EG UK Computer Graphics & Visual Computing (2018)
G. Tam and F. Vidal (Editors)

SoS TextVis: A Survey of Surveys on Text Visualization

Mohammad Alharbi and Robert S. Laramee[1]

[1] Department of Computer Science, Swansea University, United Kingdom

Abstract

Text visualization is a rapidly growing sub-field of information visualization and visual analytics. There are many approaches and techniques introduced every year to address a wide range of tasks and enable researchers from different disciplines to obtain leading-edge knowledge from digitized collections. This can be challenging particularly when the data is massive. Additionally, the sources of digital text have spread substantially in the last decades in various forms, such as web pages, blogs, twitter, email, electronic publications, and books. In response to the explosion of text visualization research literature, the first survey article was published in 2010. Furthermore, there are a growing number of surveys that review existing techniques and classify them based on text research methodology. In this work, we aim to present the first Survey of Surveys (SoS) that review all of the survey and state-of-the-art papers on text visualization techniques and provide an SoS classification. We study and compare the surveys, and categorize them into 5 groups: (1) document-centered, (2) user task analysis, (3) cross-disciplinary, (4) multi-faceted, and (5) satellite-themed. We provide survey recommendations for researchers in the field of text visualization. The result is a very unique, valuable starting point and overview of the current state-of-the-art in text visualization research literature.

1. Introduction and Motivation

Text visualization is a rapidly growing sub-field of information visualization and visual analytics. Therefore, many approaches and techniques are introduced periodically to help users and researchers with a wide range of tasks. The volume of digital text data is multiplying due to the popular demand for digital text and the digitization projects, such as those by Reddy and StClair [RS01], Andre and Eaton [AE88], and Mendelsson *et al* [MFLO14]. Literature and historical documents are digitized for further study and analysis. This volume of digital text data makes understanding and analyzing it extremely challenging. Text documents by their nature bring many challenges such as high dimensionality, irregularity, and uncertainty inherent in natural language. Thus, many advanced techniques are needed to address such challenges.

Currently, Kucher and Kerren [KK15] review over 400 text visualization approaches in their interactive web-based tool "Text Visualization Browser" (at the time of this writing and the tools are regularly updated). However, the approaches listed in the Text Visualization Browser mainly come from the data visualization community and do not include literature from other communities – particularly from digital humanities. The number of text literature surveys has grown since the first survey was published in 2010 by Šilić and Bašić [ŠB10] as shown in Figure 1. Collectively with duplicates, the surveys review 1288 text visualization approaches.

In this review, we provide a meta-survey of the existing surveys that address the exploration, analysis, and presentation of text data. Our contributions to the field include:

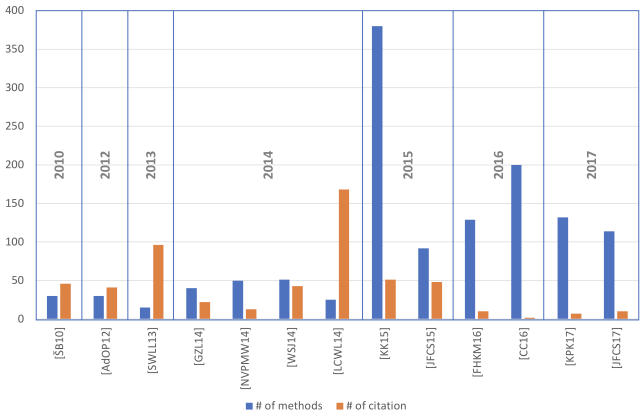

Figure 1: *The text visualization surveys from 2010 to 2017. Blue bard indicate the number of methods reviewed in each survey. Orange bars show the number of citations each survey attracts. In term of the number of surveys, 2014 dominates with 4 surveys. However, with the respect to the number of techniques, 2015 surveys review 480 methods collectively.*

- the first focus survey of surveys (SoS) in text visualization,
- a novel classification of text surveys in the reviewed literature,
- helpful survey meta-data in order to facilitate comparison of the surveys, and
- a unique, valuable starting point and comprehensive overview

DOI: 10.2312/cgvc.20181219

Conferences & Journals	Papers
The Annual EuroVis Conference / Computer Graphics Forum	3
IEEE Pacific Visualization Symposium	1
IEEE Transactions on Visualization and Computer Graphics	1
IEEE VAST Conference	0
Journal of Visual Languages & Computing	0
Information Visualization Journal	0
ACM Computing Surveys	0
Springer	4
Wiley Online Library	3
Proceedings of the Annual Conference of the Alliance of Digital Humanities Organizations	0
Literary and Linguistic Computing	0
Digital Humanities Quarterly	0
Other	1
Total	13

Table 1: *A list of literature sources searched for text visualization surveys. We mainly use IEEE Xplore [iee16], ACM Digital Library [acm16], and Google Scholar [sch16] to search for literature*

for both newcomers and experienced researchers in text visualization.

The rest of this paper is organized as follows: Section 1.1 describes the methodology used to collect related research papers and the scope of the literature. Section 1.3 introduces a previous survey of surveys and how our work differs from it. Section 1.4 presents our classification of the literature. Section 2 discusses and compares the surveys according to the classification in section 1.4. Section 3 summarizes and discusses the future challenges reported within our collection. We finish this article with conclusions and future work directions.

1.1. Literature Search Methodology

Our search methodology is a variant of the SoS by McNabb and Laramee [ML17] since they have collected many surveys in the field of information visualization and visual analytics. However, since the publication of the SoS, more recent surveys have been published in the field of text visualization and visual analytics, such as by Kucher *et al.* [KPK17] and Jänicke *et al.* [JFCS17].

In our search of the literature, we performed a manual search by looking at each journal and conference in the data visualization community and performed a keyword search e.g., 'Text Visualization Survey,' 'Text Taxonomy,' 'Text Visualization State-of-the-Art,' or 'Visual Text.' We list all the literature sources searched in Table 1. As text visualization is of interest to other communities, we searched the digital humanities (DH) digital libraries to look for surveys, however, we could not find any survey in the main DH venues (shown in Table 1).

