and text. This scenario illustrates that our visualization tool helps new researchers and practitioners to identify relevant visualization and mining techniques as well as the related papers for a certain type of data.

The last group (Fig. 1C) includes two first-level techniques that are not used very frequently: "radial visualization" (15 papers), "3D visualizations" (10 papers). Studying such techniques may help to discover the potential for rarely-used techniques and trigger the development of novel visualizations.

6.1.2 Current practices of analysis tasks

The task taxonomy consists of nine first-level concepts (Fig. 1(b)), which are divided into the following two groups.

The first group contains the most studied tasks in text visualization papers. Tasks in this group are "exploratory analysis," "natural language processing," "trend analysis," and "cluster/topic analysis." All these tasks have been studied in more than 50 visualization papers. Except for "trend analysis," all the tasks were frequently studied before 2000. Their temporal trends are also diverse. For example, "cluster/topic analysis" shows an upward trend between 2006 and 2014, while "exploratory analysis" experiences a surge after 2000.

The second group contains the less frequently studied tasks in the visualization field, namely "information retrieval," "network analysis," "classification," "outlier analysis," and "predictive analysis." We were a little surprised when we found that there were only

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2834341, IEEE Transactions on Visualization and Computer Graphics

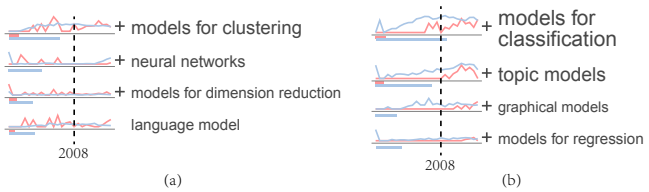


Fig. 9: Models used in text visualization papers: (a) traditional models used frequently before 2008; (b) models that attract more attention after 2008.

a few text visualization papers on classification. In our collection, there were several visualization papers on classification at IEEE VIS each year. After a careful examination of the relevant papers published at IEEE VIS in recent years, we found that most of them considered general data sources instead of textual data [206], [304]. Among the five tasks in this group, only "information retrieval" has shown a downward trend in the visualization field in recent years. The task concepts "classification," "outlier analysis," and "predictive analysis" have experienced an upward trend in recent years. However, in interpreting these trend lines one has to be cautious due to the small sample size, with fewer than 10 visualization papers for each concept.

Using our interactive visual interface, users may analyze and explore the relationships between the visualization techniques and tasks to understand which visualization techniques have been applied to support which tasks. For example, we selected the task "predictive analysis" for further investigation. Hovering over this task in the web-based visualization tool will highlight all visualization techniques that have been used to support the *predictive analysis* (Fig. 8). The visualization techniques supporting the "predictive analysis" task include: "typographic visualizations," "chart visualizations," "spatial projections," "high-dimensional visualizations," "topological maps," and "radial visualizations." Different groups of visualization techniques are applied to support various predictive analysis tasks. "Spatial projections" and "topological maps" are employed when predicting the user's demographic information. "High-dimensional visualizations," "topological maps," and "chart visualizations" are applied when predicting user actions in social media.

6.1.3 Current practices of mining techniques

The current practices of the three first-level mining techniques: "modeling," "model inference," and "data processing" are illustrated in Fig. 1(c). Most of the text visualization papers focus on "modeling" (131 papers) and "data processing" (110 papers), while fewer papers support "model inference" (17 papers). The spark lines show that the proportion of visualization papers focusing on "data processing" and "model inference" was the largest between 2000 and 2005 (Fig. 1D and Fig. 1E). After that, their popularity waned.

In contrast, "modeling" has continued to attract attention since 1995. To further study the temporal pattern, we drilled into the concept "modeling." The eight types of models used in text visualization papers appear in Fig. 9. They can be divided into two groups based on their temporal trends. The first group contains *traditional* models used frequently before 2008 (Fig. 9(a)). Models in this group include "clustering," "dimension reduction," "neural networks," and "language model." The second group consists of *trending* models that became more popular after 2008 (Fig. 9(b)). This group includes "classification," "topic models," "graphical

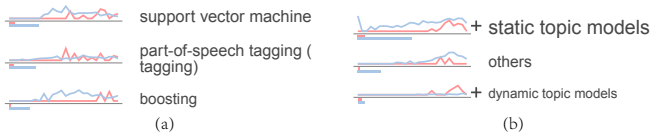


Fig. 10: Drilling into (a) models for classification; (b) topic models.

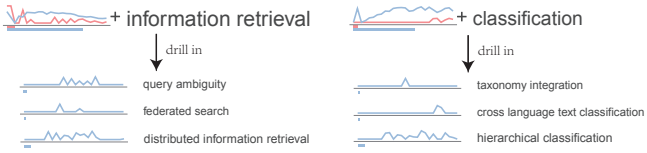


Fig. 11: Example tasks proposed by mining researchers.

models," and "regression." We further drilled into these second-level concepts to determine which specific techniques contribute to the aforementioned temporal trends. Fig. 10(a) and Fig. 10(b) show three specific models for classification and three types of topic models, respectively. The spark lines indicate that "support vector machine" and "boosting" contribute more to the trendiness of classification in contrast to "part-of-speech tagging." For topic models, the temporal trends of static and dynamic models are similar. Researchers and practitioners interested in text visualization can utilize our visualization tool to find more trending techniques and leverage these state-of-the-art methods in their work.

6.2 Investigating Research Opportunities

In this study, we illustrate how our visualization may help to identify potential research opportunities in a data-driven manner. This is achieved by revealing research gaps between text visualization and text mining research.

6.2.1 Opportunities learned by comparing analysis tasks

To understand the gaps between the text visualization and mining fields, we compared the paper distributions from these two fields under different tasks, as well as examined the connections between these concepts. In particular, we identified two main types of interesting tasks.

Tasks less frequently studied in the visualization field. Some tasks are considered by many mining papers, but by few text visualization papers, as shown in Fig. 1(b). For example, "classification" and "information retrieval" have been less studied in text visualization papers. In contrast, these two tasks have been extensively studied in the text mining field. Based on these observations, we summarize two opportunities.

Opportunity 1: Supporting tasks proposed by mining researchers. After we drilled into the hierarchy and examined more specific tasks, we recognized that many tasks proposed by text mining

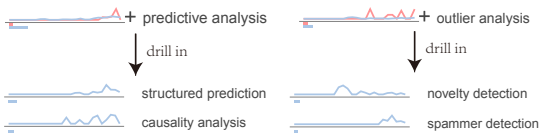


Fig. 12: Example tasks that can be better supported by tightly integrating interactive visualization with text mining techniques.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2834341, IEEE Transactions on Visualization and Computer Graphics

researchers are complex and/or interactive in nature, for which visual analytics research may be suitable. However, currently, the visualization field has not paid much attention to them. For example, by drilling into "information retrieval" (Fig. 11(a)), we identified tasks such as "query ambiguity," "federated search," and "distributed information retrieval," which have not been well studied in the visualization field. Among the children of "classification," the least studied tasks in the visualization field are "taxonomy integration," "cross language text classification," and "hierarchical classification" (Fig. 11(b)). Studying these tasks may help broaden the horizon of current visual text analytics research.

Opportunity 2: Integrating human knowledge to better support text mining tasks. When we explored the task taxonomy, tasks such as 'binary classification" and "recommendation" attracted our attention. These tasks are typical tasks for a mining paper. They have been less considered by the visualization field because they can be solved using an automatic algorithm. Usually, these tasks are well defined and the performance of the solution can be automatically evaluated. For these tasks, we believe that visual analytics can help improve the model performance by integrating human knowledge, especially when models do not work as expected. For example, to improve the performance of text classification, an interactive visualization can be developed to enable experts to effectively provide informative supervision at every stage of the classification pipeline. Such supervision can be performed through the identification of outliers in the training data, verification of labels of important data samples, and better parameter settings.

Tasks with insufficient coverage in both fields. We also noticed that several tasks had not yet attracted much attention from either the visualization or the mining field. Examples are "predictive analysis" and "outlier analysis" (Fig. 1(b)). After analyzing these tasks, we identified the following opportunity.

Opportunity 3: Supporting challenging tasks in text analysis. After drilling into "predictive analysis" and "outlier analysis," we found several tasks that were difficult to handle, even for human experts, including "causality analysis," "structured predication," and "novelty detection" (Fig. 12). Developing visual text analytics approaches that support such challenging tasks is an open research opportunity. To better support these tasks, we need to employ the full potential of both interactive visualization and text mining. One possible starting point is to study the state-of-the-art literature from both the visualization and mining fields and find a solution to tightly combine them by active learning or semi-supervised learning.

6.2.2 Opportunities learned by comparing mining techniques

We compared the visual text analytics papers with the text mining papers in terms of the mining techniques they used. Through our analysis, we identified the following three opportunities.

Opportunity 4: Incorporating state-of-the-art mining techniques. Connections between tasks and mining techniques demonstrate that a majority of mining techniques are not supported by existing text visualization papers (as shown by the bi-colored connections between (b) and (c) in Fig. 1). This gap can be observed by comparing the lengths of red segments with the lengths of blue segments. For each task, our visualization allows users to find the relevant state-of-the-art mining techniques. Leveraging these techniques may help in supporting more difficult tasks and in developing better visual analytics methods. Take topic modeling as an example. To find the state-of-the-art techniques, we drilled into a relevant mining technique: "static topic models." Examples of several static topic models are shown in Fig. 13. While the visualization field tends to use "Latent Dirichlet Allocation (LDA)," text mining researchers also use other topic models such as "spherical topic models" and a "correlated topic model." By observing the temporal trends, we discovered a recently-proposed model named "geometric Dirichlet means algorithm" [349] that is more computationally efficient than LDA and can handle larger numbers of documents. Accordingly, our analysis results can be leveraged to discover the state-of-the-art technique(s) from both the visualization and the mining fields. This can further advance the research and development of visual text analysis applications.

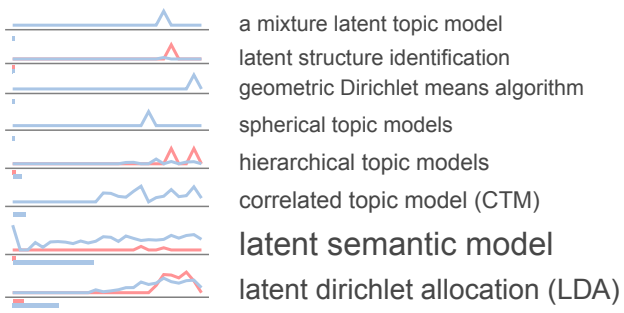


Fig. 13: Examples of static topic models and their temporal trends.

Opportunity 5: Opening the black-box of text mining. In addition, we recognized that the text mining field has produced a substantial amount of techniques that were specialized black-box models tailored to specific tasks. Examples are "mixture models," "neural networks," and "graphical models." An open research challenge that involves both text mining and visualization is to make these techniques understandable. Therefore, developing new visual analytics approaches for understanding the inner-workings of these models, which can steer users to better performance, is a gap that has a great deal of potential for innovative research.

