Eurographics Conference on Visualization (EuroVis) 2014
H. Carr, P. Rheingans, and H. Schumann
(Guest Editors)

Volume 33 (2014), Number 3

State-of-the-Art Report of Visual Analysis for Event Detection in Text Data Streams

F. Wanner, A. Stoffel, D. Jäckle, B. C. Kwon, A. Weiler, and D. A. Keim

University of Konstanz, Germany

Abstract

Event detection from text data streams has been a popular research area in the past decade. Recently, the evolution of microblogging and social network services opens up great opportunities for various kinds of knowledge-based intelligence activities which require tracking of real-time events. In a sense, visualizations in combination with analytical processes could be a viable method for such tasks because it can be used to analyze the sheer amounts of text streams. However, data analysts and visualization experts often face grand challenges stemming out of the ill-defined concept of event and various kinds of textual data. As a result, we have few guidelines on how to build successful visual analysis tools that can handle specific event types and diverse textual data sources. Our goal is to take the first step towards answering the question by organizing insights from prior research studies on event detection and visual analysis. In the scope of this report, we summarize the evolution of event detection in combination with visual analysis over the past 14 years and provide an overview of the state-of-the-art methods. Our investigation sheds light on various kinds of research areas that can be the most beneficial to the field of visual text event analytics.

1. Introduction

With the advent of Social Media, event detection from text data streams has gained popularity in the past decade. Especially for analysts event detection plays an important role for their success. Thus, it is vexing that no single, precise definition of events has been presented so far. Within this paper, events are regarded as unexpected and unique patterns extracted from text data streams, valuable to users.

In the past, many approaches have been developed to detect events from various types of text data streams. Due to recent advancements in information science and technology, event detection has been brought a big step forward by means of its main data sources. In particular, data sources evolved from a relatively limited amount of well-written news articles to rapidly generated, user written, and in some cases unstructured textual data from social media services. This trend has been initiated with the emergence of the Web 2.0. This term was publicly introduced in 2004 and identifies a new era of the web, allowing user interaction and collaboration (Social Media).

One of the very first works on event detection was presented by Yang et al. [YPC98] in 1998. The authors carried out a study where they make use of text retrieval and clustering techniques in order to detect events in chronologically ordered news story streams. Within the same year, Allan et al. [APL98] presented a modified single-pass clustering approach for on-line event detection and information filtering. This technique was then used to track events.

Accurate event detection from user-generated text data streams posits unprecedented challenges. Regardless of the used visualization and data, all approaches define an event as something surprising, abnormal, or even unexpected that can be identified within the analysis and visualization process of the data. However, the characteristics of the event definition as well as the used approaches vary. We identified some of the most important questions that rise when talking about events. For example: “How and why did the event happen?”, “Where did the event happen?”, or “How did the event evolve over time?”. Furthermore, the tasks addressed by various visualizations occur diverse. Dou et al. [DWRZ12] defined task according to “New Event Detection”, “Event Tracking”, “Event Summarization”, and “Event Associations”, but we expect that tasks can be even more diversified including geographic dimension which were not explored yet. Another challenge represents the unstructured, diverse textual data. It mandates extensive processing and preparation in order to properly employ it.

This survey aims to take first steps towards deriving meaningful insights on the issue by investigating existing visual text event detection approaches from the past decade. With this paper, we present – to the best of our knowledge – a categorization as well as an overview of a reasonable number of the most important publications, that describe any approaches associated with visual event detection in text streams. In contrast to the work of Rohrdantz et al. [ROKF11], we concentrate on research works that focus on event detection and identification in particular.

This paper is structured as follows. First, we present related work which motivates tasks and challenges within this domain. Then, we describe our methods to select and survey papers in Section 3. From the selected papers, we first derive some trends on text data sources in Section 4. Then, Sections 5, 6, and 7 summarize automatic methods for visual text event detection and diverse visualizations of events. Finally, we present various evaluation methods in Section 8. In Section 9, we summarize our findings, discuss their implications, and highlight possible future work.

2. Related Work

Event has been defined in various ways because it has different values for different purposes. Becker [Bec11] shows interesting work about event detection in social media. She divides an event using three dimensions: 1) "planned" vs. "unplanned"; 2) "trending" vs "non-trending"; 3) "exogenous" vs. "endogenous". The last dimension aims to detect events within the data in a real-life context. This is interesting, because it raises the question which events are present in the text data and *what are their specific features and why text events occur in text streams*. The basic question here is why people write about an event. That could be partially answered for the professional news creation process by having a look into communication sciences literature. The concept of *news values* or *news criteria*, respectively were introduced by Galtung and Ruge in 1965 [GR65] and were "revisited" by Harcup and O'Neill [HO01]. Research on the question why an event reflected in a text stream is newsworthy enough to be noticed by a reader did [Kep98] by presenting the concept of *selection criteria*.

There have been some surveys on text mining methods. Hotho et al., [HNP05] define text data mining in general and explain the functionality of different natural language processing, data mining and information retrieval and information extraction methods. Berry and Castellanos [BC08] discuss text mining methods like clustering, classification, filtering, and anomaly detection methods for text collections. Anomaly detection is also a kind of event detection. In the described method the anomalies are labeled with event types which come out of the text.

Visualizations coupled with data mining methods have also been reviewed by Šilić and Bašić [ŠB10]. They characterize each text document as a form of a text stream, because it consists of smaller textual components (paragraphs, sentences, etc.). Some examples of visual social media analysis is shown in Schreck and Keim [SK13]. With screenshots of the different visualizations, the authors explain the underlying data, analysis methods, and functionality of various applications in visual social media analysis. Some of the papers in this publication are included in our survey. There exists a survey on semantic sensemaking by Bontcheva and Rout [BR12]. Though their focus was on the semantic aspects, a subsection refers to visualization approaches. Some of the mentioned papers fit also in our survey.

Especially visual analytics can be a viable method to perform various domain tasks that are related to event detection. Rohrdantz et al. [ROKF11] mention tasks for the "Real-Time Visualization of Streaming Text Data". They call tasks that are relevant in terms of the scope of our paper "monitoring", "change and trend detection" and "situational awareness". After examining their examples for each particular task, we subsumed them under event detection for our purposes. The relevant visualization tasks regarding our paper are described in Section 7. Though all of these papers are highly related, they do not focus on visual analysis for event detection. Therefore, we conduct this study to investigate the delicate issues of visual analysis for event detection.

3. Methodology

This review provides an overview of the state-of-the-art techniques to detect events in text data streams. To achieve the goal, we show the evolution of event detection in text streams over the past ten years. In particular, we investigate visual analysis approaches to derive the analysis results and detect common patterns and events. Therefore, we selected research papers that show visual analysis approaches detecting any type of events in text data using visualization for analysis and communication. We particularly paid our attention to papers which include significant contributions, such as a new algorithm, a new visualization, or a new data set.

To the best of our knowledge, there is no visual analysis pipeline specifically designed for event detection. To conduct this survey, we first started reviewing papers without any particular models in mind. While reviewing event detection and visual analysis from multiple papers, we slowly conceptualized high-level components and individual task processes. Based upon our insights, we created our own the event detection and exploration pipeline in Figure 1 adapting from Keim et al. [KAF*08]. Then, we retrospectively surveyed papers to adjust and improve this pipeline. We believe that this pipeline reflects both the visual analysis process and the more interactive visual analytics process for event detection to a reasonable extent.

In the first step of the pipeline, the documents are prepared

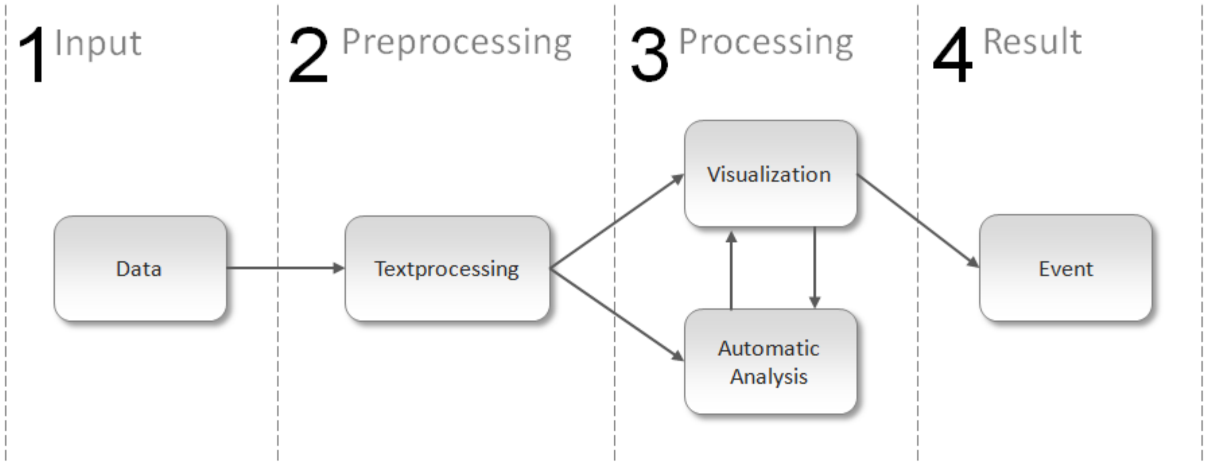


Figure 1: *The event detection and exploration pipeline used to structure this report.*