1.2. Survey Scope

In scope: We found and collected 13 surveys to include in our text SoS. We include surveys dedicated to text analysis and visualization approaches as well as surveys that explicitly feature a text visualization category in the main literature classification, such as Sun *et al.* [SWLL13] and Liu *et al* [LCWL14].
Out of scope: We restrict our literature to surveys that include a review of text visualization approaches. We do not include surveys that review text mining techniques like summarization techniques, such as Gupta and Lehal [GL10] or text clustering algorithms like Aggarwal and Zhai [AZ12]. Survey papers that focus on text recognition, such as text detection and extraction by Jung *et al.* [JKJ04] are also out of the scope of this survey.

1.3. Related Work

McNabb and Laramee [ML17] took the first step towards presenting the landscape of survey papers in information visualization. They present eight surveys which focus on analyzing and visualizing text data. They classify the papers using an adapted information visualization pipeline by Card *et al* [CMS99]. They also identify three characteristics of classifications: the dimensions that each classification of survey adopts, the structure of the classification, and the type of mapping schema the survey incorporates. Kucher and Kerren [KK15] also review five surveys that focus on text visualization and compare the visualization taxonomies used in the reviews with their proposed taxonomy.

In our review, we aim to describe the existing surveys in more depth than McNabb and Laramee [ML17] and more breadth than Kucher and Kerren [KK15]. This text SoS includes more referenced text-focused surveys and book chapters than [ML17] or [KK15]. It is, to our knowledge, the first comprehensive survey of surveys (SoS) in text visualization.

1.4. Survey Classification

In order to compare the surveys, we classify them into five categories. We study each survey classification and categorization and, group them based on the main focus themes found in each, see Table 2. Thus, we identify the following five themes:

1. **Document-centered:** we place all surveys that derive their classification based on the underlying text source in this group.
2. **Task Analysis:** we group surveys that mainly categorize their related literature based on the task analysis.
3. **multi-faceted:** here are surveys that categorize related literature into multi-faceted classifications. In this case, the survey may propose multiple classifications for a variety of characteristics.
4. **Cross-disciplinary:** we collect surveys that survey visualization techniques to support Digital Humanities.
5. **Satellite-themed:** this group contains surveys that review existing information visualization literature. We include surveys that only include text visualization as a sub-section within their classification.

Data Source	Task Analysis	multi-faceted	Cross-disciplinary	Satellite-themed
[AdOP12] [GZL*14] [NVPMW14]	[CC16] [FHKM16]	[ŠB10] [WSJ*14] [KK15] [KPK17]	[JFCS15] [JFCS17]	[SWLL13] [LCWL14]

Table 2: *Classification of our collection of surveys. There are five categories: data source, task analysis, multi-faceted, cross-disciplinary, and satellite-themed in to which the literature is grouped.*

2. Summary and Comparison

In this section, we discuss the surveys presented in Table 2 and provide our recommendations.

2.1. Document-centered Surveys

There are three surveys in this collection by Alencar *et al.* [AdOP12], Gan *et al.* [GZL*14] and Nualart-Vilaplana *et al* [NVPMW14]. Their classifications are centered around document type. This refers to classifying the document –as the central theme –to a single document, a collection of documents, or a stream of text, etc.

Alencar *et al.* [AdOP12] review visual text analysis approaches. In their classification, there are two main categories. The first is *target input material of approaches*, either a single document (TagCrowd [Ste14] and Wordle [VWF09]) or a collection of text (Cartographic Maps [Sku02], Galaxies [Wis99], InfoSky [AKS*02] and Document Cards [SOR*09]). The second category is *the focus of the approaches*, such as showing relations (CiteSpace [LH08]), highlighting temporal changes (SparkClouds [LRKC10]) and visualizing query results (TileBars [Hea95]). They describe each approach to obtain meaningful text models, how they extract information to produce representative visual designs, the user tasks supported, the interaction techniques applied and their strengths and limitations.

Gan *et al.* [GZL*14] present an overview of the concept of document visualization, the related research, and representative methods in each category of their hierarchical document classification. They classify the literature mainly based on the data source. Figure 2 shows how the methods are classified. The review then introduces several representative methods for each category which are featured based on different aspects, such as the visualized text characteristics, representative methods, design principles satisfied based on Shneiderman [Shn96].

Nualart-Vilaplana *et al.* [NVPMW14] examine 49 approaches to visualize textual data over a 19-year period spanning 1994 to 2013 in order to provide a classification of text visualization approaches. Similar to Gan *et al.* [GZL*14], Nualart-Vilaplana *et al.* [NVPMW14] start with the data source of documents and then describe analysis tasks supported in each category. The classification comprises two main categories: individual texts and collections of texts. In each category, there are heuristic subdivisions in order to understand and describe the graphs. The subdivision of the single texts and collections categories includes:

1. The sub-divisions for individual texts:
   a. Whole or sub-sets: the visualization process includes the whole text or part of it.
   b. Sequential or non-sequential: the visual layout preserves the same word sequence as that of the original text.
   c. Discourse structure or syntactic structure: the visual design uses elements from discourse structure which refers to using actual parts of the text enabling the viewer to read through visualization or syntactic structure using intrinsic elements of the text ,such as words and phrases.
   d. Search: the imagery results from a search query.
   e. Time: the text changes over time.
2. The sub-divisions for collections of texts are:
   a. Items or Aggregations: the items of the collection used individually or there is some aggregation visualized.
   b. Pure data or landscape: the text data in the collection is accompanied by a graphical content.
   c. Search: same as above.
   d. Time: same as above.

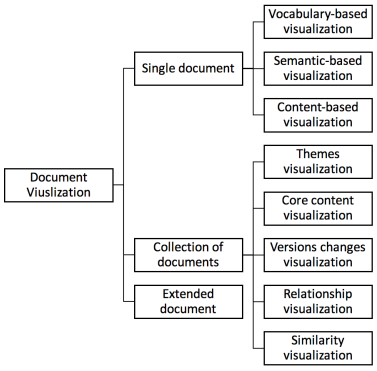

Figure 2: *Hierarical classification of document visualization methods used by Gan et al [GZL*14].*

2.2. User Task Analysis Surveys

In this category, we group surveys that mainly categorize their related literature based on user task analysis. There are two surveys in this category by Cau and Cui [CC16] and Federico *et al* [FHKM16].