Opportunity 6: Connecting big textual data with people. Textual data such as web pages, tweets, emails, instant messages, web click-streams, or CRM information, is flooding into the business world, academic community, and relevant governmental agencies. This data deluge is a large part of big data. In our exploration, we noticed that there were already some initial efforts in the visualization and mining fields for divining actionable information from the deluge. For example, the task concept "trend/pattern analysis" under "trend analysis" contains one recent paper on visually analyzing streaming textual data [207]. We also observed that there were a few initial efforts from the mining field. For example, Twitter, Inc. developed a large-scale topic modeling method for handling Twitter data [347] (under the task concept "cluster/topic analysis"). Despite the promising start in both fields, more research is needed for this research topic, especially for shaping visual analytics research that tightly integrates interactive visualization and text mining techniques to maximize the value of both in handling large-scale textual data.

7 Discussions and Reflections

Our work has studied three primary concepts, visualization techniques, mining techniques and analysis tasks. The analysis with the web-based visualization tool (Sec. 6) has disclosed the connection between mining and visualization techniques through analysis tasks,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2834341, IEEE Transactions on Visualization and Computer Graphics

as well as their temporal trends over time. By investigating the relationships between the three types of concepts, we schematically illustrated current practices and developments of visualization techniques, mining techniques, and analysis tasks. Popular research topics and potential emerging research topics were extracted by examining the connections between the three types of concepts and the gaps between them. One unique aspect of our survey is that the research question drove us to survey two distinct research fields, and the need to identify a bridge connecting them. Our goal is to provide an overview of the research related to text mining and visualization from the two fields, and foster more cross-pollination research. In this section, we introduce the by-product of our review work, lessons learned, and the limitations of our work.

7.1 Research By-Product

In addition to a comprehensive survey, our research has also delivered a research by-product, a visual-analytics-based literature analysis approach. The major feature of this approach is that it is based on the overall understanding and analysis of the major concepts (e.g., utilized techniques) mentioned in research papers. In this work, we mainly focused on analyzing two types of concepts, techniques and tasks. Inspired by Gupta and Manning's method [129] for automatically extracting key concepts, we developed a semi-automatic concept annotation method. By examining and analyzing the connections between different types of concepts, we provided a comprehensive overview of the on-going research efforts in the area of visual text analytics, including major research topics, their temporal trends, hot research topics, as well as less studied research topics. Based on the analysis of the aforementioned data, several research opportunities were identified and highlighted for the visualization domain. As the whole process is data-driven, this approach can be easily extended to other literature review work. It is particularly useful for an interdisciplinary review. Together with this approach, a visualization tool for navigating these concepts and the connections between them was also deployed as a web-based tool (http://visgroup.thss.tsinghua.edu.cn/textvis), which allows users to navigate through the major concepts in a publication dataset, their connections, and the corresponding papers.

7.2 Lessons Learned

This taxonomy was constructed in a semi-automatic way, where we iteratively and progressively extracted concepts and built a taxonomy for each type of concept. During this process, we learned several practical lessons, which are summarized in the remainder of this section.

Combination of data and knowledge. In order to provide a comprehensive overview of visual text analytics, we analyzed over 4,600 research papers and extracted around 300 keyword-based concepts. We settled on a semi-automatic method for concept extraction that is driven by both data and knowledge for two reasons. On the one hand, with such a large number of papers and concepts involved, manual labeling would have been very difficult and time-consuming. On the other hand, the accuracy of the automatic concept extraction method is not sufficient and depends on the coverage of the seed patterns. To overcome these issues, we employed a semi-automatic concept extraction method that tightly couples human knowledge with data (Fig. 4). Initially, we manually extracted a set of techniques and analysis tasks from the visualization papers that we were familiar with (knowledge-driven approach). Next, we extracted more concepts from the mining papers based on the manually extracted concepts and Gupta and Manning's method (data-driven approach). All the authors then worked together to verify the extracted concepts iteratively and resolved conflicts (knowledge-driven approach). The combination of the data- and knowledge-driven approaches improved both the quality of the concepts extracted and the labeling performance.

A similar approach that combines data and human knowledge (e.g., expert feedback) was also used to build the taxonomy. K-means clustering was employed to build the initial taxonomy. Then the authors, as well as several experts from machine learning and data mining refined and improved the taxonomy progressively. The experts preferred a balanced concept taxonomy that was not too deep or too wide. This is also consistent with a recent study [148]. Another useful lesson we learned is that human knowledge and experience are very useful to code concepts and build taxonomies. Typically, coding and taxonomy building are an iterative and progressive process. This process is most efficiently handled if experienced experts build an initial taxonomy that gets enriched by others, based on their understanding.

The **data-driven approach for finding research opportunities** is a good complement to the knowledge-based method. In the past, most literature reviews were manually carried out with the aim of examining the progress of a particular research topic and identifying emerging research opportunities. Typically, the breadth and depth of research opportunities depend on the experts' knowledge and their understanding of the research area. Our work contributes to this body of work by presenting the connections between different types of concepts, which allows experts to examine the overall trend of different research topics, as well as the gaps between different research fields in an interdisciplinary research area. This data-driven approach can be considered as a complement to the current knowledge-based approach for identifying emerging research opportunities.

7.3 Limitations

Although the developed semi-automatic approach sheds light on the research progress and emerging research opportunities in visual text analytics, we would like to note a few limitations.

When gathering the data, we tried our best to collect all relevant papers in both research fields. However, we may have missed some papers due to the large number of available venues and articles. To compensate for this, we will further extend our visualization tool to allow users to manually add papers that they believe to be relevant. We will then batch process and verify the submitted papers and merge the results into the existing concept taxonomies.

Another limitation is related to the calculations of connections between different concepts. We chose to leverage concept co-occurrences in the full-text to derive the connections between different concepts. This strategy may lead to some inaccurate connections between concepts. For example, one paper mentions using a scatterplot for an overview and a line chart for observing the temporal trend. Based on concept co-occurrences, we built all the possible connections between the four concepts, including "scatterplot" and "overview," "scatterplot" and "trend analysis," "line chart" and "overview," as well as "line chart" and "trend analysis." Here the concepts "scatterplot" and "trend analysis," as well as "line chart" and "overview," did not need to be connected. The correlation accuracy can be improved by employing more advanced approaches such as relationship extraction [112]. Another solution is to utilize a crowd-sourcing platform in order to collect

multiple annotations for each pair of connections. Accordingly, an interesting avenue of potential future work is to leverage a crowd-sourcing model such as M^3V [203] to infer the correct connection from noisy crowd-sourced labels.

8 Conclusions

In this work, we conducted a comprehensive survey based on 263 text visualization papers and 4,346 text mining papers that have been published between 1992 and 2017. With a semi-automatic, data-driven analysis, we identified and extracted three types of concepts. Two of the concepts, *visualization techniques* and *mining techniques*, summarize the research trends in the respective research fields, while the *analysis tasks* summarizes the goals of such research. Through statistically analyzing the relationships between the three types of concepts, we connected visualization techniques and mining techniques with analysis tasks serving as the bridge. In addition, a web-based visualization tool has been developed to facilitate the investigation of the major research trends in the area of text visualization, including the major techniques and tasks, their development over time, as well as the gaps between the visualization and mining fields.

We believe the data-driven analysis process developed in this work can be directly used to conduct literature analysis in other interdisciplinary research areas, such as interactive machine learning, bio-informatics visualization, or brain-inspired artificial intelligence. The key is to find an important intermediate concept that bridges the two fields. For example, for the research area of brain-inspired artificial intelligence, different types of neurons and their operating mechanisms might be a candidate intermediate concept that connects neuro-science and artificial intelligence. In this survey we have shown that using such an intermediate concept may help to narrow the gaps between two research domains and provide useful insights into an interdisciplinary area, which can foster a better understanding of the research field and opens promising avenues for future research.

References

[1] "Apache Lucene Core," https://lucene.apache.org/core/, accessed: 31-Dec-2017.

[2] S. Abbar, S. Amer-Yahia, P. Indyk, and S. Mahabadi, "Real-time recommendation of diverse related articles," in *WWW*, 2013, pp. 1–12.

[3] A. Abdul-Rahman, J. Lein, K. Coles, E. Maguire, M. Meyer, M. Wynne, C. R. Johnson, A. Trefethen, and M. Chen, "Rule-based visual mappings - with a case study on poetry visualization," *Computer Graphics Forum*, vol. 32, no. 3, pp. 381–390, 2013.

[4] ACL, "Learning bilingual word embeddings with (almost) no bilingual data," 2017, pp. 451–462.

[5] S. Afzal, R. Maciejewski, Y. Jang, N. Elmqvist, and D. S. Ebert, "Spatial text visualization using automatic typographic maps," *IEEE TVCG*, vol. 18, no. 12, pp. 2556–2564, 2012.

[6] C. C. Aggarwal, *Machine Learning for Text*. Springer, 2018.

[7] C. C. Aggarwal and C. Zhai, *Mining Text Data*. Springer Science & Business Media, 2012.

[8] S. Agrawal, K. Chakrabarti, S. Chaudhuri, V. Ganti, A. C. Konig, and D. Xin, "Exploiting web search engines to search structured databases," in *WWW*, 2009, pp. 501–510.

[9] J. Albrecht, R. Hwa, and G. E. Marai, "The Chinese room: Visualization and interaction to understand and correct ambiguous machine translation," *Computer Graphics Forum*, vol. 28, no. 3, pp. 1047–1054, 2009.

[10] A. B. Alencar, M. C. F. de Oliveira, and F. V. Paulovich, "Seeing beyond reading: A survey on visual text analytics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 476–492, 2012.

[11] E. Alexander and M. Gleicher, "Task-driven comparison of topic models," *IEEE TVCG*, vol. 22, no. 1, pp. 320–329, 2016.

[12] J. Alsakran, Y. Chen, Y. Zhao, J. Yang, and D. Luo, "STREAMIT: Dynamic visualization and interactive exploration of text streams," in *IEEE PacificVis*, 2011, pp. 131–138.

[13] E. Amitay, N. Har'El, R. Sivan, and A. Soffer, "Web-a-where: geotagging web content," in *ACM SIGIR*, 2004, pp. 273–280.

[14] R. K. Ando, "Latent semantic space: Iterative scaling improves precision of inter-document similarity measurement," in *ACM SIGIR*, 2000, pp. 216–223.

[15] K. Andrews, W. Kienreich, V. Sabol, J. Becker, G. Droschl, F. Kappe, M. Granitzer, P. Auer, and K. Tochtermann, "The InfoSky visual explorer: Exploiting hierarchical structure and document similarities," *Information Visualization*, vol. 1, no. 3–4, pp. 166–181, 2002.

[16] D. Angus, A. Smith, and J. Wiles, "Conceptual Recurrence Plots: Revealing patterns in human discourse," *IEEE TVCG*, vol. 18, no. 6, pp. 988–997, 2012.

[17] J. A. Aslam and M. Montague, "Models for metasearch," in *ACM SIGIR*, 2001, pp. 276–284.

[18] R. Atterer, M. Wnuk, and A. Schmidt, "Knowing the user's every move: User activity tracking for website usability evaluation and implicit interaction," in *WWW*, 2006, pp. 203–212.

[19] R. Baeza-Yates, "Visualization of large answers in text databases," in *AVI*, 1996, pp. 101–107.

[20] N. Bansal and N. Koudas, "BlogScope: Spatio-temporal analysis of the blogosphere," in *WWW*, 2007, pp. 1269–1270.

[21] H. Bast, A. Chitea, F. Suchanek, and I. Weber, "ESTER: efficient search on text, entities, and relations," in *ACM SIGIR*, 2007, pp. 671–678.