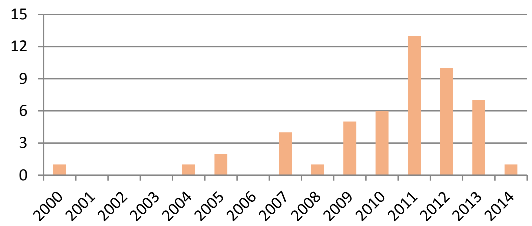


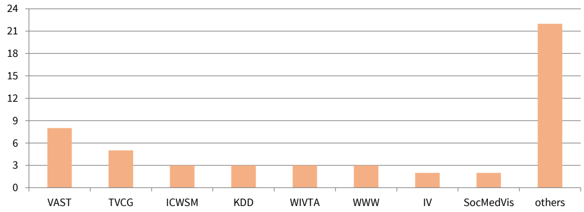
Figure 2: *Distribution of surveyed papers over publication year. The majority of papers were published between 2007 and 2013. We consider only one paper from 2014, because this report was written in early 2014.*

for the analysis. In this step the documents are parsed to get the plain texts and standard text preprocessing methods, such as sentence detection, tokenizing, and stemming and lemmatizing are applied. In addition to these standard methods, methods from the computer linguistic field can be used in the preprocessing step to annotate the texts with additional information. For instance, part-of-speech tagging, named entity extraction, or syntactic parsing can be used to identify types of words, persons and places, or structure of sentences. The standard preprocessing techniques are required to work with text data in any case. We therefore focus on the usage of computer linguistic methods in this review.

After the preprocessing step different approaches are used to detect events (see two branches in Processing in Figure 1). The first group of approaches applies automatic methods to detect patterns in the data. The detected patterns are then used to create a visual analysis interface for the data set, what we call visual analysis. The interaction between visualization and the automatic part shapes a visual analytics approach. The second group of approaches skips the automatic analysis and directly visualizes the outcome of the preprocessing, what also is only visual analysis because of the lack of interaction possibilities of a certain extent. The used visualization and interaction techniques depend on the type of data and the requirements of the different approaches. This report investigates the applied automatic data analysis methods as well as the visualization methods.

In the last step of our pipeline, users interact with the visualization and derive the requested knowledge. As our selection of papers requested visual approaches, all surveyed papers use visualizations to communicate results. The papers often show the validity and usefulness of their approach by comparing with different techniques. In this report we summarize the used evaluation techniques and give an overview of their usage. We used the pipeline to review individual papers; we derived subcategories (e.g., POS tagging for text processing methods), and checked whether individual papers include them. Following sections are structured in the same way so that readers can easily follow.

We used the following procedure to select papers for our review (Figure 4). We first archived papers from previous survey papers, e.g. [ŠB10, BR12]. Then, we used the digital libraries of IEEE Xplore, ACM and AAAI to search documents that include all the following terms in their title or abstract (for AAAI we used Google Scholar): “visual”, “text”, “event” and “analysis”. This querying process resulted in more than 280 research papers. In addition, we took research on visual text event analysis into account which does not calling it explicitly analogous. After the collection step, we refine our paper pool by filtering papers out using the following criteria: 1) papers were published from 2000; 2) papers should describe visualization methods as well as event detection (an event has to have a time dimension); 3) a newer paper was selected when multiple versions of similar methods were available from the same research group. As a result, we included 51 papers in total for our review. Figure 2 shows



VAST	IEEE Conference on Visual Analytics Science and Technology
TVCG	IEEE Transactions on Visualization and Computer Graphics
ICWSM	International AAAI Conference on Weblogs and Social Media
KDD	ACM SigKDD Conference Knowledge Discovery and Data Mining
WWW	International World Wide Web Conference
IV	IEEE International Conference on Information Visualisation
SocMedVis	Workshop on Social Media Visualization
WIVTA	Workshop on Interactive Visual Text Analytics

Figure 3: *Number of papers selected by venue. The category "others" summarizes all venues with a single paper.*

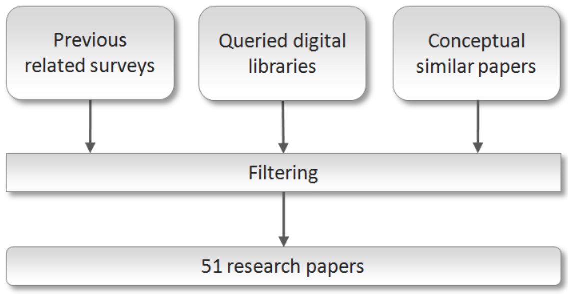


Figure 4: *The paper selection process. In a first step we collected the papers from several sources. Then we applied a filter step to get the result of 51 papers.*

the distribution of the papers over publication year and Figure 3 shows the number of papers selected by venue.

4. Text Data Sources

We derived twelve distinctive text data sources used in the 51 papers, as Table 1 shows. We summarize characteristics and trends of such data sources, which could be useful for future researchers.

4.1. Characteristics

News is a well-known text data source. News captures information of a real world event or happening. It consists of a title, often followed by a short summary and the body containing details about the event. News goes through a professional gatekeeping process which in the end forms the agenda of media. Whitney and Becker [WB80] describe the process as follows: "[...] the agenda being presented by the media audiences is influenced by the newsgathering procedures of the media and the relationships among the media." During this process, the published news has to pass different "gates". Each gate could shut down in case the news is not newsworthy enough. At the beginning of a simplified publication process the first gate is the reporter himself. As he investigates an event in detail, he has to decide if it is worth to become news or not. Later in the paper we will show some criteria for such a decision. In case the information about an event is proven newsworthy by the journalist, it has to pass the next gate which is the editorial journalist. If the news is not interesting or newsworthy enough, does not fit to the agenda or there is no space left, the news will not be published. Since news is also published on the Internet, the space limitation does not form a large barrier anymore. However, it is still limitation for the printed version of a newspaper.

A typical electronic document is email. Its ancestor is postal mail which is sent to real postboxes close to the entrance doors of houses and apartments. Emails are used for personal conversations, advertisement or business information exchange. They consist of a header and a body. The header contains information about transaction: sender, receiver, timestamp, and other meta data. The body contains the textual content of the email. An email body can be of arbitrary length which is one of its characteristics.

Weblogs, shortly named blogs are used for information purposes of a more or less undefined audience. Beside other providers, WordPress for example provides software to create, maintain and design weblogs in an easy and convenient manner (Software available at http://en.wordpress.com/features). A blog can have a specific topic or can be open for various topics. There are personal or non-personal blogs [KLS*05, EdR08]. In the scope of the survey we take personal blogs as a matter of private citizens whereas non-personal blogs are written by non-governmental organizations (NGOs), companies, enterprises or professionals in general.

RSS feeds are a standardized format to broadcast short news snippets. They consist of a title and a description. RSS feeds can be used by news agencies, newspapers and blogs. In contrast to former pull services, RSS was developed as push communication. If users want to receive a feed, they need a RSS reader and have to subscribe for a feed that they are interested in. The reader aggregates all the incoming messages. The standardized format allows the easy integration into other applications.