Cau and Cui [CC16] present a systematic review of existing text visualization techniques. The volume of the approaches cited is over 200. The overview classifies the approaches into two main categories: (1) visualization and (2) exploration or interaction. They classify the literature in the visualization category based on the tasks of the visualization (what each is developed for), such as showing similarities, contents, and sentiments. For large document collections, the review provides the most common exploration techniques which include distortion based approaches and hierarchical document exploration approaches.

Federico *et al.* [FHKM16] survey interactive visualization approaches that support search and analysis of scientific articles and patents. They classify the visualization approaches according to two orthogonal aspects: data type and analysis tasks. There are four data types identified: text, citation, authors, and meta-data. The analysis task break down [FHKM16] adopts the typology of data analysis tasks by Andrienko and Andrienko [AA06]. The four analysis tasks include elementary lookup and comparison, elementary relation seeking, synoptic tasks, and temporal patterns. Furthermore, the review also introduces a breakdown of approaches that handle multiple data types.

2.3. Multi-faceted Surveys

In this category, there are four surveys by Šilić and Bašić [ŠB10], Wanner *et al.* [WSJ*14], Kucher and Kerren [KK15], and Kucher

Method name	Basic underlying methods	Data type	Temporal	Year	Ref.
Sammon	Sammon's mapping	C	-	1969	[27]
Lin et al.	SOM	C	-	1991	[28]
BEAD	FDP	C	-	1992	[29]
Galaxy of News	ARN	C	-	1994	[30]
SPIRE / IN-SPIRE	MDS, ALS, PCA, Clustering	C/S	-	1995	[17]
TOPIC ISLANDS	MDS, Wavelets	S	N\A	1998	[18]
VxInsight	FDP, Laplacian eigenvectors	C	-	1998	[31]
WEBSOM	SOM, Random Projections	C	-	1998	[32]
Starlight	TRUST	C	-	1999	[33]
ThemeRiver	FP	C	+	2000	[34]
Kaban and Girolami	HMM	C	+	2002	[35]
InfoSky	FDP, Voronoi Tessellations	C	-	2002	[36]
Wong et al.	MDS, Wavelets	C	-	2003	[37]
NewsMap	Treemapping	SI	˜	2004	[12]
TextPool	FDP	SI	˜	2004	[13]
Document Atlas	LSI, MDS	C	-	2005	[38]
Text Map Explorer	PROJCLUS	C	-	2006	[39]
FeatureLens	FP	C	+	2007	[40]
NewsRiver, LensRiver	FP	C	+	2007	[41]
Projection Explorer (PEx)	PROJCLUS, IDMAP, LSP, PCA	C	-	2007	[42]
SDV	PCA	S	N\A	2007	[14]
Temporal-PEx	IDMAP, LSP, DTW, CDM	C	+	2007	[43]
T-Scroll	GD, Special clustering	C	+	2007	[44]
Benson et al.	Agent-based clustering	SI	˜	2008	[11]
FACT-Graph	GD	C	+	2008	[45]
Petrović et al.	CA	C	-	2009	[46]
Document Cards	Rectangle packing	S	N\A	2009	[47]
EventRiver	Clustering, 1D MDS	C	+	2009	[8]
MemeTracker	FP, Phrase clustering	C	+	2009	[16]
STORIES	GD, Term co-occurence statistics	C	+	2009	[19]

Figure 3: *Text visualization methods presented by Šilić and Bašić [ŠB10]. The table summarizes the methods, their underlying algorithms, the publication year, whether the method includes temporal presentation or not, and the data type that the method operates on (C: Collection of text, S: Single text, SI: Short intervals)*

et al. [KPK17] that include multi-faceted classifications of text visualization approaches. We consider a survey as multi-faceted if it maps approaches into multiple dimensions, such as, tasks, interaction and presentation.

Šilić and Bašić [ŠB10] introduce three categorizations of visual approaches according to the visualization process: data types, text representation, and temporal drawing. They base their classification on the underlying algorithms and data mining techniques. They provide four user interaction methodologies commonly used when exploring text datasets.

Šilić and Bašić [ŠB10] specify three data types: a collection of text, single text, and short intervals of a text stream. See Figure 3. Additionally, the survey presents the most popular feature extraction methods used to represent text features as follows:

1. Bag-of-words methods extract text features by counting the term occurrences in the text.
2. Entity recognition aims to extract proper name of entities, such as the names of persons, organizations, places, or countries.
3. Summarization methods shorten the text and present only the most relevant information.
4. Document structure parsing extracts structural information from text, such as titles, authors names, and publication dates.
5. Sentiment and affect analysis is used to identify and quantify the emotional aspects of the text.

The survey classifies the text visualization approaches into two categories:

1. Term trend approaches are based on the term frequency in the text. In such methods, feature selection is used to reduce the number of dimensions.
2. Semantic space approaches facilitate semantic methods to extract features of text(s). In most cases, features vectors representing text are high-dimensional, so more advanced dimensionality reduction algorithms are used to map these feature to 2D or 3D space.

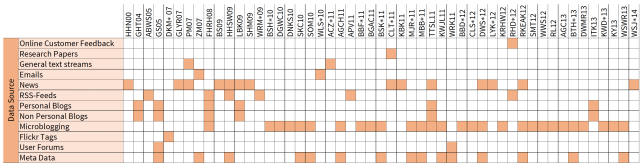

Figure 4: *Data sources classification in visual text event detection by Wanner et al [WSJ*14].*

The survey provides four exploration methodologies that help the user extract insight from the given data, as follows: brushing and linking, panning and zooming, focus-plus-context, and magic lenses.

Wanner *et al*. [WSJ*14] take a step towards defining the concept of events within text streams. They investigate the existing visual text event detection approaches and provide a event detection and exploration pipeline. An event in a text stream, as defined by [WSJ*14], is a valuable, unexpected and unique pattern extracted from the text. They classify 51 papers into different categories based on the pipeline of the event detection and exploration: text data sources, text processing methods, event detection methods, visualization methods, and supported analysis tasks. Also, the survey classifies the evaluation techniques applied in each paper.