[22] F. Beck, S. Koch, and D. Weiskopf, "Visual analysis and dissemination of scientific literature collections with SurVis," *IEEE TVCG*, vol. 22, no. 1, pp. 180–189, 2016.

[23] F. Beck and D. Weiskopf, "Word-sized graphics for scientific texts," *IEEE TVCG*, vol. 23, no. 6, pp. 1576–1587, 2017.

[24] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *JMLR*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[25] M. Berger, K. Mcdonough, and L. M. Seversky, "cite2vec: Citation-driven document exploration via word embeddings," *IEEE TVCG*, vol. 23, no. 1, pp. 691–700, 2017.

[26] T. Bergstrom and K. Karahalios, "Conversation clusters: Grouping conversation topics through human-computer dialog," in *ACM SIGCHI*, 2009, pp. 2349–2352.

[27] E. A. Bier, S. K. Card, and J. W. Bodnar, "Principles and tools for collaborative entity-based intelligence analysis," *IEEE TVCG*, vol. 16, no. 2, pp. 178–191, 2010.

[28] E. A. Bier, S. K. Card, and J. W. Bodnar, "Entity-based collaboration tools for intelligence analysis," in *IEEE VAST*, 2008, pp. 99–106.

[29] C. Bishop, J. Winn, and T. Diethe, "Model-based Machine Learning," http://mbmlbook.com/, 2015, accessed: 31-Dec-2017.

[30] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[31] D. M. Blei and P. I. Frazier, "Distance dependent Chinese restaurant processes," *JMLR*, vol. 12, no. Aug, pp. 2461–2488, 2011.

[32] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *ICML*, 2006, pp. 113–120.

[33] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *JMLR*, vol. 3, no. Jan, pp. 993–1022, 2003.

[34] M. Bögl, P. Filzmoser, T. Gschwandtner, T. Lammarsch, R. A. Leite, S. Miksch, and A. Rind, "Cycle plot revisited: Multivariate outlier detection using a distancebased abstraction," *Computer Graphics Forum*, vol. 36, no. 3, pp. 227–238.

[35] H. Bosch, D. Thom, F. Heimerl, E. Püttmann, S. Koch, R. Krüger, M. Wörner, and T. Ertl, "ScatterBlogs2: Real-time monitoring of microblog messages through user-guided filtering," *IEEE TVCG*, vol. 19, no. 12, pp. 2022–2031, 2013.

[36] S. Bostandjiev, J. O'Donovan, and T. Höllerer, "LinkedVis: Exploring social and semantic career recommendations," in *IUI*, 2013, pp. 107–116.

[37] N. Boukhelifa, F. Chevalier, and J.-D. Fekete, "Real-time aggregation of Wikipedia data for visual analytics," in *IEEE VAST*, 2010, pp. 147–154.

[38] L. Bradel, C. North, L. House, and S. Leman, "Multi-model semantic interaction for text analytics," in *IEEE VAST*, 2014, pp. 163–172.

[39] M. Brehmer, S. Ingram, J. Stray, and T. Munzner, "Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists," *IEEE TVCG*, vol. 20, no. 12, pp. 2271–2280, 2014.

[40] A. J. Brockmeier, T. Mu, S. Ananiadou, and J. Y. Goulermas, "Quantifying the informativeness of similarity measurements," *JMLR*, vol. 18, no. 76, pp. 1–61, 2017.

[41] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel, "A semantic approach to contextual advertising," in *ACM SIGIR*, 2007, pp. 559–566.

[42] M. Brooks, S. Amershi, B. Lee, S. M. Drucker, A. Kapoor, and P. Simard, "FeatureInsight: Visual support for error-driven feature ideation in text classification," in *IEEE VAST*, 2015, pp. 105–112.

[43] P. Butler, P. Chakraborty, and N. Ramakrishan, "The Deshredder: A visual analytic approach to reconstructing shredded documents," in *IEEE VAST*, 2012, pp. 113–122.

[44] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Computational intelligence magazine*, vol. 9, no. 2, pp. 48–57, 2014.

[45] R. Campos, S. Canuto, T. Salles, C. C. de Sá, and M. A. Gonçalves, "Stacking bagged and boosted forests for effective automated classification," in *ACM SIGIR*, 2017, pp. 105–114.

[46] N. Cao, Y. R. Lin, X. Sun, and D. Lazer, "Whisper: Tracing the spatiotemporal process of information diffusion in real time," *IEEE TVCG*, vol. 18, no. 12, pp. 2649–2658, 2012.

[47] N. Cao, J. Sun, Y. R. Lin, D. Gotz, S. Liu, and H. Qu, "FacetAtlas: Multifaceted visualization for rich text corpora," *IEEE TVCG*, vol. 16, no. 6, pp. 1172–1181, 2010.

[48] Z. Cao, W. Li, S. Li, and F. Wei, "Improving multi-document summarization via text classification," in *AAAI*, 2017, pp. 3053–3059.

[49] G. Carenini, R. T. Ng, and A. Pauls, "Interactive multimedia summaries of evaluative text," in *IUI*, 2006, pp. 124–131.

[50] G. Carenini and L. Rizoli, "A multimedia interface for facilitating comparisons of opinions," in *IUI*, 2009, pp. 325–334.

[51] B. Carterette, J. Allan, and R. Sitaraman, "Minimal test collections for retrieval evaluation," in *ACM SIGIR*, 2006, pp. 268–275.

[52] Q. Castellà and C. Sutton, "Word storms: Multiples of word clouds for visual comparison of documents," in *WWW*, 2014, pp. 665–676.

[53] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl, "Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition," in *IEEE VAST*, 2012, pp. 143–152.

[54] B. Chan, L. Wu, J. Talbot, M. Cammarano, and P. Hanrahan, "Vispedia: Interactive visual exploration of Wikipedia data via search-based integration," *IEEE TVCG*, vol. 14, no. 6, pp. 1213–1220, 2008.

[55] S. Chandrasegaran, S. K. Badam, L. Kisselburgh, K. Ramani, and N. Elmqvist, "Integrating visual analytics support for grounded theory practice in qualitative text analysis," *Computer Graphics Forum*, vol. 36, no. 3, pp. 201–212, 2017.

[56] M.-W. Chang and C. Collins, "Exploring entities in text with descriptive non-photorealistic rendering," in *IEEE PacificVis*, 2013, pp. 9–16.

[57] C. Chen, F. Ibekwe-Sanjuan, E. Sanjuan, and C. Weaver, "Visual analysis of conflicting opinions," in *IEEE VAST*, 2006, pp. 59–66.

[58] F. Chen, P. Chiu, and S. Lim, "Topic modeling of document metadata for visualizing collaborations over time," in *IUI*, 2016, pp. 108–117.

[59] S. Chen, S. Chen, Z. Wang, J. Liang, X. Yuan, N. Cao, and Y. Wu, "D-Map: Visual analysis of ego-centric information diffusion patterns in social media," in *IEEE VAST*, 2017, pp. 41–50.

[60] Y. Chen, L. Wang, M. Dong, and J. Hua, "Exemplar-based visualization of large document corpus," *IEEE TVCG*, vol. 15, no. 6, pp. 1161–1168, 2009.

[61] F. Chevalier, S. Huot, and J.-D. Fekete, "WikipediaViz: Conveying article quality for casual Wikipedia readers," in *IEEE PacificVis*, 2010, pp. 49–56.

[62] M.-T. Chi, S.-S. Lin, S.-Y. Chen, C.-H. Lin, and T.-Y. Lee, "Morphable word clouds for time-varying text data visualization," *IEEE TVCG*, vol. 21, no. 12, pp. 1415–1426, 2015.

[63] I. Cho, W. Dou, D. X. Wang, E. Sauda, and W. Ribarsky, "VAiRoma: A visual analytics system for making sense of places, times, and events in Roman history," *IEEE TVCG*, vol. 22, no. 1, pp. 210–219, 2016.

[64] F. Y. Choi, "Advances in domain independent linear text segmentation," in *ACL*, 2000, pp. 26–33.

[65] J. Choo, C. Lee, C. K. Reddy, and H. Park, "UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization," *IEEE TVCG*, vol. 19, no. 12, pp. 1992–2001, 2013.

[66] J.-K. Chou and C.-K. Yang, "PaperVis: Literature review made easy," *Computer Graphics Forum*, vol. 30, no. 3, pp. 721–730, 2011.

[67] J. Chu-Carroll, J. Prager, K. Czuba, D. Ferrucci, and P. Duboue, "Semantic search via XML fragments: a high-precision approach to IR," in *ACM SIGIR*, 2006, pp. 445–452.

[68] J. Chuang, C. D. Manning, and J. Heer, "Termite: Visualization techniques for assessing textual topic models," in *AVI*, 2012, pp. 74–77.

[69] J. Chuang, D. Ramage, C. Manning, and J. Heer, "Interpretation and trust: Designing model-driven visualizations for text analysis," in *ACM SIGCHI*, 2012, pp. 443–452.

[70] B. chul Kwon, W. Javed, S. Ghani, N. Elmqvist, J. S. Yi, and D. S. Ebert, "Evaluating the role of time in investigative analysis of document collections," *IEEE TVCG*, vol. 18, no. 11, pp. 1992–2004, 2012.

[71] H. Chung, S. Yang, N. Massjouni, C. Andrews, R. Kanna, and C. North, "VizCept: Supporting synchronous collaboration for constructing visualizations in intelligence analysis," in *IEEE VAST*, 2010, pp. 107–114.

[72] C. L. Clarke, "Controlling overlap in content-oriented XML retrieval," in *ACM SIGIR*, 2005, pp. 314–321.

[73] A. Cockburn, C. Gutwin, and J. Alexander, "Faster document navigation with space-filling thumbnails," in *ACM SIGCHI*, 2006, pp. 1–10.

[74] C. Collins, S. Carpendale, and G. Penn, "Visualization of uncertainty in lattices to support decision-making," in *EuroVis*, 2007, pp. 51–58.

[75] C. Collins, S. Carpendale, and G. Penn, "DocuBurst: Visualizing document content using language structure," *Computer Graphics Forum*, vol. 28, no. 3, pp. 1039–1046, 2009.

[76] C. Collins, G. Penn, and S. Carpendale, "Bubble sets: Revealing set relations with isocontours over existing visualizations," *IEEE TVCG*, vol. 15, no. 6, pp. 1009–1016, 2009.

[77] C. Collins, F. B. Viégas, and M. Wattenberg, "Parallel tag clouds to explore and analyze faceted text corpora," in *IEEE VAST*, 2009, pp. 91–98.

[78] M. Correll, M. Witmore, and M. Gleicher, "Exploring collections of tagged text for literary scholarship," *Computer Graphics Forum*, vol. 30, no. 3, pp. 731–740, 2011.

[79] M. A. Correll, E. C. Alexander, and M. Gleicher, "Quantity estimation in visualizations of tagged text," in *ACM SIGCHI*, 2013, pp. 2697–2706.

[80] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[81] P. J. Crossno, D. M. Dunlavy, and T. M. Shead, "LSAView: A tool for visual exploration of latent semantic modeling," in *IEEE VAST*, 2009, pp. 83–90.

[82] A. M. Cuadros, F. V. Paulovich, R. Minghim, and G. P. Telles, "Point placement by phylogenetic trees and its application to visual analysis of document collections," in *IEEE VAST*, 2007, pp. 99–106.