Recently, microblogging providers are becoming more and more popular. The messages are limited with respect to their length of 140 characters. So-called "hashtags" are used in order to characterize the membership of a tweet to a certain topic. In addition, more meta data is provided, e.g. geolocation, author, place etc. That depends on whether users have these meta data fields enabled. The largest service is Twitter (https://twitter.com/). Since its birth in March 2006, it has more than one billion of registered users [Koe13] and more than 232 million monthly active users [Con13]. The to-

Data Source	HHN00	GHT04	ABWS05	GS05	DKM∗07	GLYR07	PM07	ZM07	FHRH08	BS09	HHSW09	LBK09	SHM09	WRM∗09	BSH∗10	DGWC10	DNKS10	SKC10	SOM10	WLS∗10	ACZ∗11	AGCH11	APV11	BBF∗11	BGAC11	BSK∗11
Online Customer Feedback																										
Research Papers																										
General text streams							x														x					
Emails								x												x						
News	x					x	x			x	x	x	x													
RSS-Feeds			x						x		x			x									x			
Personal Blogs		x		x					x			x														
Non Personal Blogs		x		x					x			x														
Microblogging									x						x	x	x	x	x			x		x	x	x
Flickr Tags					x																					
User Forums				x																						
Meta Data				x				x			x							x	x			x				x

Data Source	CLT∗11	KBK11	MJR∗11	MBB∗11	TTSL11	KWJL11	WRK11	BBD∗12	CLS∗12	DWS∗12	LYK∗12	KRHW12	RHD∗12	RKEAK12	SMT12	WWS12	RL12	AGC13	BTH∗13	DWMR13	ITK13	KWD∗13	KY13	WSWR13	WSJ∗14
Online Customer Feedback													x												
Research Papers	x																								
General text streams																									
Emails																									
News	x	x			x					x	x			x											x
RSS-Feeds													x												
Personal Blogs					x																x				
Non Personal Blogs					x																x				
Microblogging			x	x	x	x		x	x	x		x		x	x	x	x	x	x	x		x	x	x	
Flickr Tags																									
User Forums							x																		
Meta Data			x	x		x	x			x				x					x					x	

Table 1: *Data sources used in visual text event detection. Before microblogging sources got popular in 2010 news data were the most common data source for visual detect event. Since 2010 many techniques are developed for microblogging data. Interesting to see is that visual event detection is only used once on customer feedback data.*

tal number of tweets exceeds 300 billion, the average number of daily tweets is about 500 million and about 75 % of users access Twitter by using mobile devices [sec13]. In order to receive tweets a user has to be a follower of a sender.

User forums often have hierarchical structure. A message within the forum is a post and is not strictly restricted with respect to its length. Posts which belong to the same topic shape a so called thread. On the other hand several threads often belong to a sub-forum within the main forum. The purpose of a forum is the discussion on specific issues and topics regarding the its main topic. That is, why they are also called discussion forums. There is usually one or more moderators who monitor the discussions in order to guarantee the compliance with the forum regulations. Authors have to be registered users who are logged-in. They can post and comment on former entries while visitors typically have only read access to articles. Depending on the forum it can also be exclusively accessible for registered users.

Modern customer-care systems often ask each customer to fill out a feedback form after a purchase. This form (often digital, reachable through the internet) gives the customer the opportunity to provide issues directly to the vendor. The information is a valuable source which allows the seller to react fast and adequately to issues being raised by customers. The more purchases the more feedback forms the company receives. Often these forms are semi-structured, which means they have checkboxes for predefined questions and provide a free text field for further comments.

Images and video sequences can be uploaded on sharing sites such as Flickr (https://www.flickr.com/). Users can tag their content with text. These tags and little text snippets typically describe the content in a short manner or express an emotional state being associated with the photo. In Flickr the photos can be displayed on a map according to their geographical location.

4.2. Trends

Visual text event detection is often application-driven. So the application determines the text data source which can be either static or dynamic. In terms of this categorization static means a document collection which is not characterized by frequent updates, e.g., a set of domain specific research papers. In contrast, there are in general any other text streams which are characterized by many updates and rather short textual content.

Almost half of the papers use microblogging data namely Twitter. It is obvious in Table 1 that in 2010 a shift towards microblogging happened. It is also noticeable that meta data (geolocations, author information) is often used in conjunction with microblogs. This trend is still continuing. Beside these two dominant sources, news is the third one that provides text input. Another fact is the minor occurrence of comprehensive text sources like user discussion forums and customer feedback data. We think that in the future deeper text analysis methods could be applied and tested there. In addition, the text of news and microblogs is typically longer which tends to convey more structural and syntactical information. That can provide further useful features in future visual event detection scenarios.

5. Text Processing Methods

Analyzing text data requires several transformation steps to split text into meaningful units and/or calculate descriptive metrics. The text processing step in the event detection and exploration pipeline (see Figure 1) is typically a sequence of different text transformations. In this report we do not consider basic text processing methods because these are required steps for almost all text analyses. For instance, most approaches of text data mining require sentence or token splitting techniques. Instead of such basic processing methods, this study focuses on advanced techniques collected from our paper pool. Table 2 summarizes the techniques used by the surveyed papers.

		HHN00	GHT04	ABWS05	GS05	DKM∗07	GLYR07	PM07	ZM07	FHRH08	BS09	HHSW09	LBK09	SHM09	WRM∗09	BSH∗10	DGWC10	DNKS10
Text Processing	Shallow syntactic parsing				■													
	Full syntactic parsing																	
	Typed-dependency parsing							■										
	Named Entity Recognition		■					■						■				
	Coreference Resolution																	
	Polarity extraction				■										■			■
	POS tagging		■		■			■						■		■		
	Disambiguation							■				■						

		SKC10	SOM10	WLS∗10	ACZ∗11	AGCH11	APV11	BBF∗11	BGAC11	BSK∗11	CLT∗11	KBK11	MJR∗11	MBB∗11	TTSL11	KWJL11	WRK11	BBD∗12
Text Processing	Shallow syntactic parsing										■							
	Full syntactic parsing																	
	Typed-dependency parsing																	
	Named Entity Recognition										■	■	■		■			
	Coreference Resolution																	
	Polarity extraction						■		■					■	■		■	
	POS tagging										■					■		
	Disambiguation																	

		CLS∗12	DWS∗12	LYK∗12	KRHW12	RHD∗12	RKEAK12	SMT12	WWS12	RL12	AGC13	BTH∗13	DWMR13	ITK13	KWD∗13	KY13	WSWR13	WSJ∗14
Text Processing	Shallow syntactic parsing																	■
	Full syntactic parsing																	■
	Typed-dependency parsing													■				■
	Named Entity Recognition		■	■									■		■			■
	Coreference Resolution			■														
	Polarity extraction				■	■	■		■		■							■
	POS tagging		■	■	■	■	■	■	■				■			■	■	■
	Disambiguation																	

Table 2: *Text processing methods. The most popular text processing methods are part-of-speech tagging, polarity extraction, and named entity recognition. Twenty papers do not apply any text processing except for standard tokenizing or lemmatization. Only in exceptional cases, more complex linguistic techniques, such as syntactic or dependency parsing, are used.*

5.1. Applied Techniques

Part-of-speech (POS) tagging detects the word type of tokens. This is a basic technique from computational linguistics to disambiguate grammatical types and relations of words within text chunks. POS tags are used to determine attributes assigned to nouns or nouns describing a document. We consider this as advanced technique because it was rarely used in the surveyed papers published before 2012.

Syntactic parsing determines the grammatical structure of sentences. As many other computational linguistics methods do, surveyed papers also use full and shallow algorithms. Full syntactic parsing uses grammars and build up a complete parse tree for a sentence. Due to ambiguities in language, several valid parse trees for a single sentence can exist and usually the most likely one is used for further processing. Shallow parsing creates meaningful chunks and avoids the complexity of full parsing. Shallow parsing algorithms are therefore computationally less expensive. Interestingly, we found only three papers using syntactic parsing probably because the majority of surveyed papers process text on token levels without considering grammatical structures.

Typed-dependency parsing determines the type of relations between words in a sentence. The created dependency network describes semantic connections between words. An example relation could be that one word negates the meaning of another one. Typed-dependency parsing is used to extract semantics in text, which is not possible with syntactic parsing, but it is computationally expensive. We conjecture that this is why typed-dependency parsing is used in only one of the surveyed paper.

Coreference resolution creates connection between referring expression, such as pronouns, and subjects in a text. A correct resolution of referring expressions could improve text mining results, e.g., polarity extraction would benefit from correctly resolved referring expressions. Due to the complexity of the problem, coreference resolution is seldom used; it is often substituted by a simple heuristic yielding comparable results.

Named entity recognition (NER) detects and labels names of, e.g., persons, locations, events, or dates in texts. Grammars or machine learning techniques can be used for NER, but very often are dictionary-based approaches in use. NER techniques are, for instance, used to extract character networks from documents or determine geographic locations that documents describe. The surveyed papers using NER are mainly extracting persons or location from data, which are later on used to identify events or generate visualizations.

Polarity extraction or determines the attitude (positive vs. negative) of the writer about a subject. Usually word-based sentiment dictionaries are used to compute sentiment scores. This sentiment score is then used to determine an overall polarity of a document or a topic. The quality of sentiment analysis highly depends on the sentence structure and language use. It is often impacted by how popular issues like negation or anaphora resolution are resolved.

Words can have different senses, which may mislead simple text mining algorithms that consider tokens literally. Word-sense disambiguation techniques use the context of words to determine the correct sense of tokens. Text mining methods, such as topic modeling or latent semantic analysis, can substitute for word-sense disambiguation by considering contexts of tokens. We only observed one paper using word sense disambiguation.