Wanner *et al*. [WSJ*14] derive twelve data sources as shown in Figure 4. The figure clearly reveals a pattern. Since 2010 microblogging is the most common data source for visual event detection. In contrast, there is only one paper that detects and visualizes events in online customer feedback. Figure 5 shows the visualization approaches used within the investigated literature (Figure 5a) and these approaches along with event detection techniques applied (Figure 5b). We can observe that all of the clustering based techniques are mainly presented using the river metaphor. Most of the papers in Wanner *et al*. [WSJ*14] rely on use cases for evaluation (35 out of 51). On the other hand, only four papers present user studies. They suggest that more involvement from end users is encouraged.

Kucher and Kerren [KK15] present a visual survey of text visualization techniques. They classify text visualization into five top-level categories (shown in Figure 6):

- Analytic tasks include the techniques that support high-level analytic tasks.
- Visualization tasks include techniques that support lower-level representation and interaction tasks.
- Domain describes the techniques that are developed for a specific domain.
- Data consists of two subcategories, source and properties, that describe the data source and the special properties of data used by the techniques.
- Visualization contains three subcategories to describe the properties of visual representations, dimensionality, representation, and alignment.

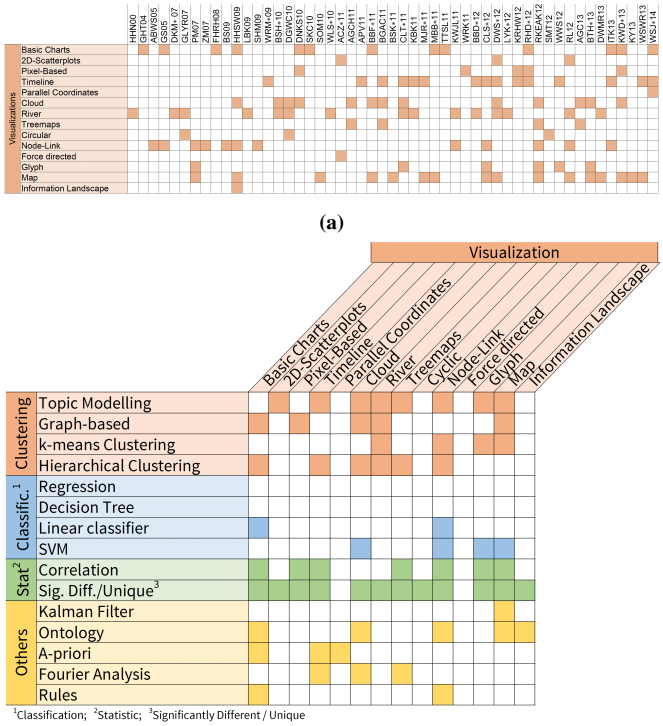

Figure 5: *(a) Visualization methods used within the surveyed collection by Wanner et al [WSJ*14]. (b) Usage of visualization and event detection techniques. [WSJ*14]*

In this section, we also include the survey by Kucher *et al.* [KPK17] which generally uses the same taxonomy as [KK15] with a focus on techniques that visualize sentiment and opinions from text data.

2.4. Cross-disciplinary Surveys

In this section we present surveys that support Digital Humanities tasks. There are two surveys which review the literature in the field of visualization that support close and distant reading of textual data by Jänicke *et al.* [JFCS15] and an extended version of this survey by Jänicke *et al.* [JFCS17]

Jänicke *et al.* [JFCS15] provide an overview of the last ten years of advancements in the field of visualization that support Digital Humanities tasks. They classify the literature based on the representation: whether it supports close reading or distant reading as proposed by Moretti [Mor05]. Close reading aims at providing direct access to the original textual content in its sequential order while distant reading does not retain the source text and provides an overview of its global features. The large availability of digital texts introduced by web portals, such as Google Books [Goo04] leads to new possibilities of close reading techniques and collaborative tools.

Jänicke *et al.* [JFCS15] classify the methods found in their collection based on task supported (close, distant or combined) reading. Furthermore, the review classifies each paper based on the underlying source text (single text, parallel, and corpus) with an extended subdivision in each category. Below, we summarize the classification proposed. Figure 7 shows a summary table of the proposed classification.

Close Reading Techniques: There are a number of techniques that have been applied in the 46 papers included in the research paper collection that provide visual features for close reading visualization as follows:

- Color is used to show a great variety of features, e.g., classification, similarity or importance.
- Font size is also used to convey text features, e.g., word frequency or significance.
- Glyphs are used to present some aspects of the text that are difficult to express using other techniques and are mostly used in poems to draw phonetic units.
- Connections help illustrate the relationship between text entities, e.g., to show subsequent words to track variation among various text editions or to convey sentence structure.

Distant Reading Techniques: 81 research papers in the collection provide an abstract distant reading view of text. There are several approaches used to visualize summarized information as the following:

- Structural overviews illustrate the hierarchy of document or collection of documents.
- Heat maps are usually used to show textual patterns, such as similarities.
- Tag clouds encode word occurrence frequency within a text using variable font size.
- Maps display geospatial information contained in a text.
- Timelines are used to visualize text that conveys temporal information. Such a technique could use the text's meta-data and support the temporal analysis of use of a word over time.
- Graphs usually use nodes and edges to visualize certain structural features of a text corpus.
- Miscellaneous methods are used to explore specific aspects within text interactively.

Techniques for Combining Close and Distant Reading: There are still some visual designs that provide both close and distant reading by preserving direct access to the source text. The 26 papers in the collection that use hybrid techniques and serve this purpose are grouped into three categories as follows:

- Top-down approaches implement the information seeking mantra "overview first, zoom and filter, details-on-demand" [Shn96]. Initially, an overview of the textual data is shown, and then the user interacts with the graphics by filtering or zooming, and finally, clicking on the interesting sub-set to obtain details-on-demand.
- Bottom-up methods start with the desired text or part of it and then generate an overview layout which relates to the given section or text.
- Top-down and bottom-up methods provide a mechanism of switching between close (text view) and distant reading (structural overview).