[83] Q. Cui, B. Gao, J. Bian, S. Qiu, H. Dai, and T.-Y. Liu, "KNET: A general framework for learning word embedding using morphological knowledge," *ACM TOIS*, vol. 34, no. 1, pp. 4:1–4:25, Aug. 2015.

[84] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. J. Gao, X. Tong, and H. Qu, "TextFlow: Towards better understanding of evolving topics in text," *IEEE TVCG*, vol. 17, no. 12, pp. 2412–2421, 2011.

[85] W. Cui, Y. Wu, S. Liu, F. Wei, M. Zhou, and H. Qu, "Context-preserving, dynamic word cloud visualization," *IEEE Computer Graphics and Applications*, vol. 30, no. 6, pp. 42–53, 2010.

[86] W. Cui, S. Liu, Z. Wu, and H. Wei, "How hierarchical topics evolve in large text corpora," *IEEE TVCG*, vol. 20, no. 12, pp. 2281–2290, 2014.

[87] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *ICML*, 2017, pp. 933–941.

[88] M. Deller, S. Agne, A. Ebert, A. Dengel, H. Hagen, B. Klein, M. Bender, T. Bernardin, and B. Hamann, "Managing a document-based information space," in *IUI*, 2008, pp. 119–128.

[89] Z.-H. Deng, H. Yu, and Y. Yang, "Identifying sentiment words using an optimization model with L1 regularization," in *AAAI*, 2016, pp. 115–121.

[90] S. Dezdar and A. Sulaiman, "Successful enterprise resource planning implementation: Taxonomy of critical factors," *Industrial Management & Data Systems*, vol. 109, no. 8, pp. 1037–1052, 2009.

[91] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," in *IEEE VAST*, 2010, pp. 115–122.

[92] E. Diemert and G. Vandelle, "Unsupervised query categorization using automatically-built concept graphs," in *WWW*, 2009, pp. 461–470.

[93] M. Dörk, S. Carpendale, C. Collins, and C. Williamson, "VisGets: Coordinated visualizations for web-based information exploration and discovery," *IEEE TVCG*, vol. 14, no. 6, pp. 1205–1212, 2008.

[94] W. Dou, I. Cho, O. Eltayeby, J. Choo, X. Wang, and W. Ribarsky, "DemographicVis: Analyzing demographic information based on user generated content," in *IEEE VAST*, 2015, pp. 57–64.

[95] W. Dou and S. Liu, "Topic- and time-oriented visual text analysis," *IEEE computer graphics and applications*, vol. 36, no. 4, pp. 8–13, 2016.

[96] W. Dou, X. Wang, R. Chang, and W. Ribarsky, "ParallelTopics: A probabilistic approach to exploring document collections," in *IEEE VAST*, 2011, pp. 231–240.

[97] W. Dou, X. Wang, D. Skau, and W. Ribarsky, "LeadLine: Interactive visual analysis of text data through event identification and exploration," in *IEEE VAST*, 2012, pp. 93–102.

[98] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky, "HierarchicalTopics: Visually exploring large text collections using topic hierarchies," *IEEE TVCG*, vol. 19, no. 12, pp. 2002–2011, 2013.

[99] N. Du, Y. Liang, M.-F. Balcan, M. Gomez-Rodriguez, H. Zha, and L. Song, "Scalable influence maximization for multiple products in continuous-time diffusion networks," *JMLR*, vol. 18, no. 2, pp. 1–45, 2017.

[100] Y. Du, C. Xu, and D. Tao, "Privileged matrix factorization for collaborative filtering," in *IJCAI*, 2017, pp. 1610–1616.

[101] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins, "Visualizing tags over time," *ACM TWEB*, vol. 1, no. 2, 2007.

[102] S. Dumais and H. Chen, "Hierarchical classification of web content," in *ACM SIGIR*, 2000, pp. 256–263.

[103] S. G. Eick, J. Mauger, and A. Ratner, "Visualizing the performance of computational linguistics algorithms," in *IEEE VAST*, 2006, pp. 151–157.

[104] M. El-Assady, V. Gold, C. Acevedo, C. Collins, and D. Keim, "ConToVi: Multi-party conversation exploration using topic-space views," *Computer Graphics Forum*, vol. 35, pp. 431–440, 2016.

[105] M. El-Assady, R. Sevastjanova, B. Gipp, D. Keim, and C. Collins, "NEREx: Named-entity relationship exploration in multi-party conversations," *Computer Graphics Forum*, vol. 36, no. 3, pp. 213–225, 2017.

[106] M. El-Assady, R. Sevastjanova, F. Sperrle, D. Keim, and C. Collins, "Progressive learning of topic modeling parameters: A visual analytics framework," *IEEE TVCG*, vol. 24, no. 1, pp. 382–391, 2018.

[107] A. Endert, P. Fiaux, and C. North, "Semantic interaction for visual text analytics," in *ACM SIGCHI*, 2012, pp. 473–482.

[108] A. Endert, S. Fox, D. Maiti, S. Leman, and C. North, "The semantics of clustering: Analysis of user-generated spatializations of text documents," in *AVI*, 2012, pp. 555–562.

[109] S. G. Esparza, M. P. O'Mahony, and B. Smyth, "CatStream: Categorising tweets for user profiling and stream filtering," in *IUI*, 2013, pp. 25–36.

[110] P. Federico, F. Heimerl, S. Koch, and S. Miksch, "A survey on visual approaches for analyzing scientific literature and patents," *IEEE TVCG*, vol. 23, no. 9, pp. 2179–2198, 2017.

[111] C. Felix, A. V. Pandey, and E. Bertini, "TextTile: An interactive visualization tool for seamless exploratory analysis of structured data and unstructured text," *IEEE TVCG*, vol. 23, no. 1, pp. 161–170, 2017.

[112] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," in *ACL*, 2005, pp. 363–370.

[113] D. Fisher, A. Hoff, G. Robertson, and M. Hurst, "Narratives: A visualization to track narrative events as they develop," in *IEEE VAST*, 2008, pp. 115–122.

[114] D. Fried and S. G. Kobourov, "Maps of computer science," in *IEEE PacificVis*, 2014, pp. 113–120.

[115] S. Fu, J. Zhao, W. Cui, and H. Qu, "Visual analysis of MOOC forums with iForum," *IEEE TVCG*, vol. 23, no. 1, pp. 201–210, 2016.

[116] N. Fuhr and K. Großjohann, "XIRQL: A query language for information retrieval in XML documents," in *ACM SIGIR*, 2001, pp. 172–180.

[117] K. Fujimura, S. Fujimura, T. Matsubayashi, T. Yamada, and H. Okuda, "Topigraphy: Visualization for large-scale tag clouds," in *WWW*, 2008, pp. 1087–1088.

[118] J. Fulda, M. Brehmer, and T. Munzner, "TimeLineCurator: Interactive authoring of visual timelines from unstructured text," *IEEE TVCG*, vol. 22, no. 1, pp. 300–309, 2016.

[119] S. Gad, W. Javed, S. Ghani, N. Elmqvist, T. Ewing, K. N. Hampton, and N. Ramakrishnan, "ThemeDelta: Dynamic segmentations over temporal topic models," *IEEE TVCG*, vol. 21, no. 5, pp. 672–685, 2015.

[120] Q. Gan, M. Zhu, M. Li, T. Liang, Y. Cao, and B. Zhou, "Document visualization: An overview of current research," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 6, no. 1, pp. 19–36, 2014.

[121] T. Gao, J. R. Hullman, E. Adar, B. Hecht, and N. Diakopoulos, "NewsViews: An automated pipeline for creating custom geovisualizations for news," in *ACM SIGCHI*, 2014, pp. 3005–3014.

[122] M. Glueck, M. P. Naeini, F. Doshi-Velez, F. Chevalier, A. Khan, D. Wigdor, and M. Brudno, "PhenoLines: Phenotype comparison visualizations for disease subtyping via topic models," *IEEE TVCG*, vol. 24, no. 1, pp. 371–381, 2018.

[123] V. Gold, C. Rohrdantz, and M. El-Assady, "Exploratory text analysis using lexical episode plots," in *EuroVis*, 2015, pp. 85–89.

[124] E. Gomez-Nieto, R. F. San, P. Pagliosa, W. Casaca, E. S. Helou, M. C. de Oliveira, and L. G. Nonato, "Similarity preserving snippet-based visualization of web search results," *IEEE TVCG*, vol. 20, no. 3, pp. 457–470, 2014.

[125] C. Görg, Z. Liu, J. Kihm, J. Choo, H. Park, and J. Stasko, "Combining computational analyses and interactive visualization for document exploration and sensemaking in Jigsaw," *IEEE TVCG*, vol. 19, no. 10, pp. 1646–1663, 2013.

[126] E. Graells-Garrido, M. Lalmas, and R. Baeza-Yates, "Data portraits and intermediary topics: Encouraging exploration of politically diverse profiles," in *IUI*, 2016, pp. 228–240.

[127] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation," in *IJCAI*, 2017, pp. 1753–1759.

[128] Y. Guo and M. Xiao, "Cross language text classification via subspace co-regularized multi-view learning," *ICML*, pp. 1615–1622, 2012.

[129] S. Gupta and C. D. Manning, "Analyzing the dynamics of research by extracting key aspects of scientific papers," in *IJCNLP*, 2011, pp. 1–9.

[130] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," *Journal of emerging technologies in web intelligence*, vol. 1, no. 1, pp. 60–76, 2009.

[131] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. MIT press, 2001.

[132] K. Hartmann, S. Schlechtweg, R. Helbing, and T. Strothotte, "Knowledge-supported graphical illustration of texts," in *AVI*, 2002, pp. 300–307.

[133] T. H. Haveliwala, "Topic-sensitive pagerank," in *WWW*, 2002, pp. 517–526.

[134] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "ThemeRiver: Visualizing thematic changes in large document collections," *IEEE TVCG*, vol. 8, no. 1, pp. 9–20, 2002.

[135] S. Havre, B. Hetzler, and L. Nowell, "ThemeRiver: Visualizing theme changes over time," in *IEEE InfoVis*, 2000, pp. 115–123.

[136] S. Havre, E. Hetzler, K. Perrine, E. Jurrus, and N. Miller, "Interactive visualization of multiple query results," in *IEEE InfoVis*, 2001, pp. 105–112.

[137] J. He, Z. Hu, T. Berg-Kirkpatrick, Y. Huang, and E. P. Xing, "Efficient correlated topic modeling with topic embedding," in *ACM SIGKDD*, 2017, pp. 225–233.

[138] M. A. Hearst, "TileBars: Visualization of term distribution information in full text information access," in *SIGCHI*, 1995, pp. 59–66.

[139] M. A. Hearst and J. O. Pedersen, "Visualizing information retrieval results: a demonstration of the TileBar interface," in *ACM SIGCHI*, 1996, pp. 394–395.

[140] F. Heimerl, S. Koch, H. Bosch, and T. Ertl, "Visual classifier training for text document retrieval," *IEEE TVCG*, vol. 18, no. 12, pp. 2839–2848, 2012.

[141] F. Heimerl, Q. Han, S. Koch, and T. Ertl, "CiteRivers: Visual analytics of citation patterns," *IEEE TVCG*, vol. 22, no. 1, pp. 190–199, 2016.