5.2. Trends

We confirm that text processing methods are used very sparingly. As Table 2 shows, 31 out of 51 papers used one or more text processing methods we described. Since 2000 until the end of 2011, only 17 out of 33 papers utilized any of the methods. The 16 papers with no text processing methods solved the event detection tasks with visualization. In the year of 2012, the trend changed dramatically; 14 out of 18 papers have used text processing methods in the papers published since then. It is also noticeable that part-of-speech tagging and polarity extraction have gained popularity since 2012 as well. We may conjecture a reason for such shift in trend. Many prior studies in the early 2000s focused on the global overview from large text corpus. To achieve their goals, many applied simple keyword extraction algorithms such as *tf-idf*. Since we considered the eight named methods (as shown in the second column of Table 2), many other elementary methods (e.g., tf-idf) have been excluded. Diverse research topics emerged recently, which require in-depth analysis of text, perhaps about the quality of a subset of text (e.g., topics, sentiments). Thus, we believe that many research papers started absorbing more natural language processing techniques to further generate their event metrics.

6. Automatic Methods for Text Event Detection

We observed four categories of event detection techniques; they are further subdivided into 15 categories (see Table 3). In a big picture, we have seen few papers that adopt any automatic event detection methods. The two most popular method categories were clustering and statistical methods. The popularity may be highly correlated to how the event is defined in the papers. For instance, many research papers defined the event as an irregular activities compared with average signs across the temporal dimension in given data. Thus, such papers seek to find statistically significant difference in textual data, which naturally leads to the popularity of statistical methods. Given that many topic modeling techniques were also adapted with respect to temporal dimension to narrow down events in fine granularity (e.g., Zhao and Mitra [ZM07]). However, very few papers incorporated classification, prediction, knowledge modeling, or pattern mining.

6.1. Event Detection Methods

A widely used family of algorithms to detect events is based on clustering techniques. Clusters are generated for different time windows based different properties in the document, e.g., co-occurrence of terms, frequency in time, or metadata. Events are generated when the set of clusters changes, e.g., a new cluster arise or two existing clusters merge. With clustering based methods users do not need to specify the type of events but the algorithms detect changes in the data.

Other approaches to detect specific events are based on classifiers. Users provide a set of example documents and classifiers learn to detect the annotated events. Classifier-based techniques are used in similar cases with rule-based ones, but have the advantage that users do not need to create rules by themselves. In our surveyed papers, only two papers use classifiers to detect user specific events. This accords with the finding of the rare usage of rule-based methods and indicates that visual analysis approaches are mainly used to explore new events in text streams.

Statistical methods such as correlation or detection of outliers and significant difference are used to identify events. Correlation based methods examine collection between terms or between terms and time and detect events by changes in the correlation measures. A different type of statistical methods calculate term-wise deviation from an expected value or use other measures to identify rare or unique occurrence of terms. These statistical abnormalities are then assessed and events are identified.

Prediction-based methods predict the occurrence of following documents based upon past history. We found one paper using Kalman filters [SOM10] for prediction. Based upon many geo-tagged documents, predictions can be made on the location of newly added documents. Thus, this method can detect abnormal events which deviate from prediction. Thus, prediction-based methods require archived past events to precisely anticipate the future events. We found only one paper that applies a prediction method to detect natural disasters.

Methods based on ontologies [HHSW09] are suitable for event analysis in single domains. Specific ontologies are generated with full- or semi-automatic methods. The only paper in our set using ontologies aims to identify concepts appearing in documents and time frames. Using this type of methods, events can then be detected from changes in activated concepts.

Pattern mining algorithms, such as the A-priori algorithm of Wu and Chen [WC09] applied to text in [WSJ*14], are used to extract common sequential patterns in document streams. Patterns can be found based on documents themselves or time intervals. In both cases, features extracted from documents are then used to define patterns. Instead of identifying single events, pattern mining algorithms aim to derive recurring patterns in data. In our selected papers, we found only a single paper using pattern mining technique, which does not only identify events but characterizes these events based upon their pattern.

Models of the recurring characteristics of data steams can be used to detect events. An event is detected when a stream deviates significantly from its expected characteristics. We only found one paper [BGAC11] using Fourier analysis to model the base frequencies of a data stream and using them as model to detect events.

Rule-based approaches detect events with manually created rules. For instance, users specify rules based on terms and/or frequency to detect a particular event. The events that can be detected by these approaches are only limited by the expressiveness of rule languages. In contrast to other techniques are events detected with these approaches predictable for users but require precise knowledge of the data and the events. To our surprise is only one paper [GS05] using rule-based definition of events and allows expert users to define events by queries.

6.2. Trends

We cannot find any correlation between text analysis and event detection methods. All event detection methods are used with basic text tokens and also with text processing results. Interestingly, only two papers build classifiers

Group	Method	HHN00	GHT04	ABWS05	GS05	DKM*07	GLYR07	PM07	ZM07	FHRH08	BS09	HHSW09	LBK09	SHM09	WRM*09	BSH*10	DGWC10	DNKS10	SKC10	SOM10	WLS*10	ACZ*11	AGCH11	APV11	BBF*11	BGAC11	BSK*11
Clustering	Topic Modelling							x	x							x					x			x			
Clustering	Graph-based												x	x													
Clustering	k-means Clustering							x													x						
Clustering	Hierarchical Clustering								x												x		x			x	
Classific.[1]	Regression																										
Classific.[1]	Decision Tree																										
Classific.[1]	Linear classifier				x																						
Classific.[1]	SVM							x																			
Stat[2]	Correlation									x																	
Stat[2]	Sig. Diff./Unique[3]		x			x	x				x	x					x								x		
Others	Kalman Filter																			x							
Others	Ontology											x															
Others	A-priori																										
Others	Fourier Analysis																									x	
Others	Rules				x																						

Group	Method	CLT*11	KBK11	MJR*11	MBB*11	TTSL11	KWJL11	WRK11	BBD*12	CLS*12	DWS*12	LYK*12	KRHW12	RHD*12	RKEAK12	SMT12	WWS12	RL12	AGC13	BTH*13	DWMR13	ITK13	KWD*13	KY13	WSWR13	WSJ*14
Clustering	Topic Modelling	x									x				x					x	x					
Clustering	Graph-based																						x			
Clustering	k-means Clustering																									
Clustering	Hierarchical Clustering											x							x			x				
Classific.[1]	Regression																									
Classific.[1]	Decision Tree																									
Classific.[1]	Linear classifier																									
Classific.[1]	SVM																			x						
Stat[2]	Correlation													x	x											
Stat[2]	Sig. Diff./Unique[3]				x						x		x	x	x									x	x	
Others	Kalman Filter																									
Others	Ontology																									
Others	A-priori																									x
Others	Fourier Analysis																									
Others	Rules																									

[1]Classification; [2]Statistic; [3]Significantly Different / Unique

Table 3: *Automatic event detection methods. Statistics and clustering are the most important methods to detect events in text data. Rules, classification, and prediction techniques do not play a role.*

and use these classification models to detect events. Scatterblogs2 [BTH*13] allows interactive creation of classifiers with relevance feedback. This allows an analyst to define a specific type of event based on textual content or meta information. For the rest of cases in our paper selection, the systems are designed to detect and analyze events with a predefined characteristic.

7. Visualization of Events in Text Data

We derived 14 different visualization techniques that are used for event detection in text (see Table 4). We excluded document lists (a view showing list of documents) or text views (a view showing text) because all of our surveyed papers use them. We found that all visualization methods except for the cloud method are techniques for general data visualization. In total, 21 papers use a time-oriented visualization (river, timeline, cyclic) to visualize the evolution of the data over time. Time-oriented visualizations are often combined with additional visualizations to show non-time dependent information. Maps visualization came up with microblogging data and use mainly geographic references in the meta information of the microblogs for visualization.

A problem for all visualizations is the question how to visually represent text data. This problem is usually solved by selecting meaningful keywords that are either generated by frequency or by another scoring technique such as topic models. The selected keywords are then used directly in the visualization for annotations, used as labels in a legend, or shown in word clouds.

Basic visualizations (e.g. line or bar charts) are mainly used to give an overview of the data set by showing the time dependent relations of events. They are used to visualize the data volumes or frequencies over time, for instance, of detected topics, named entities, or keywords. An alternative visualization for these tasks are either rivers or timeline based visualizations. We therefore did not observe any papers using river visualizations together with standard charts. The decision whether river or timeline visualizations seems to depend on the purpose of the visualization. Timeline visualization use one or multiple timelines and place glyphs or shapes on these timelines to indicate single data items, densities, or volumes. Timeline visualizations are therefore preferred over river visualizations when single or rare items should be tracked, because a river visualization put the focus on high frequent events. River metaphors are often used to visualize outcomes of cluster algorithms. Although timeline techniques could be used, rivers provide a space saving overview and give a better visual impression of the distributions of the clusters and the overall amount of data.