Jänicke *et al.* extend the survey in 2017 [JFCS17]. In terms of

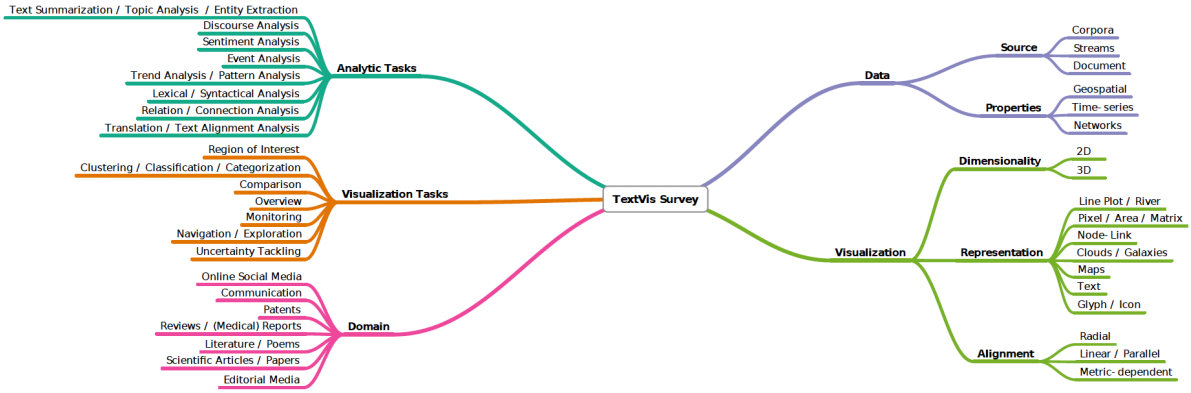

Figure 6: *The classification of text visualization techniques used in the survey by Kucher and Kerren [KK15].*

			Close Reading					Distant Reading						
			Plain	Color	Font size	Glyphs	Connections	Structure	Heat maps	Tag clouds	Maps	Timelines	Graphs	Miscellaneous
Single Text Analysis	enhanced text views	[Pie10], [CGM*12], [Pie13], [GWFI14]				x								
		[PSA*06], [CTA*13], [Ben14], [BJ14]		x										
		[ARLC*13]				x	x							
		[WMN*14]		x	x									
	both	[VCPK09], [BGHJ*14], [KJW*14]		x				x		x				
		[WJ13b], [CMLM14], [KZ14]		x			x							
		[Cay05]	x					x						
		[CDP*07]		x					x					
		[WV08]		x									x	
		[MFM13]	x										x	
		[RSDCD*13]	x											x
	abstract text views	[KO07], [FSI11], [CTA*13], [OKK13], [Ben14]							x					
		[Pie05]						x						
		[PBD14]												x
Parallel Text Analysis	section alignments	[WH11], [HKTK14]		x										
		[Cor13], [WJ13b]		x			x							
		[JRS*09]		x				x						
		[GCL*13]		x					x					
		[JGBS14b]		x			x		x					x
	sentence alignments	[BGHE10]		x			x							
		[JGBS14a]			x		x							
Corpus Analysis	statistics for textual entities	[Bea08], [Bea11], [Bea12], [BJ14]								x				
		[WJ13a], [HCC14]												x
		[CWG11]		x	x				x					
		[Mur11]		x					x					
		[FKT14]	x						x	x				
	relationships between texts	[EX10], [Gal11], [WH11], [Joc12], [CEJ*14], [Ede14]											x	
		[RRRG05]							x					
		[OST*10]		x										x
		[Wol13]		x									x	
	relationships between textual entities	[RRRG05], [AGL*07], [vHWV09], [KKL*11], [MLSU13], [WJ13a], [Arm14]											x	
		[GZ12], [RFH14]		x									x	
		[MH13]		x					x				x	
		[AKV*14]		x					x					
	social networks	[Cob05], [CSV08], [BDF*10], [RD10], [BHW11], [Kle12], [Boo13], [KOTM13], [Tót13], [Pet14]											x	
		[KLB14]	x										x	
	space and time	[JHSS12], [JW13], [DNCM14], [GDMF*14], [ÓML14]									x	x		
		[Wea08]							x		x	x		
		[BPBI10]		x							x	x		
		[DWS*12]	x							x	x	x		
		[HACQ14]		x						x	x	x		
	space	[MBL*06], [DFM*08], [Tra09], [GH11b], [EJ14]									x			
	time	[KBK11], [ARR*12], [LWW*13]											x	
		[CLT*11], [CLW14]								x		x		
		[HSC08]	x									x	x	
		[DWS*12]	x							x	x	x		
		[ESK14]								x		x		x
		[HPR14]		x								x		

Figure 7: *Hierarchical classification of research papers reviewed. At the top-right, the intended tasks supported by the visual design and the techniques implemented. On the left, the rows show the paper classification organization by Jänicke et al [JFCS15].*

classification, they add a categorization of the text analysis techniques which includes 22 more papers than the original version. The text analysis taxonomy has five main categories: named entities, topics, similar patterns, text of interest, and corpus analysis. They also, extend the discussion of collaboration experiences and future challenges.

2.5. Satellite-themed Surveys

There are two surveys that review the broader information visualization literature and then consider text visualization as a subsection in their overview classification by Sun *et al.* [SWLL13] and Liu *et al.* [LCWL14]. This is in contrast to focusing on text only like other surveys.

Sun *et al.* [SWLL13] review the recent developments in the field of visual analytics. They propose a 2D classification which they call Analytics Space. The first dimension is an applications category which includes: space & time, multivariate, text, graph and others. The second dimension is motivated by the visual analytics model proposed by Keim *et al.* [KEM10] which includes: visual mapping, model-based analysis, and user interaction. With respect to text classification, Sun *et al.* provide two categories to organize methods that process text data. The first category includes topic-based approaches which mostly leverage algorithms from Natural Language Processing (NLP). In this category, the methods that involve topics or event extraction from the text data are included e.g., TextFlow [CLT*11] and EventReader [LYK*12]. The second category is feature-based approaches which use text features to visualize text e.g., Wordle [VWF09] and FacetAtlas [CSL*10].