[142] F. Heimerl, M. John, Q. Han, S. Koch, and T. Ertl, "DocuCompass: Effective exploration of document landscapes," in *IEEE VAST*, 2016, pp. 11–20.

[143] B. Hetzler, P. Whitney, L. Martucci, and J. Thomas, "Multi-faceted insight through interoperable visual information analysis paradigms," in *IEEE InfoVis*, 1998, pp. 137–144.

[144] E. G. Hetzler, V. L. Crow, D. A. Payne, and A. E. Turner, "Turning the bucket of text into a pipe," in *IEEE InfoVis*, 2005, pp. 89–94.

[145] U. Hinrichs, S. Forlini, and B. Moynihan, "Speculative practices: Utilizing InfoVis to explore untapped literary collections," *IEEE TVCG*, vol. 22, no. 1, pp. 429–438, 2016.

[146] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015.

[147] Q. M. Hoang, T. N. Hoang, and K. H. Low, "A generalized stochastic variational Bayesian hyperparameter learning framework for sparse spectrum Gaussian process regression," in *AAAI*, 2017, pp. 2007–2014.

[148] T. Höllt, N. Pezzotti, V. van Unen, F. Koning, B. P. F. Lelieveldt, and A. Vilanova, "CyteGuide: Visual guidance for hierarchical single-cell analysis," *IEEE TVCG*, vol. 24, no. 1, pp. 739–748, 2018.

[149] E. Hoque and G. Carenini, "ConVis: A visual text analytic system for exploring blog conversations," vol. 33, no. 3, pp. 221–230, 2014.

[150] E. Hoque and G. Carenini, "MultiConVis: A visual text analytics system for exploring a collection of online conversations," in *IUI*, 2016, pp. 96–107.

[151] E. Hoque, S. Joty, L. Marquez, and G. Carenini, "CQAVis: Visual text analytics for community question answering," in *IUI*, 2017, pp. 161–172.

[152] M. S. Hossain, P. Butler, A. P. Boedihardjo, and N. Ramakrishnan, "Storytelling in entity networks to support intelligence analysts," in *ACM SIGKDD*, 2012, pp. 1375–1383.

[153] C.-K. Hsieh, L. Yang, Y. Cui, T.-Y. Lin, S. Belongie, and D. Estrin, "Collaborative metric learning," in *WWW*, 2017, pp. 193–201.

[154] J. Hu, G. Wang, F. Lochovsky, J.-t. Sun, and Z. Chen, "Understanding user's query intent with Wikipedia," in *WWW*, 2009, pp. 471–480.

[155] M. Hu, K. Wongsuphasawat, and J. Stasko, "Visualizing social media content with SentenTree," *IEEE TVCG*, vol. 23, no. 1, pp. 621–630, 2017.

[156] J. Huang, J. Gao, J. Miao, X. Li, K. Wang, F. Behr, and C. L. Giles, "Exploring web scale language models for search query processing," in *WWW*, 2010, pp. 451–460.

[157] J. Hullman, N. Diakopoulos, and E. Adar, "Contextifier: Automatic generation of annotated stock visualizations," in *ACM SIGCHI*, 2013, pp. 2707–2716.

[158] F. Ieva, A. M. Paganoni, and N. Tarabelloni, "Covariance-based clustering in multivariate and functional data analysis," *JMLR*, vol. 17, no. 1, pp. 4985–5005, 2016.

[159] Indratmo, J. Vassileva, and C. Gutwin, "Exploring blog archives with interactive visualization," in *AVI*, 2008, pp. 39–46.

[160] N. Ireson, F. Ciravegna, M. E. Califf, D. Freitag, N. Kushmerick, and A. Lavelli, "Evaluating machine learning for information extraction," in *ICML*, 2005, pp. 345–352.

[161] E. Isaacs, K. Damico, S. Ahern, E. Bart, and M. Singhal, "Footprints: A visual search tool that supports discovery and coverage tracking," *IEEE TVCG*, vol. 20, no. 12, pp. 1793–802, 2014.

[162] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. D. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko, "vispubdata. org: A metadata collection about IEEE Visualization (VIS) publications," *IEEE TVCG*, vol. 23, no. 9, pp. 2199–2206, 2017.

[163] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller, "Visualization as seen through its research paper keywords," *IEEE TVCG*, vol. 23, no. 1, pp. 771–780, 2017.

[164] K. Ishiguro, I. Sato, and N. Ueda, "Averaged collapsed variational bayes inference," *JMLR*, vol. 18, no. 1, pp. 1–29, 2017.

[165] M. Itoh, N. Yoshinaga, M. Toyoda, and M. Kitsuregawa, "Analysis and visualization of temporal changes in bloggers' activities and interests," in *IEEE PacificVis*, 2012, pp. 57–64.

[166] S. Jameel, Z. Bouraoui, and S. Schockaert, "MEmbER: Max-Margin based embeddings for entity retrieval," in *ACM SIGIR*, 2017, pp. 783–792.

[167] S. Jänicke, J. Focht, and G. Scheuermann, "Interactive visual profiling of musicians," *IEEE TVCG*, vol. 22, no. 1, pp. 200–209, 2016.

[168] S. Jänicke, G. Franzini, M. Cheema, and G. Scheuermann, "Visual text analysis in digital humanities," *Computer Graphics Forum*, vol. 36, no. 6, pp. 226–250, 2017.

[169] S. Jänicke and D. J. Wrisley, "Interactive visual alignment of medieval text versions," 2017.

[170] M. Jankowska, V. Kešelj, and E. Milios, "Relative N-gram signatures: Document visualization at the level of character N-grams," in *IEEE VAST*, 2012, pp. 103–112.

[171] A. Jatowt, Y. Kawai, and K. Tanaka, "Visualizing historical content of web pages," in *WWW*, 2008, pp. 1221–1222.

[172] X. Ji and H. Zha, "Domain-independent text segmentation using anisotropic diffusion and dynamic programming," in *ACM SIGIR*, 2003, pp. 322–329.

[173] R. Johnson and T. Zhang, "Semi-supervised convolutional neural networks for text categorization via region embedding," in *NIPS*, 2015, pp. 919–927.

[174] B. Joshi, M. R. Amini, I. Partalas, F. Iutzeler, and Y. Maximov, "Aggressive sampling for multi-class to binary reduction with applications to text classification," in *NIPS*, 2017, pp. 4159–4168.

[175] I. Jusufi, M. Milrad, and X. Legaspi, "Interactive exploration of student generated content presented in blogs," in *EuroVis Posters*, 2016, pp. 53–55.

[176] I.-H. Kang and G. Kim, "Query type classification for web document retrieval," in *ACM SIGIR*, 2003, pp. 64–71.

[177] K. Kaugars, "Integrated multi scale text retrieval visualization," in *ACM SIGCHI*, 1998, pp. 307–308.

[178] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, "Visual analytics: Definition, process, and challenges," in *Information visualization*, 2008, pp. 154–175.

[179] D. A. Keim, F. Mansmann, C. Panse, J. Schneidewind, and M. Sips, "Mail explorer—spatial and temporal exploration of electronic mail," in *EuroVis*, 2005, pp. 247–254.

[180] D. A. Keim and D. Oelke, "Literature fingerprinting: A new method for visual literary analysis," in *IEEE VAST*, 2007, pp. 115–122.

[181] M. Kim, K. Kang, D. Park, J. Choo, and N. Elmqvist, "TopicLens: Efficient multi-level visual topic exploration of large-scale document collections," *IEEE TVCG*, vol. 23, no. 1, pp. 151–160, 2017.

[182] Y. Kim, J. Han, and C. Yuan, "TOPTRAC: Topical trajectory pattern mining," in *ACM SIGKDD*, 2015, pp. 587–596.

[183] Y.-H. Kim, S.-Y. Hahn, and B.-T. Zhang, "Text filtering by boosting naive bayes classifiers," in *ACM SIGIR*, 2000, pp. 168–175.

[184] S. Koch, H. Bosch, M. Giereth, and T. Ertl, "Iterative integration of visual insights during scalable patent search and analysis," *IEEE TVCG*, vol. 17, no. 5, pp. 557–569, 2011.

[185] S. Koch, M. John, M. Wörner, A. Müller, and T. Ertl, "VarifocalReader – In-depth visual analysis of large text documents," *IEEE TVCG*, vol. 20, no. 12, pp. 1723–1732, 2014.

[186] A. Kochtchi, T. von Landesberger, and C. Biemann, "Networks of names: Visual exploration and semi-automatic tagging of social networks from newspaper articles," *Computer Graphics Forum*, vol. 33, no. 3, pp. 211–220, 2014.

[187] K. Koh, B. Lee, B. Kim, and J. Seo, "ManiWordle: Providing flexible control over Wordle," *IEEE TVCG*, vol. 16, no. 6, pp. 1190–1197, 2010.

[188] M. Krstajic, E. Bertini, and D. Keim, "CloudLines: Compact display of event episodes in multiple time-series," *IEEE TVCG*, vol. 17, no. 12, pp. 2432–2439, 2011.

[189] K. Kucher and A. Kerren, "Text visualization techniques: Taxonomy, visual survey, and community insights," in *IEEE PacificVis*, 2015, pp. 117–121.

[190] B. C. Kwon, S.-H. Kim, S. Lee, J. Choo, J. Huh, and J. S. Yi, "VisOHC: Designing visual analytics for online health communities," *IEEE TVCG*, vol. 22, no. 1, pp. 71–80, 2016.

[191] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *AAAI*, 2015, pp. 2267–2273.

[192] B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering," in *ACM SIGKDD*, 1999, pp. 16–22.

[193] T. M. V. Le and H. W. Lauw, "Semantic visualization for spherical representation," in *ACM SIGKDD*, 2014, pp. 1007–1016.

[194] B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale, "SparkClouds: Visualizing trends in tag clouds," *IEEE TVCG*, vol. 16, no. 6, pp. 1182–1189, 2010.

[195] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park, "iVisClustering: An interactive visual document clustering via topic modeling," *Computer Graphics Forum*, vol. 31, no. 3, pp. 1155–1164, 2012.

[196] J. Lee, C. Heaukulani, Z. Ghahramani, L. F. James, and S. Choi, "Bayesian inference on random simple graphs with power law degree distributions," in *ICML*, 2017, pp. 2004–2013.

[197] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *ACM SIGKDD*, 2009, pp. 497–506.

[198] A. Leuski and J. Allan, "Lighthouse: Showing the way to relevant information," in *IEEE InfoVis*, 2000, pp. 125–129.

[199] M. D. Lieberman and H. Samet, "Adaptive context features for toponym resolution in streaming news," in *ACM SIGIR*, 2012, pp. 731–740.

[200] X. Lin, "Visualization for the document space," in *IEEE VIS*, 1992, pp. 274–281.

[201] C. Y. Liu, M. S. Chen, and C. Y. Tseng, "IncreSTS: Towards real-time incremental short text summarization on comment streams from social network services," *IEEE TKDE*, vol. 27, no. 11, pp. 2986–3000, 2015.

[202] M. Liu, S. Liu, X. Zhu, Q. Liao, F. Wei, and S. Pan, "An uncertainty-aware approach for exploratory microblog retrieval," *IEEE TVCG*, vol. 22, no. 1, pp. 250–259, 2016.

[203] M. Liu, L. Jiang, J. Liu, X. Wang, J. Zhu, and S. Liu, "Improving learning-from-crowds through expert validation," in *IJCAI*, 2017, pp. 2329–2336.