7.1. Supported Analysis Tasks

Selection of visualization methods often depend upon users' intended tasks. In Table 5 the different tasks supported by the different systems are shown.

Almost all surveyed systems provide an overview of the text collection. Overview visualization give users a summary of the document collection. Common are textual summaries based on frequent terms or topic models that describe the topics found in the collection. These summaries serve as navigation support and are often used as starting point for further analysis.

It is also common to provide users with abilities to search for keywords or allow filtering of the data by meta information. Both tasks reduce the number of item or extracted events in the analysis or visualization. In search and retrieval tasks users are able to specify open queries based on content or meta information. Only items or events matching the query are used in analyses or shown in visualizations. Filters, in contrast to search and retrieval, are based on predefined data properties and are usually implemented with specific controls. Filter abilities are designed to support predefined

Visualizations	HHN00	GHT04	ABWS05	GS05	DKM* 07	GLYR07	PM07	ZM07	FHRH08	BS09	HHSW09	LBK09	SHM09	WRM*09	BSH*10	DGWC10	DNKS10	SKC10	SOM10	WLS*10	ACZ*11	AGCH11	APV11	BBF*11	BGAC11
Basic Charts		x		x					x		x						x	x						x	
2D-Scatterplots																					x				
Pixel-Based																	x								
Timeline														x									x		x
Parallel Coordinates																									
Cloud											x				x	x	x					x		x	x
River	x				x	x						x			x	x				x					
Treemaps																						x			x
Cyclic						x										x									
Node-Link			x	x			x	x		x	x														
Force directed																					x				
Glyph							x																		
Map							x				x								x					x	
Information Landscape											x														

Visualizations	BSK*11	CLT*11	KBK11	MJR*11	MBB*11	TTSL11	KWJL11	WRK11	BBD*12	CLS*12	DWS*12	LYK*12	KRHW12	RHD*12	RKEAK12	SMT12	WWS12	RL12	AGC13	BTH*13	DWMR13	ITK13	KWD*13	KY13	WSWR13	WSJ*14
Basic Charts					x	x								x								x	x			x
2D-Scatterplots											x							x								
Pixel-Based								x					x	x									x			
Timeline		x	x	x					x	x	x		x	x			x					x			x	x
Parallel Coordinates																										x
Cloud		x									x								x	x			x			
River		x	x				x		x		x	x						x			x					
Treemaps															x				x							
Cyclic																x										
Node-Link							x			x					x			x				x				
Force directed																										
Glyph		x								x					x		x			x						
Map	x			x	x					x	x				x			x		x	x		x	x	x	
Information Landscape																										

Table 4: *Visualization methods for event detection. Word clouds are the only visualization technique based on textual data. The majority of papers use time-based visualizations (river, timeline, cyclic) together with additional visualization of meta information. With geographic information in microblogging data, map visualization beame more popular recently.*

Tasks	HHN00	GHT04	ABWS05	GS05	DKM* 07	GLYR07	PM07	ZM07	FHRH08	BS09	HHSW09	LBK09	SHM09	WRM*09	BSH*10	DGWC10	DNKS10	SKC10	SOM10	WLS*10	ACZ*11	AGCH11	APV11	BBF*11	BGAC11
Overview over Documents	x		x	x	x	x	x	x		x	x	x		x	x	x	x			x	x	x	x	x	x
Search and Retrieval		x				x	x		x		x				x	x	x			x		x		x	x
Filtering			x	x		x	x			x	x			x	x	x	x			x	x		x	x	x
Monitoring	x	x		x		x	x		x	x	x			x	x	x	x	x	x	x	x	x	x	x	x
Relations between Events	x	x		x		x	x	x	x		x	x		x	x	x				x	x	x		x	x
Relations to Other Data																	x							x	
Details on Demand				x		x	x		x	x				x	x					x	x			x	

Tasks	BSK*11	CLT*11	KBK11	MJR*11	MBB*11	TTSL11	KWJL11	WRK11	BBD*12	CLS*12	DWS*12	LYK*12	KRHW12	RHD*12	RKEAK12	SMT12	WWS12	RL12	AGC13	BTH*13	DWMR13	ITK13	KWD*13	KY13	WSWR13	WSJ*14
Overview over Documents		x	x			x	x	x	x	x	x	x	x	x	x	x		x	x		x	x	x	x	x	
Search and Retrieval				x	x		x	x	x		x	x	x	x		x	x		x	x						
Filtering	x	x		x	x	x		x	x	x	x	x			x					x	x		x			x
Monitoring	x			x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x		x	
Relations between Events	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x		x	
Relations to Other Data	x					x																				x
Details on Demand		x	x	x	x			x	x	x	x	x	x	x		x				x	x	x		x		x

Table 5: *The tasks supported by the different systems. Almost all systems can generate an overview of the data sets, allow monitoring of single events, and show relations between the events. Searching for keywords or filtering data according to meta information are also common interaction techniques. Detail-on-demand techniques are not implemented in all cases.*

analyses tasks and users are not able to change the supported set of filters dynamically.

Monitoring tasks are the second most frequent tasks supported by the surveyed systems. Users monitoring a data source are interested in the evolution of events in a changing data source. Time-based visualizations (e.g., timeline, river, cyclic) are often used for monitoring task, because they show the temporal development of events in data sources.

In many cases relations between different detected events are visualized. The most frequent shown relations are relation in content, time, and volume. For instance, a river visualization shows time and volume relations between different streams and with additional annotations also relations in content can be shown. Until 2011, usually relations of time and data volume were of interests in many papers. Since 2011 map visualization get more and more popular and additionally also geographic relations between events are investigated. We observed many papers showing relations between events from the same data source. Only a view papers extract events and relate them with different data sources, for instance, Diakopoulos et al. [DNKS10] use Twitter to detect events in broadcast news.

Surprisingly, almost half of the surveyed papers only provide an overview and show their analytic or visual results but do not implement a details-on-demand functionality. These papers do not even provide an ability to access the papers associated with an event or data point in visualizations.

7.2. Visualization and Event Detection Methods

We also investigated which visualization methods and event detection methods are used together in Table 6. The statistical methods are combined with any type of visualizations. Interestingly, clustering methods are often visualized by river visualizations. Exceptionally, topic modeling techniques are not only used with time dependent visualizations but also with other visualizations such as treemaps or geographic visualizations. This pattern appears because topic models are clustering methods that return a ranked list of terms representing single topics, which are often used in visualizations to label data and find names for clusters.

8. Evaluation Methods

Evaluation plays a crucial role when it comes to validate benefits of newly introduced techniques and algorithms. Methods for visual text event detection have been evaluated and classified according to the table we designed (see Table 7). In particular, we identified nine different evaluation

		Visualization													
		Basic Charts	2D-Scatterplots	Pixel-Based	Timeline	Parallel Coordinates	Cloud	River	Treemaps	Cyclic	Node-Link	Force directed	Glyph	Map	Information Landscape
Clustering	Topic Modelling		■		■		■	■	■		■		■	■	
	Graph-based	■		■			■	■						■	
	k-means Clustering							■			■		■	■	
	Hierarchical Clustering	■			■		■	■	■		■				
Classific.[1]	Regression														
	Decision Tree														
	Linear classifier	■									■				
	SVM						■				■		■	■	
Stat[2]	Correlation	■		■	■				■		■		■	■	
	Sig. Diff./Unique[3]	■	■	■	■		■	■	■	■	■		■	■	■
Others	Kalman Filter													■	
	Ontology	■					■				■			■	■
	A-priori	■			■	■									
	Fourier Analysis				■		■		■						
	Rules	■									■				

[1]Classification; [2]Statistic; [3]Significantly Different / Unique

Table 6: *Usage of visualization and event detection techniques. Statistical event detection methods are used with any kind of visualization. In contrast, the clustering based methods are mainly visualized with the river metaphor. The exception are topic modeling techniques, which are often used to generate labels in visualizations or legends.*

categories across a representative number of papers that address visual text event detection. We subdivided these categories into two classes: *qualitative* and *quantitative* evaluation methods.

The table reveals one common trend at first glance: Use cases are very popular, especially since 2011. In this chapter we will highlight the main differences between common evaluation methods and unfold significant patterns.

We subdivide qualitative methods into the following categories: *case study, usability evaluation, use case, and anecdotal evaluation*. Table 7 accentuates the popular usage of use cases; except for 16 of all considered papers the authors make use of this method. Typically, a use case validates through the description of a fictitious scenario that pinpoints main features whereas a case study involves a domain expert and therefore is more time-consuming [DNKS10, MBB*11]. This may be one of the reasons why use cases are more popular. Anecdotal evaluation describes how the suggested system could be used, but do not provide sufficient evidence to judge the general efficacy of the presented technique. Usability evaluations involve users performing particular tasks with the given system and asks for comments on usability.