Similarly, Liu *et al.* [LCWL14] include a category of application within their information visualization taxonomy that includes four categories: empirical methodologies, interactions, frameworks, and application. In the application category, they [LCWL14] include four applications to classify: graph, text, map and multivariate data visual designs. There are two categories assigned to the text visualization collection. The first category is applications that visualize static textual information. In this category, they discuss and classify techniques that visualize the time-invariant content of the document(s). The second category is the visualization of dynamic textual information. In this category, they present designs that visualize temporal changes within a document or collection of documents. In both categories, Liu *et al.* group the techniques, similarly to Sun *et al.* [SWLL13], into two categories: feature-based and topic-based approaches.

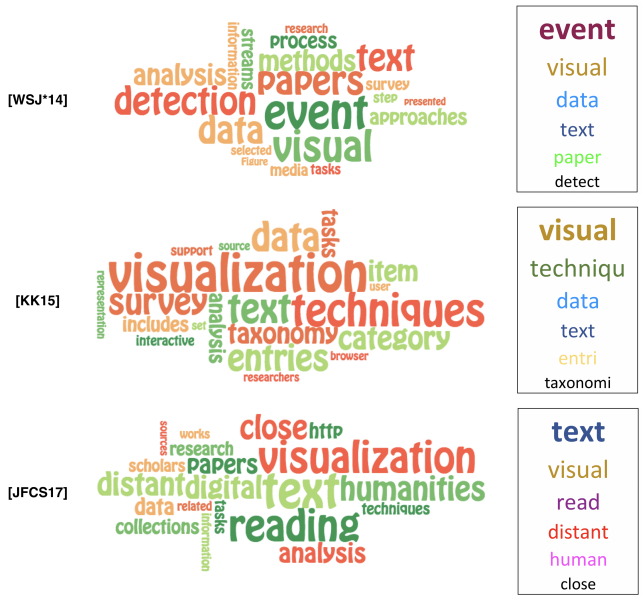

Figure 8: *A word cloud representation of the surveys [WSJ*14], [KK15], and [JFCS17] to illustrate the vocabulary used in each one. We apply the word clouds using the EdWordle online tool [WCB*18]. On the right are the most important terms in each survey using TF-IDF weighting metrics.*

2.6. Survey Recommendations

As a starting point, we think that the multi-faceted surveys are a good place to begin. Wanner *et al.* [WSJ*14] and Kucher and Kerren [KK15] provide well-crafted taxonomies. The former survey provides a guide for researchers interested in extracting events from text. The taxonomy itself is not complicated and is built on the literature they collected. It also provides a classification of the evaluation techniques that are used in each approach. Wanner *et al.* identify trends, research directions, and untouched areas in the discussion of their taxonomy which may also be beneficial for readers.

On the other hand, the Kucher and Kerren [KK15] classification covers many aspects of text visual analytics. We recommend it for researchers who would like to explore or contribute to the field of text visualization. It provides the most comprehensive and up-to-date summary of text visualization [CC16] out of the surveys. The survey's associated text visualization browser enables the user to explore and filter the collection based on the classification.

For researchers interested in the digital humanity we recommend Jänicke *et al* [JFCS17]. They provide a comprehensive overview and discussion of text visualization techniques in a humanities context.

Figure 8 illustrates the vocabulary used in those three surveys. The word clouds are produced using the EdWordle online tool [WCB*18]. We also apply TF-IDF as a term-weighting to measure the significance of each word in the surveys [SB88]. In the preprocessing phase, we remove stopwords and apply stemming algorithm to reduce inflected words. A sub-set of the result is shown in the figure on the right of each word cloud. The figure clearly illustrates the theme of each paper. Wanner *et al.* [WSJ*14] show a significant use of words, such as detection, event, and data. In Kucher and Kerren [KK15] we can see multiple terms, such as visualization, text, techniques and more importantly the term taxonomy. On the other hand, an obvious change of vocabulary in Jänicke *et al* [JFCS17] which discuss the approaches within a different context. There is more use of terms that convey digital humanities purposes that such as humanities, distant, close and reading.

3. Comparisons, Discussion of Future Challenges

In this section, we summarize the future challenges that are identified in the collection. In Table 3 we list the challenges along with the surveys. We add the McNabb and Laramee survey [ML17] to identify the overlapping challenges reported by them. We identify four unique challenges that are not reported by McNabb and Laramee since their challenges are derived from a wider perspective, These challenges are adopting advanced text mining techniques, lacking the cognitive and/or psychological analysis, lacking clear boundaries of concepts, and the need of a collaboration framework between multidisciplinary scholars.

There are nine challenges that are common to two or more surveys. Federico *et al.* [FHKM16] identify 10 future challenges, three of them are unique. The eldest two surveys by Šilić and Bašić [ŠB10] and Alancer *et al.* [AdOP12] and the two multi-faceted surveys Wanner *et al.* [WSJ*14] and Kucher and Kerren [KK15] do not feature any unique future challenges. This reflects that these surveys do not focus on a specific discipline or task. On the other hand, Jänicke *et al* [JFCS15] [JFCS17] identify two unique future challenges which indicate that they have a distinctive theme.

The most common future challenges are the need for an in-depth, effective quantitative or qualitative evaluation. This challenge is mentioned in eight surveys of the collection. Most of the surveys report a lack of in-depth evaluation of the proposed approaches. Advanced and formal evaluation provide valuable user feedback and facilitate identification of the potential problem with the systems [SWLL13]. Wanner *et al.* [WSJ*14] expect a rise of user study evaluation to verify the strength and weakness of novel visual designs. We believe that further research in the effectiveness of text visualization evaluation is encouraged.

Also, scalability and handling huge volumes of data is a common challenge. Approaches usually use various aggregations, projections, or multiple views to address this issue. However, further investigation is needed to validate the usefulness and effectiveness of such approaches, especially for scientific literature [FHKM16]. This challenge is generally associated with the challenge of adopting advanced text mining and linguistics algorithms.