[204] S. Liu, W. Zhu, N. Xu, F. Li, X.-q. Cheng, Y. Liu, and Y. Wang, "Co-training and visualizing sentiment evolvement for tweet events," in *WWW*, 2013, pp. 105–106.

[205] S. Liu, Y. Chen, H. Wei, J. Yang, K. Zhou, and S. M. Drucker, "Exploring topical lead-lag across corpora," *IEEE TKDE*, vol. 27, no. 1, pp. 115–129, 2015.

[206] S. Liu, J. Xiao, J. Liu, X. Wang, J. Wu, and J. Zhu, "Visual diagnosis of tree boosting methods," *IEEE TVCG*, vol. 24, no. 1, pp. 163–173, 2018.

[207] S. Liu, J. Yin, X. Wang, W. Cui, K. Cao, and J. Pei, "Online visual analytics of text streams," *IEEE TVCG*, vol. 22, no. 11, pp. 2451–2466, 2016.

[208] X. Liu, A. Xu, L. Gou, H. Liu, R. Akkiraju, and H.-W. Shen, "SocialBrands: Visual analysis of public perceptions of brands on social media," in *IEEE VAST*, 2016, pp. 71–80.

[209] Y. Liu, A. Niculescu-Mizil, and W. Gryc, "Topic-link LDA: Joint models of topic and author community," in *ICML*, 2009, pp. 665–672.

[210] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text classification using string kernels," *JMLR*, vol. 2, no. Feb, pp. 419–444, 2002.

[211] Y. Lu, R. Kr uger, D. Thom, F. Wang, S. Koch, T. Ertl, and R. Maciejewski, "Integrating predictive analytics and social media," in *IEEE VAST*, 2014, pp. 193–202.

[212] Y. Lu, M. Steptoe, S. Burke, H. Wang, J.-Y. Tsai, H. Davulcu, D. Montgomery, S. R. Corman, and R. Maciejewski, "Exploring evolving

media discourse through event cueing," *IEEE TVCG*, vol. 22, no. 1, pp. 220–229, 2016.

[213] Z. Lu, L. Wang, and J.-R. Wen, "Direct semantic analysis for social image classification," in *AAAI*, 2014, pp. 1258–1264.

[214] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. A. Keim, "EventRiver: Visually exploring text collections with temporal references," *IEEE TVCG*, vol. 18, no. 1, pp. 93–105, 2012.

[215] M. Luo, F. Nie, X. Chang, Y. Yang, A. G. Hauptmann, and Q. Zheng, "Probabilistic non-negative matrix factorization and its robust extensions for topic modeling," in *AAAI*, 2017, pp. 2308–2314.

[216] S. J. Luo, L. T. Huang, B. Y. Chen, and H.-W. Shen, "EmailMap: Visualizing event evolution and contact interaction within email archives," in *IEEE PacificVis*, 2014, pp. 320–324.

[217] Y. Lv, T. Moon, P. Kolari, Z. Zheng, X. Wang, and Y. Chang, "Learning to model relatedness for news recommendation," in *WWW*, 2011, pp. 57–66.

[218] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," in *IJCAI*, 2016, pp. 3818–3824.

[219] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford, "SensePlace2: GeoTwitter analytics support for situational awareness," in *IEEE VAST*, 2011, pp. 181–190.

[220] K. Madhavan, N. Elmqvist, M. Vorvoreanu, X. Chen, Y. Wong, H. Xian, Z. Dong, and A. Johri, "DIA2: Web-based cyberinfrastructure for visual analysis of funding portfolios," *IEEE TVCG*, vol. 20, no. 12, pp. 1823–1832, 2014.

[221] T. Maekawa, T. Hara, and S. Nishio, "Image classification for mobile web browsing," in *WWW*, 2006, pp. 43–52.

[222] J. Mahmud, G. Fei, A. Xu, A. Pal, and M. Zhou, "Predicting attitude and actions of Twitter users," in *IUI*, 2016, pp. 2–6.

[223] Y. Mao, J. Dillon, and G. Lebanon, "Sequential document visualization," *IEEE TVCG*, vol. 13, no. 6, pp. 1208–1215, 2007.

[224] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller, "Twitinfo: Aggregating and visualizing microblogs for event exploration," in *ACM SIGCHI*, 2011, pp. 227–236.

[225] A. Martinez and W. Martinez, "At the interface of computational linguistics and statistics," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 7, no. 4, pp. 258–274, 2015.

[226] D. Mashima, S. Kobourov, and Y. Hu, "Visualizing dynamic data with maps," *IEEE TVCG*, vol. 18, no. 9, pp. 1424–1437, 2012.

[227] N. McCurdy, J. Lein, K. Coles, and M. Meyer, "Poemage: Visualizing the sonic topology of a poem," *IEEE TVCG*, vol. 22, no. 1, pp. 439–448, 2016.

[228] Q. Mei, D. Cai, D. Zhang, and C. Zhai, "Topic modeling with network regularization," in *WWW*, 2008, pp. 101–110.

[229] Q. Mei and C. Zhai, "A mixture model for contextual text mining," in *ACM SIGKDD*, 2006, pp. 649–655.

[230] N. E. Miller, P. C. Wong, M. Brewster, and H. Foote, "TOPIC ISLANDS^{TM}—a wavelet-based text visualization system," in *IEEE VIS*, 1998, pp. 189–196.

[231] A. Mishra, K. Dey, and P. Bhattacharyya, "Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network," in *ACL*, 2017, pp. 377–387.

[232] F. Morstatter, S. Kumar, H. Liu, and R. Maciejewski, "Understanding Twitter data with TweetXplorer," in *ACM SIGKDD*, 2013, pp. 1482–1485.

[233] S. Mukherjea, K. Hirata, and Y. Hara, "Visualizing the results of multimedia Web search engines," in *IEEE InfoVis*, 1996, pp. 64–65, 122.

[234] S. Muthiah, B. Huang, J. Arredondo, D. Mares, L. Getoor, G. Katz, and N. Ramakrishnan, "Planned protest modeling in news and social media," in *AAAI*, 2015, pp. 3920–3927.

[235] M. Nacenta, U. Hinrichs, and S. Carpendale, "FatFonts: Combining the symbolic and visual aspects of numbers," in *AVI*, 2012, pp. 407–414.

[236] M. Nafari and C. Weaver, "Augmenting visualization with natural language translation of interaction: A usability study," *Computer Graphics Forum*, vol. 32, no. 3, pp. 391–400, 2013.

[237] T. N. Nguyen and J. Zhang, "A novel visualization model for web search results," *IEEE TVCG*, vol. 12, no. 5, pp. 981–988, 2006.

[238] A. Nocaj and U. Brandes, "Organizing search results with a reference map," *IEEE TVCG*, vol. 18, no. 12, pp. 2546–2555, 2012.

[239] D. Oelke, D. Kokkinakis, and D. A. Keim, "Fingerprint Matrices: Uncovering the dynamics of social networks in prose literature," *Computer Graphics Forum*, vol. 32, no. 3, pp. 371–380, 2013.

[240] D. Oelke, D. Spretke, A. Stoffel, and D. A. Keim, "Visual readability analysis: How to make your writings easier to read," in *IEEE VAST*, 2010, pp. 123–130.

[241] D. Oelke, H. Strobelt, C. Rohrdantz, I. Gurevych, and O. Deussen, "Comparative exploration of document collections: A visual analytics approach," *Computer Graphics Forum*, vol. 33, no. 3, pp. 201–210, 2014.

[242] D. Oelke, P. Bak, D. A. Keim, M. Last, and G. Danon, "Visual evaluation of text features for document summarization and analysis," in *IEEE VAST*, 2008, pp. 75–82.

[243] D. Oelke and I. Gurevych, "A study on human-generated tag structures to inform tag cloud layout," in *AVI*, 2014, pp. 297–304.

[244] D. Oelke, M. Hao, C. Rohrdantz, D. A. Keim, U. Dayal, L.-E. Haug, and H. Janetzko, "Visual opinion analysis of customer feedback data," in *IEEE VAST*, 2009, pp. 187–194.

[245] P. Oesterling, G. Scheuermann, S. Teresniak, G. Heyer, S. Koch, T. Ertl, and G. H. Weber, "Two-stage framework for a topology-based projection and visualization of classified document collections," in *IEEE VAST*, 2010, pp. 91–98.

[246] P. Ogilvie and J. Callan, "Combining document representations for known-item search," in *ACM SIGIR*, 2003, pp. 143–150.

[247] S. Pan, M. X. Zhou, Y. Song, W. Qian, F. Wang, and S. Liu, "Optimizing temporal topic segmentation for intelligent text visualization," in *IUI*, 2013, pp. 339–350.

[248] H. Park and J. Choi, "V-model: A new innovative model to chronologically visualize narrative clinical texts," in *ACM SIGCHI*, 2012, pp. 453–462.

[249] S. Park, S. Lee, and J. Song, "Aspect-level news browsing: Understanding news events from multiple viewpoints," in *IUI*, 2010, pp. 41–50.

[250] R. Patel and W. Furr, "ReadN'Karaoke: Visualizing prosody in children's books for expressive oral reading," in *ACM SIGCHI*, 2011, pp. 3203–3206.

[251] F. V. Paulovich and R. Minghim, "HiPP: A novel hierarchical point placement strategy and its application to the exploration of document collections," *IEEE TVCG*, vol. 14, no. 6, pp. 1229–1236, 2008.

[252] F. V. Paulovich, L. G. Nonato, R. Minghim, and H. Levkowitz, "Least Square Projection: A fast high-precision multidimensional projection technique and its application to document mapping," *IEEE TVCG*, vol. 14, no. 3, pp. 564–575, 2008.

[253] F. V. Paulovich, F. M. B. Toledo, G. P. Telles, R. Minghim, and L. G. Nonato, "Semantic wordification of document collections," *Computer Graphics Forum*, vol. 31, no. 3, pp. 1145–1153, 2012.

[254] A. Perer and M. A. Smith, "Contrasting portraits of email practices: Visual approaches to reflection and analysis," in *AVI*, 2006, pp. 389–395.

[255] A. J. Perotte, F. Wood, N. Elhadad, and N. Bartlett, "Hierarchically supervised latent Dirichlet allocation," in *NIPS*, 2011, pp. 2609–2617.

[256] D. Pinto, A. McCallum, X. Wei, and W. B. Croft, "Table extraction using conditional random fields," in *ACM SIGIR*, 2003, pp. 235–242.

[257] P. Pirolli and S. K. Card, "Information foraging models of browsers for very large document spaces," in *AVI*, 1998, pp. 83–93.

[258] P. Pirolli, E. Wollny, and B. Suh, "So you know you're getting the best possible information: A tool that increases Wikipedia credibility," in *ACM SIGCHI*, 2009, pp. 1505–1508.

[259] J. Prager, E. Brown, A. Coden, and D. Radev, "Question-answering by predictive annotation," in *ACM SIGIR*, 2000, pp. 184–191.

[260] J. Proskurnia, M.-A. Cartright, L. Garcia-Pueyo, I. Krka, J. B. Wendt, T. Kaufmann, and B. Miklos, "Template induction over unstructured email corpora," in *WWW*, 2017, pp. 1521–1530.