Despite the fact that many systems suggests automatic event detection methods, quantitative evaluation are seldom used. The most prominent quantitative evaluation methods are comparisons of the detected events with a ground truth set. Often event databases are used as ground truth that are enriched by the authors with missing entries. Missing reliable ground truth might be the major reason why algorithms are in many cases not evaluated. A different evaluation form of algorithms are comparison with existing algorithms and reporting quality measures. In some cases not the results of the algorithms are evaluated but the performance in the sense of runtime or memory consumption is assessed, which is important for systems working in near real-time scenarios. We also found only four papers using a user study for evaluation. We expected more papers using user studies, because many systems present novel visualization techniques and user studies can verify the strength and weakness of the application [HHN00, LYK*12, RHD*12].

We come to the conclusion that qualitative evaluation methods, especially use cases, are used frequently, since papers with respect to visual systems are application-driven and demand validation through description of fictitious scenarios. We also believe that there need more user evaluation, especially when testing the efficacy of systems for time-critical tasks.

		HHN00	GHT04	ABWS05	GS05	DKM* 07	GLYR07	PM07	ZM07	FHRH08	BS09	HHSW09	LBK09	SHM09	WRM*09	BSH*10	DGWC10	DNKS10	SKC10	SOM10	WLS*10	ACZ*11	AGCH11	APV11	BBF*11	BGAC11
Qualitativ	Case Study																	x								
	Usability Evaluation	x																								
	Use Case	x			x		x			x	x	x	x		x					x	x	x	x		x	
	Anecdotal Evaluation		x			x			x								x		x		x					
Quantitative	User Study															x								x		
	Algorithm Comparison							x								x					x					
	Ground Truth comparison										x					x				x						
	Algorithm Performance					x															x	x				
	Others													x												

		BSK*11	CLT*11	KBK11	MJR*11	MBB*11	TTSL11	KWJL11	WRK11	BBD*12	CLS*12	DWS*12	LYK*12	KRHW12	RHD*12	RKEAK12	SMT12	WWS12	RL12	AGC13	BTH*13	DWMR13	ITK13	KWD*13	KY13	WSWR13	WSJ*14
Qualitativ	Case Study					x					x																
	Usability Evaluation					x		x					x		x												
	Use Case	x	x	x	x		x		x		x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x
	Anecdotal Evaluation									x											x				x		
Quantitative	User Study														x					x							
	Algorithm Comparison												x														
	Ground Truth comparison					x							x		x							x					
	Algorithm Performance																				x						
	Others																										

Table 7: *Evaluation categorization of papers addressing visual text event detection. Qualitative use cases are the most common evaluation technique. Quantitative evaluations are seldom used. Interesting is that many systems dealing with potentially real-time data streams do not evaluate the performance or scalability of their algorithms.*

9. Conclusions

One thing we noticed was that data sources have dramatically changed from news to social media since 2010. Mainly due to the burst of social media, many research studies used text data streams generated out of Facebook or Twitter.

To our surprise, only one research study deals with event detection from customer feedback data, which are often generated from email, discussion thread, product review, or survey. Such data can be valuable resources to detect events that potentially influence brand and corporate reputation. Based on the analysis on detected events, one can also build strategies tailored to specific geographic regions and seasons. Thus, we point this area for future research.

Some data sources – like for instance discussion forums – are underused than others. Discussion forums are traditional methods to collect opinions from many people, but few research topics investigate data because they are asynchronous and slow to build up in nature. Despite these limitations, they also have a strength: archival history of some topics. Several discussion forums include years of textual conversation between multiple users on a single topic. For instance, this longitudinal conversation can be used to detect certain noticeable shifts in a specific user group's opinions on political issues over some months or years.

We found that a majority of visualization techniques are based on meta information, such as time or author, and are not visualizing text directly. Furthermore, most visualizations incorporate only basic text processing techniques such as *tf-idf*. Despite the lack of data processing methods and data mining methods used in the past decade, various research topics are slowly demanding more sophisticated data analytics techniques (e.g., topic modeling). We believe that this trend will continue.

More importantly, visualizations were primarily used as presentation, but had no interaction possible to steer the underlying data processing algorithm in order to further analyze data in a different angle. This limitation can prevent users from providing their insights back into the visualizations. We call for research to explore interactivity of visual analytics in various contexts.

Evaluations tend to rely on use cases for now. This trend seems to prohibit researchers from confirming the efficacy of different techniques in detecting events. In particular, many papers claim real-time analysis, but they do not evaluate time or memory consumption of their algorithm. We may need to generate example data with annotated ground truth for more in-depth evaluation and comparisons of algorithms. Data analysis communities should call for creation of such data through open-access competition, such as the well-known VAST Challenge. It seems also to be necessary to include users in the evaluations of systems. Users studies are only one way to go that is able to evaluate visualizations or automatic detection algorithms. Different possibilities to involve end users in the evaluation of systems are design studies or paired analytics. With theses techniques it is even possible to evaluate whole systems that might be to complex to evaluate with a single user study.

Almost all papers use different event definitions. Especially for news, *news criteria* (also known as news values) [GR65, HO01] can help find and develop new features for content-based feature detection. They are only mentioned once in our whole bulk of surveyed news analysis research papers [DNKS10]. According to [GR65], news criteria are: frequency, threshold, unambiguity, meaningfulness, consonance, unexpectedness, continuity, composition, reference to elite nations, reference to elite people, reference to persons, and reference to something negative. In [HO01] the authors show another set which modified Galtung's news values: the power elite, celebrity, entertainment, surprise, bad news, good news, magnitude, relevance, follow-up and newspaper agenda. There are several news criteria that extend the lists above and could be a guideline to improve the event detection algorithms and visual interfaces. In step 2 of Figure 5 news criteria are applied by a human (e.g. journalist, blogger, etc) in order to decide whether a detected event is newsworthy or not.

Events are sometimes endogenous within the given data [Bec11], which means that there is no real world event related in space and time. To overcome such problems, according to [Kep98] user's *selection criteria* could be taken into account. In step 3 of Figure 5 selection criteria are applied by a consumer (reader) in order to decide whether a broadcasted event is newsworthy or not. Especially in news analysis could those criteria of news creation and selection lead to new filters expressing user's intents.

In general, news and selection criteria could be merged into one concept we call *event values*. Event values are a concept including the text data producer's and user's perspectives. They could be implemented in the data analysis process by means of new features (feature engineering) and interactive elements, which comes along with the call for more visual analytics functionality. In (visual) text event analysis applications, this concept could help to improve and generalize the underlying event definition and the process of event detection. It could also enable different user groups to adapt the same application for their needs. Then, event criteria would help to understand which events are in the data and what is relevant to certain users. Anyway, further interdisciplinary research is needed on how a general concept can be developed and which news and selection criteria can be transferred to social media analysis [Jah12].

As a side note we did not find any paper within our selection judging the trustworthiness or credibility of the textual documents.

These lessons from our reviews shed light on various areas we need to fill in the next decade. Our review shows that we have an ill-defined concept of event, which may distract the focused effort from this community. On the other hand, we also show some interesting perspectives of news and selection criteria that can be used to detect events that are defined by users. We believe that this paper takes the first step towards clarifying this concept.