Because natural language often comes with ambiguity, uncertainness, and/or errors, five surveys report this as a challenging task. Many approaches do not consider uncertainty and that could affect the analysis results. Appropriate uncertainty visualization approaches should be developed [FHKM16]. In the Text Visualization Browser [KK15], there are 25 articles that include visualization of uncertainty and ambiguity, 12 of them were published in 2016 and 2017. Jänicke *et al.* [JFCS15] [JFCS17] specifically consider the temporal and geospatial uncertainty in literature as an important

Future challenges	SoS by [ML17]	[ŠB10]	[AdOP12]	[SWLL13]	[GZL*14]	[NVPMW14]	[WSJ*14]	[LCWL14]	[KK15]	[JFCS15]	[FHKM16]	[CC16]	[KPK17]	[JFCS17]
Scalability	✓	✓	✓	✓	✓			✓			✓	✓		
Adopting advanced linguistics techniques		✓	✓	✓	✓	✓	✓				✓			
Lacking in-depth/effective quantitative or qualitative evaluation	✓	✓	✓	✓	✓		✓			✓	✓			✓
Lacking of cognitive and/or psychological analysis		✓	✓											
Natural language ambiguity and uncertainty	✓			✓				✓		✓	✓			✓
Designing general models for different tasks (versatility)	✓	✓		✓										
Lacking user interactivity that support the analysis tasks	✓		✓				✓				✓			
Lacking well-defined general concepts						✓	✓							
Multidisciplinary framework										✓	✓			✓

Table 3: *Summary of the future challenges reported in our collection of surveys. This list contains the common challenges among the surveys.*

future task. Uncertainty modeling and visualization research is expected to rise.

Another common challenge is the lack of the user interaction in order to support the analysis process. Many approaches represent the outcome of the analysis process visually, and do not provide a mean for the user to steer the underlying algorithms to further analyze the data [WSJ*14] [FHKM16]. We expect future work in the interactivity of visual analytics.

Jänicke *et al.* [JFCS15] [JFCS17] and Federico *et al.* [FHKM16] reported multidisciplinary as a challenging research topic. They suggest a systematic approach that guides and steers the work between scientist and domain experts. The former two surveys by Jänicke *et al* summarize the experiences reported regarding collaborations between visualization scientist and humanities scholars.

Lacking the cognitive and/or psychological analysis that verifies how users perceive and preserve information and incorporate it into the decision-making process is a challenging task reported in two surveys [ŠB10] and [AdOP12].

Many of the visual designs are targeted towards a specific point and do not support multiple tasks. Gan *et al.* [GZL*14] believes that it is important to design general visualization models for different tasks. Alencar *et al.* [AdOP12] confirm that it is very difficult to approach a problem without a domain-specific solution, however, users might have different goals or needs and the visual design should accommodate that.

In the text visualization community, experts always face the challenge of an ill-defined concept of 'event' and other general elements of textual data [WSJ*14]. Such a problem may distract the devoted effort of experts. Nualart-Vilaplana *et al.* [NVPMW14] also believe that the boundaries of the discipline in data visualization are not well-defined yet.

Since the surveys vary in terms of global goals and targets, there are specific challenges reported within a given context. Jänicke *et al.* report the lack of visualization approaches that represent a transposition of textual entities on all text hierarchy levels using close and distant reading. Federico *et al.* [FHKM16] expect a rise in the approaches that integrate citation analysis and other text mining techniques, such as sentiment analysis in order to reason the citation and enrich the analysis. Nualart-Vilaplana *et al.* [NVPMW14] pose an interesting question about the long-term availability of the tools. They argue that if the tool is no longer available and is not maintained for use, perhaps the tool is not effective.

4. Conclusion

In this SoS we present a meta-survey of the reviews of literature in the field of text visualization. We classify the survey collection based on five themes. We summarize each survey classification and features in order to facilitate comparisons of the surveys. Then, we provide surveys recommendations for researcher in the field of text visualization. The survey discusses and compares the field challenges reported within the collection, and examines potential future trends. This review offers a unique, valuable starting point and comprehensive overview for both newcomers and experienced researchers in text visualization.

5. Acknowledgements

We would like to thank the Technical and Vocational Training Corporation (TVTC) and the Saudi Cultural Bureau for funding and supporting this research endeavour. We would also like to thank Richard Roberts, Liam McNabb, and Dylan Rees for proofreading the manuscript.

References

[AA06] ANDRIENKO N., ANDRIENKO G.: *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer Science & Business Media, 2006. 3

[acm16] ACM digital library. http://dl.acm.org/, 2016. Accessed: 2017-5-26. 2

[AdOP12] ALENCAR A. B., DE OLIVEIRA M. C. F., PAULOVICH F. V.: Seeing beyond reading: a survey on visual text analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2* (2012), 476–492. 2, 3, 7, 8

[AE88] ANDRE P. Q., EATON N. L.: National agricultural text digitizing project. *Library Hi Tech 6*, 3 (1988), 61–66. 1

[AKS*02] ANDREWS K., KIENREICH W., SABOL V., BECKER J., DROSCHL G., KAPPE F., GRANITZER M., AUER P., TOCHTERMANN K.: The infosky visual explorer: exploiting hierarchical structure and document similarities. *Information Visualization 1*, 3-4 (2002), 166–181. 3

[AZ12] AGGARWAL C. C., ZHAI C.: A survey of text clustering algorithms. In *Mining text data*. Springer, 2012, pp. 77–128. 2

[CC16] CAO N., CUI W.: Overview of text visualization techniques. In *Introduction to Text Visualization*. Springer, 2016, pp. 11–40. 2, 3, 7, 8

[CLT*11] CUI W., LIU S., TAN L., SHI C., SONG Y., GAO Z., QU H., TONG X.: Textflow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2412–2421. 6

[CMS99] CARD S. K., MACKINLAY J. D., SHNEIDERMAN B.: *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999. 2

[CSL*10] CAO N., SUN J., LIN Y.-R., GOTZ D., LIU S., QU H.: Facetatlas: Multifaceted visualization for rich text corpora. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1172–1181. 6

[FHKM16] FEDERICO P., HEIMERL F., KOCH S., MIKSCH S.: A survey on visual approaches for analyzing scientific literature and patents. *IEEE Transactions on Visualization and Computer Graphics* (2016). 2, 3, 7, 8

[GL10] GUPTA V., LEHAL G. S.: A survey of text summarization extractive techniques. *Journal of emerging technologies in web intelligence 2*, 3 (2010), 258–268. 2

[Goo04] GOOGLE: Google books, 2004. [accessed 17-4-2017]. URL: https://books.google.com. 5