[261] H. Raghavan and J. Allan, "An interactive algorithm for asking and incorporating feature feedback into support vector machines," in *ACM SIGIR*, 2007, pp. 79–86.

[262] M. M. Rahman and H. Wang, "Hidden topic sentiment model," in *WWW*, 2016, pp. 155–165.

[263] L. Ramaswamy, A. Iyengar, L. Liu, and F. Douglis, "Automatic detection of fragments in dynamically generated web pages," in *WWW*, 2004, pp. 443–454.

[264] M. D. Reid and R. C. Williamson, "Information, divergence and risk for binary experiments," *JMLR*, vol. 12, no. Mar, pp. 731–817, 2011.

[265] D. Ren, X. Zhang, Z. Wang, J. Li, and X. Yuan, "WeiboEvents: A crowd sourcing Weibo visual analytic system," in *IEEE PacificVis*, 2014, pp. 330–334.

[266] Z. Ren, M.-H. Peetz, S. Liang, W. van Dolen, and M. de Rijke, "Hierarchical multi-label classification of social text streams," in *ACM SIGIR*, 2014, pp. 213–222.

[267] R. L. Ribler and M. Abrams, "Using visualization to detect plagiarism in computer science classes," in *IEEE InfoVis*, 2000, pp. 173–178.

[268] N. H. Riche, B. Lee, and F. Chevalier, "iChase: Supporting exploration and awareness of editing activities on Wikipedia," in *AVI*, 2010, pp. 59–66.

[269] P. Riehmann, M. Potthast, B. Stein, and B. Froehlich, "Visual assessment of alleged plagiarism cases," *Computer Graphics Forum*, vol. 34, no. 3, pp. 61–70, 2015.

[270] P. Riehmann, H. Gruendl, B. Froehlich, M. Potthast, M. Trenkmann, and B. Stein, "The NETSPEAK WORDGRAPH: Visualizing keywords in context," in *IEEE PacificVis*, 2011, pp. 123–130.

[271] P. Riehmann, H. Gruendl, M. Potthast, M. Trenkmann, B. Stein, and B. Froehlich, "WORDGRAPH: Keyword-in-context visualization for NETSPEAK's wildcard search," *IEEE TVCG*, vol. 18, no. 9, pp. 1411–1423, 2012.

[272] C. Rohrdantz, M. Hund, T. Mayer, and D. A. Keim, "The world's languages explorer: Visual analysis of language features in genealogical and areal contexts," *Computer Graphics Forum*, vol. 31, no. 3, pp. 935–944, 2012.

[273] R. M. Rohrer, D. S. Ebert, and J. L. Sibert, "The shape of Shakespeare: visualizing text using implicit surfaces," in *IEEE InfoVis*, 1998, pp. 121–129.

[274] S. Rose, S. Butner, W. Cowley, M. Gregory, and J. Walker, "Describing story evolution from dynamic information streams," in *IEEE VAST*, 2009, pp. 99–106.

[275] D. A. Rushall and M. R. Ilgen, "DEPICT: Documents evaluated as pictures. Visualizing information using context vectors and self-organizing maps," in *IEEE InfoVis*, 1996, pp. 100–107.

[276] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim, "Visual interaction with dimensionality reduction: A structured literature analysis," *IEEE TVCG*, vol. 23, no. 1, pp. 241–250, 2017.

[277] U. Sapkota, T. Solorio, M. Montes-y Gómez, and S. Bethard, "Domain adaptation for authorship attribution: Improved structural correspondence learning," in *ACL*, 2016, pp. 2226–2235.

[278] S. Schlechtweg, P. Schulze-Wollgast, and H. Schumann, "Interactive treemaps with detail on demand to support information search in documents," in *EuroVis*, 2004, pp. 121–128.

[279] E. Schubert, M. Weiler, and H.-P. Kriegel, "SigniTrend: scalable detection of emerging topics in textual streams by hashed significance thresholds," in *ACM SIGKDD*, 2014, pp. 871–880.

[280] H. Schütze, "Word space," in *NIPS*, 1993, pp. 895–902.

[281] H. Schütze and C. Silverstein, "Projections for efficient document clustering," in *ACM SIGIR*, 1997, pp. 74–81.

[282] E. Segel and J. Heer, "Narrative visualization: Telling stories with data," *IEEE TVCG*, vol. 16, no. 6, pp. 1139–1148, 2010.

[283] P. Sen, "Collective context-aware topic models for entity disambiguation," in *WWW*, 2012, pp. 729–738.

[284] C. Shah and W. B. Croft, "Evaluating high accuracy retrieval techniques," in *ACM SIGIR*, 2004, pp. 2–9.

[285] D. Shahaf, C. Guestrin, and E. Horvitz, "Metro maps of science," in *ACM SIGKDD*, 2012, pp. 1122–1130.

[286] D. Shahaf, J. Yang, C. Suen, J. Jacobs, H. Wang, and J. Leskovec, "Information cartography: Creating zoomable, large-scale maps of information," in *ACM SIGKDD*, 2013, pp. 1097–1105.

[287] D. Shen, Z. Chen, Q. Yang, H.-J. Zeng, B. Zhang, Y. Lu, and W.-Y. Ma, "Web-page classification through summarization," in *ACM SIGIR*, 2004, pp. 242–249.

[288] Q. Shen, T. Wu, H. Yang, Y. Wu, H. Qu, and W. Cui, "NameClarifier: A visual analytics system for author name disambiguation," *IEEE TVCG*, vol. 23, no. 1, pp. 141–150, 2017.

[289] Y. Shen, W. Rong, N. Jiang, B. Peng, J. Tang, and Z. Xiong, "Word embedding based correlation model for question/answer matching." in *AAAI*, 2017, pp. 3511–3517.

[290] L. Shi, F. Wei, S. Liu, L. Tan, X. Lian, and M. X. Zhou, "Understanding text corpora with multiple facets," in *IEEE VAST*, 2010, pp. 99–106.

[291] M. Shokouhi, L. Azzopardi, and P. Thomas, "Effective query expansion for federated search," in *ACM SIGIR*, 2009, pp. 427–434.

[292] L. Si and J. Callan, "Using sampled data and regression to merge search engine results," in *ACM SIGIR*, 2002, pp. 19–26.

[293] M. A. Smith and A. T. Fiore, "Visualization components for persistent conversations," in *ACM SIGCHI*, 2001, pp. 136–143.

[294] J.-W. Son, J. Jeon, A. Lee, and S.-J. Kim, "Spectral clustering with brainstorming process for multi-view data," in *AAAI*, 2017, pp. 2548–2554.

[295] Y. Song, S. Pan, S. Liu, F. Wei, M. X. Zhou, and W. Qian, "Constrained coclustering for textual documents," in *AAAI*, 2010, pp. 581–586.

[296] D. Spina, J. Gonzalo, and E. Amigó, "Learning similarity functions for topic detection in online reputation monitoring," in *ACM SIGIR*, 2014, pp. 527–536.

[297] A. Spoerri, "InfoCrystal: A visual tool for information retrieval & management," in *IEEE VIS*, 1993, pp. 150–157.

[298] J. Stasko, C. Görg, Z. Liu, and K. Singhal, "Jigsaw: Supporting investigative analysis through interactive visualization," in *IEEE VAST*, 2007, pp. 131–138.

[299] F. Stoffel, W. Jentner, M. Behrisch, J. Fuchs, and D. Keim, "Interactive ambiguity resolution of named entities in fictional literature," *Computer Graphics Forum*, vol. 36, no. 3, pp. 189–200, 2017.

[300] H. Strobelt, D. Oelke, C. Rohrdantz, A. Stoffel, D. A. Keim, and O. Deussen, "Document Cards: A top trumps visualization for documents," *IEEE TVCG*, vol. 15, no. 6, pp. 1145–1152, 2009.

[301] H. Strobelt, M. Spicker, A. Stoffel, D. Keim, and O. Deussen, "Rolled-out Wordles: A heuristic method for overlap removal of 2d data representatives," *Computer Graphics Forum*, vol. 31, no. 3, pp. 1135–1144, 2012.

[302] G. Sun, Y. Wu, S. Liu, T. Q. Peng, J. J.H.Zhu, and R. Liang, "EvoRiver: Visual analysis of topic coopetition on social media," *IEEE TVCG*, vol. 20, no. 12, pp. 1753–1762, 2014.

[303] D. A. Szafir, D. Stuffer, Y. Sohail, and M. Gleicher, "TextDNA: Visualizing word usage with configurable colorfields," *Computer Graphics Forum*, vol. 35, no. 3, pp. 421–430, 2016.

[304] G. K. Tam, V. Kothari, and M. Chen, "An analysis of machine-and human-analytics in classification," *IEEE TVCG*, vol. 23, no. 1, pp. 71–80, 2017.

[305] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *ACM SIGKDD*, 2009, pp. 807–816.

[306] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[307] D. Thom, R. Krüger, and T. Ertl, "Can Twitter save lives? A broad-scale study on visual social media analytics for public safety," *IEEE TVCG*, vol. 22, no. 7, pp. 1816–1829, 2016.

[308] D. Thom, H. Bosch, S. Koch, M. Wörner, and T. Ertl, "Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages," in *IEEE PacificVis*, 2012, pp. 41–48.

[309] N. UzZaman, J. P. Bigham, and J. F. Allen, "Multimodal summarization of complex sentences," in *IUI*, 2011, pp. 43–52.

[310] F. van Ham, M. Wattenberg, and F. B. Viégas, "Mapping text with Phrase Nets," *IEEE TVCG*, vol. 15, no. 6, pp. 1169–1176, 2009.

[311] K. Verbert, D. Parra, P. Brusilovsky, and E. Duval, "Visualizing recommendations to support exploration, transparency and controllability," in *IUI*, 2013, pp. 351–362.

[312] F. B. Viégas, M. Wattenberg, and J. Feinberg, "Participatory visualization with Wordle," *IEEE TVCG*, vol. 15, no. 6, pp. 1137–1144, 2009.

[313] F. Viégas, M. Wattenberg, J. Hebert, G. Borggaard, A. Cichowlas, J. Feinberg, J. Orwant, and C. Wren, "Google+ Ripples: A native visualization of information flow," in *WWW*, 2013, pp. 1389–1398.

[314] F. B. Viégas, S. Golder, and J. Donath, "Visualizing email content: Portraying relationships from conversational histories," in *ACM SIGCHI*, 2006, pp. 979–988.

[315] R. Vuillemot, T. Clement, C. Plaisant, and A. Kumar, "What's being said near Martha? Exploring name entities in literary text collections," in *IEEE VAST*, 2009, pp. 107–114.

[316] S. Wan and C. Paris, "Improving government services with social media feedback," in *IUI*, 2014, pp. 27–36.

[317] C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *IJCAI*, 2011, pp. 1541–1546.

[318] C. Wang, Y. Wang, P.-S. Huang, A. Mohamed, D. Zhou, and L. Deng, "Sequence modeling via segmentations," in *ICML*, 2017, pp. 3674–3683.

[319] D. Wang, T. Li, S. Zhu, and C. Ding, "Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization," in *ACM SIGIR*, 2008, pp. 307–314.

[320] F. Y. Wang, A. Sallaberry, K. Klein, M. Takatsuka, and M. Roche, "SentiCompass: Interactive visualization for exploring and comparing the sentiments of time-varying Twitter data," in *IEEE PacificVis*, 2015, pp. 129–133.