References

[ABWS05] ALBRECHT-BUEHLER C., WATSON B., SHAMMA D. A.: Visualizing live text streams using motion and temporal pooling. *Computer Graphics and Applications, IEEE 25*, 3 (2005), 52–59. 12

[ACZ*11] ALSAKRAN J., CHEN Y., ZHAO Y., YANG J., LUO D.: Streamit: Dynamic visualization and interactive exploration of text streams. In *Pacific Visualization Symposium (PacificVis), 2011 IEEE* (2011), IEEE, pp. 131–138. 12

[AGC13] ARCHAMBAULT D., GREENE D., CUNNINGHAM P.: Twittercrowds: Techniques for exploring topic and sentiment in microblogging data. *CoRR abs/1306.3839* (2013). 12

[AGCH11] ARCHAMBAULT D., GREENE D., CUNNINGHAM P., HURLEY N.: Themecrowds: Multiresolution summaries of twitter usage. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents* (2011), ACM, pp. 77–84. 12

[APL98] ALLAN J., PAPKA R., LAVRENKO V.: On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (1998), SIGIR '98, ACM, pp. 37–45. 1

[APV11] ADAMS B., PHUNG D., VENKATESH S.: Eventscapes: visualizing events over time with emotive facets. In *Proceedings of the 19th ACM international conference on Multimedia* (2011), ACM, pp. 1477–1480. 12

[BBD*12] BEST D. M., BRUCE J., DOWSON S., LOVE O., MCGRATH L.: Web-based visual analytics for social media. In *Proceedings of the Workshop on Social Media Visualization (SocMedVis) at ICWSM 2012* (June 2012), pp. 2–5. 12

[BBF*11] BERTINI E., BUCHMULLER J., FISCHER F., HUBER S., LINDEMEIER T., MAASS F., MANSMANN F., RAMM T., REGENSCHEIT M., ROHRDANTZ C., ET AL.: Visual analytics of terrorist activities related to epidemics. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (2011), IEEE, pp. 329–330. 12

[BC08] BERRY M. W., CASTELLANOS M.: *Survey of Text Mining II: Clustering, Classification, and Retrieval*, vol. 2. Springer, 2008. 2

[Bec11] BECKER H.: *Identification and Characterization of Events in Social Media.* PhD thesis, COLUMBIA UNIVERSITY, 2011. 2, 11

[BGAC11] BREW A., GREENE D., ARCHAMBAULT D., CUNNINGHAM P.: Deriving insights from national happiness indices. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops* (Washington, DC, USA, 2011), ICDMW '11, IEEE Computer Society, pp. 53–60. 7, 12

[BR12] BONTCHEVA K., ROUT D.: Making sense of social media streams through semantics: a survey. *Semantic Web* (2012). 2, 3

[BS09] BERENDT B., SUBASIC I.: Stories in time: a graph-based interface for news tracking and discovery. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 03* (2009), IEEE Computer Society, pp. 531–534. 12

[BSH*10] BERNSTEIN M. S., SUH B., HONG L., CHEN J., KAIRAM S., CHI E. H.: Eddi: interactive topic-based browsing of social status streams. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology* (2010), ACM, pp. 303–312. 12

[BSK*11] BRYAN K., SANTOS Y., KIM B., ET AL.: Twitter-reporter: Breaking news detection and visualization through the geo-tagged twitter network. In *Proceedings of the 26th International Conference on Computers and Their Applications (CATA-2011)* (2011). 12

[BTH*13] BOSCH H., THOM D., HEIMERL F., PUTTMANN E., KOCH S., KRUGER R., WORNER M., ERTL T.: Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *Visualization and Computer Graphics, IEEE Transactions on 19*, 12 (2013), 2022–2031. 8, 12

[CLS*12] CAO N., LIN Y.-R., SUN X., LAZER D., LIU S., QU H.: Whisper: Tracing the spatiotemporal process of information diffusion in real time. *IEEE Trans. Vis. Comput. Graph. 18*, 12 (2012), 2649–2658. 12

[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. *Visualization and Computer Graphics, IEEE Transactions on 17*, 12 (2011), 2412–2421. 12

[Con13] CONSTINE J.: Twitter user growth decelerating: +6% in q3 to 231.7 million now vs +10% in q1. *TechCrunch* (Oct. 2013). Accessed April 13, 2014. URL: http://techcrunch.com/2013/10/15/twitter-growth-decelerating/. 4

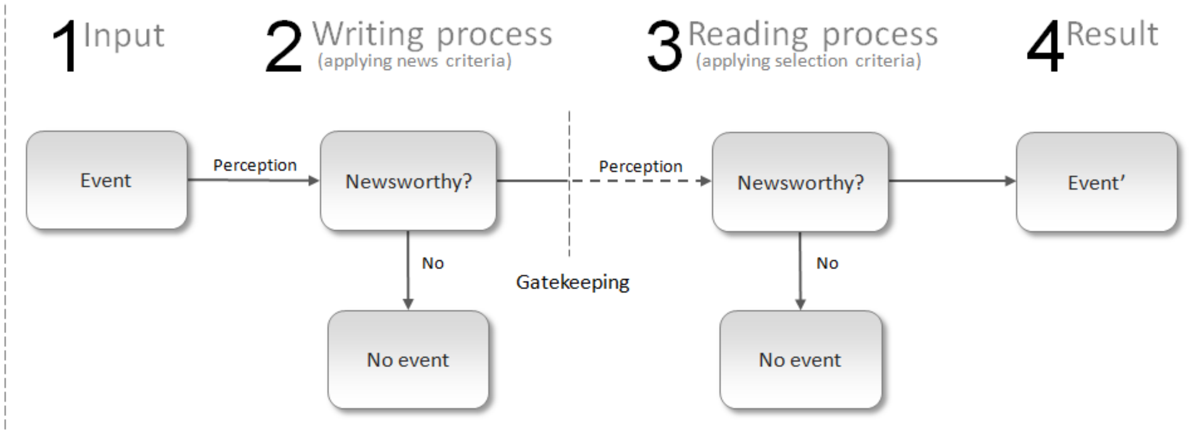


Figure 5: *A simplified process of news creation and consumption. During news creation (step 2) news criteria [GR65, HO01] are used to judge newsworthiness of an event by a human. During reception (step 3) a reader applies selection criteria [Kep98] to decide whether the broadcasted event is newsworthy enough to get attention. The process is appropriate for conventional news and could also be appropriate for social media, but not without further research [Jah12]. Note, social media have typically no gatekeeping process.*

[DGWC10] DORK M., GRUEN D., WILLIAMSON C., CARPENDALE S.: A visual backchannel for large-scale events. *Visualization and Computer Graphics, IEEE Transactions on 16*, 6 (2010), 1129–1138. 12

[DKM*07] DUBINKO M., KUMAR R., MAGNANI J., NOVAK J., RAGHAVAN P., TOMKINS A.: Visualizing tags over time. *ACM Transactions on the Web (TWEB) 1*, 2 (2007), 7. 12

[DNKS10] DIAKOPOULOS N., NAAMAN M., KIVRAN-SWAINE F.: Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology* (2010). 9, 10, 11, 12

[DWMR13] DOU W., WANG D. X., MA Z., RIBARSKY W.: Discover diamonds-in-the-rough using interactive visual analytics system: Tweets as a collective diary of the occupy movement. In *Seventh International AAAI Conference on Weblogs and Social Media* (2013). 12

[DWRZ12] DOU W., WANG X., RIBARSKY W., ZHOU M.: Event detection in social media data. In *IEEE VisWeek Workshop on Interactive Visual Text Analytics - Task Driven Analytics of Social Media Content* (2012). 1

[DWS*12] DOU W., WANG X., SKAU D., RIBARSKY W., ZHOU M. X.: Leadline: Interactive visual analysis of text data through event identification and exploration. In *IEEE Conference on Visual Analytics Science and Technology* (2012). 12

[EdR08] ELGERSMA E., DE RIJKE M.: Personal vs non-personal blogs: Initial classification experiments. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2008), SIGIR '08, ACM, pp. 723–724. 4

[FHRH08] FISHER D., HOFF A., ROBERTSON G., HURST M.: Narratives: A visualization to track narrative events as they develop. In *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on* (2008), IEEE, pp. 115–122. 12

[GHT04] GLANCE N., HURST M., TOMOKIYO T.: Blogpulse: Automated trend discovery for weblogs. In *WWW 2004 workshop on the weblogging ecosystem: Aggregation, analysis and dynamics* (2004), vol. 2004. 12

[GLYR07] GHONIEM M., LUO D., YANG J., RIBARSKY W.: Newslab: Exploratory broadcast news video analysis. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on* (2007), IEEE, pp. 123–130. 12

[GR65] GALTUNG J., RUGE M. H.: The structure of foreign news the presentation of the congo, cuba and cyprus crises in four norwegian newspapers. *Journal of peace research 2*, 1 (1965), 64–90. 2, 11, 13

[GS05] GLANCE N., SIEGLER M.: Deriving marketing intelligence from online discussion. In *In KDD* (2005), pp. 419–428. 7, 12

[HHN00] HAVRE S., HETZLER B., NOWELL L.: Themeriver: Visualizing theme changes over time. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on* (2000), IEEE, pp. 115–123. 10, 12

[HHSW09] HUBMANN-HAIDVOGEL A., SCHARL A., WEICHSELBRAUN A.: Multiple coordinated views for searching and navigating web content repositories. *Information Sciences 179*, 12 (2009), 1813–1821. 7, 12

[HNP05] HOTHO A., NÜRNBERGER A., PAASS G.: A brief survey of text mining. In *Ldv Forum* (2005), vol. 20, pp. 19–62. 2

[HO01] HARCUP T., O'NEILL D.: What is news? galtung and ruge revisited. *Journalism studies 2*, 2 (2001), 261–280. 2, 11, 13

[ITK13] ITOH M., TOYODA M., KITSUREGAWA M.: Visualizing time-varying topics via images and texts for inter-media analysis. In *IV* (2013), IEEE, pp. 568–576. 12