[GZL*14] GAN Q., ZHU M., LI M., LIANG T., CAO Y., ZHOU B.: Document visualization: An overview of current research. *Wiley Interdisciplinary Reviews: Computational Statistics 6* (2014), 19–36. 2, 3, 8

[Hea95] HEARST M. A.: Tilebars: visualization of term distribution information in full text information access. In *Proc. of the SIGCHI conference on Human factors in computing systems* (1995), ACM Press/Addison-Wesley Publishing Co., pp. 59–66. 3

[iee16] IEEE Xplore. http://ieeexplore.ieee.org/Xplore/home.jsp, 2016. Accessed: 2017-2-26. 2

[JFCS15] JÄNICKE S., FRANZINI G., CHEEMA M. F., SCHEUERMANN G.: On close and distant reading in digital humanities: A survey and future challenges. In *Eurographics Conference on Visualization (EuroVis)-STARs. The Eurographics Association* (2015). 2, 5, 6, 7, 8

[JFCS17] JÄNICKE S., FRANZINI G., CHEEMA M., SCHEUERMANN G.: Visual text analysis in digital humanities. In *Computer Graphics Forum* (2017), vol. 36, Wiley Online Library, pp. 226–250. 2, 5, 7, 8

[JKJ04] JUNG K., KIM K. I., JAIN A. K.: Text information extraction in images and video: a survey. *Pattern recognition 37*, 5 (2004), 977–997. 2

[KEM10] KEIM D., ELLIS G., MANSMANN F.: Mastering the information age solving problems with visual analytics. In *Eurographics* (2010), vol. 2, p. 5. 6

[KK15] KUCHER K., KERREN A.: Text visualization techniques: Taxonomy, visual survey, and community insights. In *Proc. IEEE Pacific Visualization Symposium, PacificVis* (2015), IEEE, pp. 117–121. 1, 2, 3, 4, 5, 6, 7, 8

[KPK17] KUCHER K., PARADIS C., KERREN A.: The state of the art in sentiment visualization. In *Computer Graphics Forum* (2017), Wiley Online Library. 2, 4, 5, 8

[LCWL14] LIU S., CUI W., WU Y., LIU M.: A survey on information visualization: recent advances and challenges. *The Visual Computer 30*, 12 (2014), 1373–1393. 2, 6, 8

[LH08] LIU J.-W., HUANG L.-C.: Detecting and visualizing emerging trends and transient patterns in fuel cell scientific literature. In *International Conference on Wireless Communications, Networking and Mobile Computing* (2008), IEEE, pp. 1–4. 3

[LRKC10] LEE B., RICHE N. H., KARLSON A. K., CARPENDALE S.: Sparkclouds: Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1182–1189. 3

[LYK*12] LUO D., YANG J., KRSTAJIC M., RIBARSKY W., KEIM D.: Eventriver: Visually exploring text collections with temporal references. *IEEE Transactions on Visualization and Computer Graphics 18*, 1 (2012), 93–105. 6

[MFLO14] MENDELSSON D., FALK E., L. OLIVER A.: The albert einstein archives digitization project: opening hidden treasures. *Library Hi Tech 32*, 2 (2014), 318–335. 1

[ML17] MCNABB L., LARAMEE R. S.: Survey of Surveys (SoS) - Mapping The Landscape of Survey Papers in Information Visualization. *Computer Graphics Forum* (2017). 2, 7, 8

[Mor05] MORETTI F.: *Graphs, maps, trees: abstract models for a literary history*. Verso, 2005. 5

[NVPMW14] NUALART-VILAPLANA J., PÉREZ-MONTORO M., WHITELAW M.: How we draw texts: a review of approaches to text visualization and exploration. *El profesional de la información 23* (2014), 221–235. 2, 3, 8

[RS01] REDDY R., STCLAIR G.: The million book digital library project. *Computer Science Presentation* (2001). 1

[SB88] SALTON G., BUCKLEY C.: Term-weighting approaches in automatic text retrieval. *Information processing & management 24*, 5 (1988), 513–523. 7

[ŠB10] ŠILIĆ A., BAŠIĆ B.: Visualization of text streams: A survey. *Knowledge-based and intelligent information and engineering systems* (2010), 31–43. 1, 2, 3, 4, 7, 8

[sch16] Google scholar. https://scholar.google.co.uk/, 2016. Accessed: 2017-1-20. 2

[Shn96] SHNEIDERMAN B.: The eyes have it: a task by data type taxonomy for information visualizations. In *Proc. IEEE Symposium on Visual Languages* (1996), pp. 336–343. 3, 5

[Sku02] SKUPIN A.: A cartographic approach to visualizing conference abstracts. *IEEE Computer Graphics and Applications 22*, 1 (2002), 50–58. 3

[SOR*09] STROBELT H., OELKE D., ROHRDANTZ C., STOFFEL A., KEIM D. A., DEUSSEN O.: Document cards: A top trumps visualization for documents. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 1145–1152. 3

[Ste14] STEINBOCK D.: Tagcrowd. *Internet URL: http://www. tagcrowd. com/blog/about/[accessed 2018-2-13]* (2014). 3

[SWLL13] SUN G.-D., WU Y.-C., LIANG R.-H., LIU S.-X.: A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology 28*, 5 (2013), 852–867. 2, 6, 7, 8

[VWF09] VIEGAS F. B., WATTENBERG M., FEINBERG J.: Participatory visualization with wordle. *IEEE Transactions on Visualization and Computer Graphics 15* (2009). 3, 6

[WCB*18] WANG Y., CHU X., BAO C., ZHU L., DEUSSEN O., CHEN B., SEDLMAIR M.: Edwordle: Consistency-preserving word cloud editing. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 647–656. 7

[Wis99] WISE J. A.: The ecological approach to text visualization. *Journal of the Association for Information Science and Technology 50*, 13 (1999), 1224. 3

[WSJ*14] WANNER F., STOFFEL A., JÄCKLE D., KWON B. C., WEILER A., KEIM D. A., ISAACS K. E., GIMÉNEZ A., JUSUFI I., GAMBLIN T., ET AL.: State-of-the-art report of visual analysis for event detection in text data streams. In *Computer Graphics Forum* (2014), vol. 33, Citeseer. 2, 3, 4, 5, 7, 8