[321] H. Wang, H. Huang, F. Nie, and C. Ding, "Cross-language web page classification via dual knowledge transfer using Nonnegative Matrix Tri-factorization," in *ACM SIGIR*, 2011, pp. 933–942.

[322] S. Wang, Z. Chen, B. Liu, and S. Emery, "Identifying search keywords for finding relevant social media posts," in *AAAI*, 2016, pp. 3052–3058.

[323] X. Wang, W. Dou, Z. Ma, J. Villalobos, Y. Chen, T. Kraft, and W. Ribarsky, "I-SI : Scalable architecture for analyzing latent topical-

level information from social media data," *Computer Graphics Forum*, vol. 31, no. 3, pp. 1275–1284, 2012.
[324] X. Wang, B. Janssen, and E. Bier, "Finding business information by visualizing enterprise document activity," in *AVI*, 2010, pp. 41–48.
[325] X. Wang, S. Liu, Y. Chen, T. Q. Peng, J. Su, J. Yang, and B. Guo, "How ideas flow across multiple social groups," in *IEEE VAST*, 2016, pp. 51–60.
[326] X. Wang, S. Liu, J. Liu, J. Chen, J. Zhu, and B. Guo, "TopicPanorama: A full picture of relevant topics," *IEEE TVCG*, vol. 22, no. 12, pp. 2508–2521, 2016.
[327] Y. Wang, H. Huang, C. Feng, Q. Zhou, J. Gu, and X. Gao, "Cse: Conceptual sentence embeddings based on attention model," in *ACL*, 2016, pp. 505–515.
[328] Y. Wang, G. Williams, E. Theodorou, and L. Song, "Variational policy for guiding point processes," in *ICML*, 2017, pp. 3684–3693.
[329] M. Wattenberg and F. B. Viégas, "The Word Tree, an interactive visual concordance," *IEEE TVCG*, vol. 14, no. 6, pp. 1221–1228, 2008.
[330] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang, "TIARA: A visual exploratory text analytic system," in *ACM SIGKDD*, 2010, pp. 153–162.
[331] J. Wieting and K. Gimpel, "Revisiting recurrent networks for paraphrastic sentence embeddings," in *ACL*, 2017, pp. 2078–2088.
[332] P. C. Wong, H. Foote, D. Adams, W. Cowley, and J. Thomas, "Dynamic visualization of transient data streams," in *IEEE InfoVIs*, 2003, pp. 97–104.
[333] K. Wongsuphasawat and D. Gotz, "Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization," *IEEE TVCG*, vol. 18, no. 12, pp. 2659–2668, 2012.
[334] J. Wood, J. Dykes, A. Slingsby, and K. Clarke, "Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geo-visualization mashup," *IEEE TVCG*, vol. 13, no. 6, pp. 1176–1183, 2007.
[335] F. Wu and Y. Huang, "Sentiment domain adaptation with multiple sources," in *ACL*, 2016, pp. 301–310.
[336] Y. Wu, S. Liu, K. Yan, and M. Liu, "OpinionFlow: Visual analysis of opinion diffusion on social media," *IEEE TVCG*, vol. 20, no. 12, pp. 1763–1772, 2014.
[337] Y. Wu, T. Provan, F. Wei, S. Liu, and K.-L. Ma, "Semantic-preserving word clouds by seam carving," *Computer Graphics Forum*, vol. 30, no. 3, pp. 741–750, 2011.
[338] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu, "OpinionSeer: Interactive visualization of hotel customer feedback," *IEEE TVCG*, vol. 16, no. 6, pp. 1109–1118, 2010.
[339] C. Xiao, P. Zhang, W. A. Chaovalitwongse, J. Hu, and F. Wang, "Adverse drug reaction prediction with symbolic latent Dirichlet allocation," in *AAAI*, 2017, pp. 1590–1596.
[340] M. Xiao and Y. Guo, "A novel two-step method for cross language representation learning," in *NIPS*, 2013, pp. 1259–1267.
[341] W. Xie, Y. Peng, and J. Xiao, "Cross-view feature learning for scalable social image analysis," in *AAAI*, 2014, pp. 201–207.
[342] J. Xu, Y. Tao, H. Lin, R. Zhu, and Y. Yan, "Exploring controversy via sentiment divergences of aspects in reviews," in *IEEE PacificVis*, 2017, pp. 240–249.
[343] J. Xu and J. Callan, "Effective retrieval with distributed collections," in *ACM SIGIR*, 1998, pp. 112–120.
[344] P. Xu, Y. Wu, E. Wei, T. Q. Peng, S. Liu, J. J. H. Zhu, and H. Qu, "Visual analysis of topic competition on social media," *IEEE TVCG*, vol. 19, no. 12, pp. 2012–2021, 2013.
[345] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *ACM SIGIR*, 2003, pp. 267–273.
[346] J. Yabe, S. Takahashi, and E. Shibayama, "Automatic animation of discussions in USENET," in *AVI*, 2000, pp. 84–91.
[347] S.-H. Yang, A. Kolcz, A. Schlaikjer, and P. Gupta, "Large-scale high-precision topic modeling on Twitter," in *IEEE SIGKDD*, 2014, pp. 1907–1916.
[348] S. Yu, D. Cai, J.-R. Wen, and W.-Y. Ma, "Improving pseudo-relevance feedback in web information retrieval using web page segmentation," in *WWW*, 2003, pp. 11–18.
[349] M. Yurochkin and X. Nguyen, "Geometric Dirichlet means algorithm for topic inference," in *NIPS*, 2016, pp. 2505–2513.
[350] H. Zha, X. He, C. Ding, M. Gu, and H. D. Simon, "Spectral relaxation for k-means clustering," in *NIPS*, 2002, pp. 1057–1064.
[351] C. Zhai and J. Lafferty, "Two-stage language models for information retrieval," in *ACM SIGIR*, 2002, pp. 49–56.
[352] C. Zhang, L. Shou, K. Chen, and G. Chen, "See-to-retrieve: Efficient processing of spatio-visual keyword queries," in *ACM SIGIR*, 2012, pp. 681–690.
[353] D. Zhang and W. S. Lee, "Web taxonomy integration using support vector machines," in *WWW*, 2004, pp. 472–481.
[354] J. Zhang, Y. Song, C. Zhang, and S. Liu, "Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora," in *ACM SIGKDD*, 2010, pp. 1079–1088.
[355] J. Zhang, B. Ahlbrand, A. Malik, J. Chae, Z. Min, S. Ko, and D. S. Ebert, "A visual analytics framework for microblog data analysis at multiple scales of aggregation," *Computer Graphics Forum*, vol. 35, no. 3, pp. 441–450, 2016.
[356] N. L. Zhang and L. K. Poon, "Latent tree analysis." in *AAAI*, 2017, pp. 4891–4898.
[357] Y. Zhang, J. Callan, and T. Minka, "Novelty and redundancy detection in adaptive filtering," in *ACM SIGIR*. ACM, 2002, pp. 81–88.
[358] J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin, and C. Collins, "#FluxFlow: Visual analysis of anomalous information spreading on social media," *IEEE TVCG*, vol. 20, no. 12, pp. 1773–1782, 2014.
[359] J. Zhao, F. Chevalier, C. Collins, and R. Balakrishnan, "Facilitating discourse analysis with interactive visualization," *IEEE TVCG*, vol. 18, no. 12, pp. 2639–2648, 2012.
[360] J. Zhao, C. Collins, F. Chevalier, and R. Balakrishnan, "Interactive exploration of implicit and explicit relations in faceted datasets," *IEEE TVCG*, vol. 19, no. 12, pp. 2080–2089, 2013.
[361] J. Zhao, L. Gou, F. Wang, and M. Zhou, "PEARL: An interactive visual analytic tool for understanding personal emotion style derived from social media," in *IEEE VAST*, 2014, pp. 203–212.
[362] J. Zhao, L. Dong, J. Wu, and K. Xu, "MoodLens: An emoticon-based sentiment analysis system for Chinese tweets," in *ACM SIGKDD*, 2012, pp. 1528–1531.
[363] G. Zheng, J. Guo, L. Yang, S. Xu, S. Bao, Z. Su, D. Han, and Y. Yu, "Mining topics on participations for community discovery," in *ACM SIGIR*, 2011, pp. 445–454.
[364] G. Zhou, Z. Xie, J. X. Huang, and T. He, "Bi-transferring deep neural networks for domain adaptation," in *ACL*, 2016, pp. 322–332.
[365] X. Zhou, X. Wan, and J. Xiao, "Cross-lingual sentiment classification with bilingual document representation learning," in *ACL*, 2016, pp. 1403–1412.
[366] "New Power BI custom visuals enable browsing and analyzing collections of text," https://powerbi.microsoft.com/en-us/blog/new-power-bi-custom-visuals-for-browsing-and-analyzing-collections-of-text/, accessed: 31-Dec-2017.

Shixia Liu is an associate professor at Tsinghua University. Her research interests include visual text analytics, visual social analytics, interactive machine learning, and text mining. She worked as a research staff member at IBM China Research Lab and a lead researcher at Microsoft Research Asia. She received a B.S. and M.S. from Harbin Institute of Technology, a Ph.D. from Tsinghua University. She is an associate editor of IEEE TVCG and Information Visualization.

Xiting Wang is now an associate researcher at Microsoft Research Asia. Her research interests include text mining, visual text analytics, and explainable recommendation. She received a BS degree and a Ph.D. degree from Tsinghua University.

Christopher Collins received the PhD degree from University of Toronto in 2010. He is currently the Canada Research Chair in Linguistic Information Visualization and Associate Professor at the University of Ontario Institute of Technology. His research focus combines information visualization and human-computer interaction with natural language processing. He is a member of the IEEE, a past member of the executive of the IEEE VGTC and has served several roles on the IEEE VIS Conference Organizing Committee.

Wenwen Dou Wenwen Dou is an assistant professor at the University of North Carolina at Charlotte. Her research interests include Visual Analytics, Text Mining, and Human Computer Interaction. Dou has worked with various analytics domains in reducing information overload and providing interactive visual means to analyzing unstructured information. She has experience in turning cutting-edge research into technologies that have broad societal impacts.

Fangxin Ouyang is now a master student at Tsinghua University. Her research interest is visual text analytics. She received a BS degree in the School of Software, Tsinghua University.

Mennatallah El-Assady is a research associate in the group for Data Analysis and Visualization at the University of Konstanz (Germany) and in the Visualization for Information Analysis lab at the University of Ontario Institute of Technology (Canada). She received her M.Sc. degree in Information Engineering from the University of Konstanz in 2015. Her research interests include visual text analytics, user-steerable topic modeling, and discourse/conversational text analysis.

Liu Jiang is now a Ph.D. student at Tsinghua University. His research interest includes learning from crowds and interactive machine learning. He received a BS degree in Electronics Engineering, University of Science and Technology of China.

Daniel A. Keim is professor and head of the Information Visualization and Data Analysis Research Group in the Computer Science Department at the University of Konstanz, Germany. He has been actively involved in data analysis and information visualization research for more than 25 years and developed novel visual analysis techniques for large data sets, including textual data. Dr. Keim got his Ph.D. degree in Computer Science from the University of Munich, Germany. Before joining the University of Konstanz, Dr. Keim was associate professor at the University of Halle, Germany and Technology Consultant at AT&T Shannon Research Labs, NJ, USA.