[Jah12] JAHN C.: *Selektion und Rezeption im Internet: Eine Metaanalyse zu Nachrichtenfaktoren im Online-Journalismus*. GRIN Verlag, 2012. 12, 13

[KAF*08] KEIM D., ANDRIENKO G., FEKETE J.-D., GÖRG C., KOHLHAMMER J., MELANÇON G.: *Visual analytics: Definition, process, and challenges*. Springer, 2008. 2

[KBK11] KRSTAJIC M., BERTINI E., KEIM D.: Cloudlines:

compact display of event episodes in multiple time-series. *Visualization and Computer Graphics, IEEE Transactions on 17*, 12 (2011), 2432–2439. 12

[Kep98] KEPPLINGER H. M.: Der nachrichtenwert der nachrichtenfaktoren. In *Wie die Medien die Welt erschaffen und wie die Menschen darin leben* (1998), pp. 19–38. 2, 12, 13

[KLS*05] KOH A., LIM A., SOON N. E., DETENBER B. H., CENITE M.: Ethics in blogging, Aug. 2005. Accessed April 13, 2014. URL: http://unpan1.un.org/intradoc/groups/public/documents/apcity/unpan026247.pdf. 4

[Koe13] KOETSIER J.: How twitter plans to make its 750M 'users' like its 250M real users. *VentureBeat* (Sept. 2013). Accessed April 13, 2014. URL: http://venturebeat.com/2013/09/16/how-twitter-plans-to-make-its-750m-users-like-its-250m-real-users/. 4

[KRHW12] KRSTAJIC M., ROHRDANTZ C., HUND M., WEILER A.: Getting there first: Real-time detection of real-world incidents on twitter. In *2nd IEEE Workshop on Interactive Visual Text Analytics "Task-Driven Analysis of Social Media" as part of the IEEE VisWeek 2012* (Seattle, Washington, USA, 2012). 12

[KWD*13] KRAFT T., WANG D. X., DELAWDER J., DOU W., LI Y., RIBARSKY W.: Less after-the-fact: Investigative visual analysis of events from streaming twitter. In *LDAV* (2013), Geveci B., Pfister H., Vishwanath V., (Eds.), IEEE, pp. 95–103. 12

[KWJL11] KRAKER P., WAGNER C., JEANQUARTIER F., LINDSTAEDT S.: On the way to a science intelligence: visualizing tel tweets for trend detection. *Towards Ubiquitous Learning* (2011), 220–232. 12

[KY13] KANEKO T., YANAI K.: Visual event mining from geo-tweet photos. In *ICME Workshops* (2013), IEEE, pp. 1–6. 12

[LBK09] LESKOVEC J., BACKSTROM L., KLEINBERG J.: Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (2009), ACM, pp. 497–506. 12

[LYK*12] LUO D., YANG J., KRSTAJIC M., RIBARSKY W., KEIM D. A.: Eventriver: Visually exploring text collections with temporal references. *Visualization and Computer Graphics, IEEE Transactions on 18*, 1 (2012), 93–105. 10, 12

[MBB*11] MARCUS A., BERNSTEIN M. S., BADAR O., KARGER D. R., MADDEN S., MILLER R. C.: Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (2011), ACM, pp. 227–236. 10, 12

[MJR*11] MACEACHREN A. M., JAISWAL A., ROBINSON A. C., PEZANOWSKI S., SAVELYEV A., MITRA P., ZHANG X., BLANFORD J.: Senseplace2: Geotwitter analytics support for situational awareness. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (2011), IEEE, pp. 181–190. 12

[PM07] PAN C.-C., MITRA P.: Femarepviz: Automatic extraction and geo-temporal visualization of fema national situation updates. In *IEEE VAST* (2007), IEEE, pp. 11–18. 12

[RHD*12] ROHRDANTZ C., HAO M. C., DAYAL U., HAUG L.-E., KEIM D. A.: Feature-based Visual Sentiment Analysis of Text Document Streams. *ACM Transactions on Intelligent Systems and Technology, Special Issue on Intelligent Visual Interfaces for Text Analysis 3*, 2 (2012), 26. 10, 12

[RKEAK12] ROHRDANTZ C., KRSTAJIC M., EL ASSADY M., KEIM D. A.: What's Going On? How Twitter and Online News Can Work in Synergy to Increase Situational Awareness. In *Published at the 2nd IEEE Workshop on Interactive Visual Text Analytics "Task-Driven Analysis of Social Media" as part of the IEEE VisWeek 2012, October 15th, 2012, Seattle, Washington, USA* (2012). 12

[RL12] RIOS M., LIN J.: Distilling massive amounts of data into simple visualizations: Twitter case studies. In *Proceedings of the Workshop on Social Media Visualization (SocMedVis) at ICWSM 2012* (June 2012), pp. 22–25. 12

[ROKF11] ROHRDANTZ C., OELKE D., KRSTAJIC M., FISCHER F.: Real-time visualization of streaming text data: Tasks and challenges. In *Workshop on Interactive Visual Text Analytics for Decision-Making at the IEEE VisWeek* (2011), vol. 201. 2

[ŠB10] ŠILIĆ A., BAŠIĆ B. D.: Visualization of text streams: a survey. In *Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2010, pp. 31–43. 2, 3

[sec13] Twitter registration document under united states securities and exchange commission, Oct. 2013. Accessed April 13, 2014. URL: http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm. 5

[SHM09] SAYYADI H., HURST M., MAYKOV A.: Event detection and tracking in social streams. In *Proceedings of International Conference on Weblogs and Social Media (ICWSM)* (2009). 12

[SK13] SCHRECK T., KEIM D.: Visual analysis of social media data. *Computer 46*, 5 (2013), 68–75. 2

[SKC10] SHAMMA D. A., KENNEDY L., CHURCHILL E. F.: Conversational shadows: Describing live media events using short messages. In *ICWSM* (2010). 12

[SMT12] SHIROI S., MISUE K., TANAKA J.: Chronoview: Visualization technique for many temporal data. In *Information Visualisation (IV), 2012 16th International Conference on* (2012), IEEE, pp. 112–117. 12

[SOM10] SAKAKI T., OKAZAKI M., MATSUO Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (2010), ACM, pp. 851–860. 7, 12

[TTSL11] TJONDRONEGORO D., TAO X., SASONGKO J., LAU C. H.: Multi-modal summarization of key events and top players in sports tournament videos. In *WACV* (2011), IEEE Computer Society, pp. 471–478. 12

[WB80] WHITNEY D. C., BECKER L. B.: The effects of wire news. *Mass communication review yearbook* (1980), 407. 4

[WC09] WU S.-Y., CHEN Y.-L.: Discovering hybrid temporal patterns from sequences consisting of point-and interval-based events. *Data & Knowledge Engineering 68*, 11 (2009), 1309–1330. 7

[WLS*10] WEI F., LIU S., SONG Y., PAN S., ZHOU M. X., QIAN W., SHI L., TAN L., ZHANG Q.: Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (2010), ACM, pp. 153–162. 12

[WRK11] WANNER F., RAMM T., KEIM D. A.: ForAVis: explorative user forum analysis. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics* (2011), WIMS '11. 12

[WRM*09] WANNER F., ROHRDANTZ C., MANSMANN F., OELKE D., KEIM D. A.: Visual sentiment analysis of rss news feeds featuring the us presidential election in 2008. In *Workshop* *on Visual Interfaces to the Social and the Semantic Web (VISSW)* (2009). 12

[WSJ*14] WANNER F., SCHRECK T., JENTNER W., SHARALIEVA L., KEIM D. A.: Relating interesting quantitative time series patterns with text events and text features. In *SPIE 2014 Conference on Visualization and Data Analysis (VDA 2014), Best Paper Award* (2014). 7, 12

[WSWR13] WEILER A., SCHOLL M. H., WANNER F., ROHRDANTZ C.: Event identification for local areas using social media streaming data. In *Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks* (2013), ACM, pp. 1–6. 12

[WWS12] WANNER F., WEILER A., SCHRECK T.: Topic Tracker: Shape-based visualization for trend and sentiment tracking in twitter. In *2nd IEEE Workshop on Interactive Visual Text Analytics. Task-Driven Analysis of Social Media. IEEE VisWeek, Seattle, WA, USA* (2012). 12

[YPC98] YANG Y., PIERCE T., CARBONELL J.: A study of retrospective and on-line event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (1998), SIGIR '98, ACM, pp. 28–36. 1

[ZM07] ZHAO Q., MITRA P.: Event detection and visualization for social text streams. *Proceedings of ICWSM'2007* (2007), 26–28. 7, 12

Biography

Franz Wanner is PhD student in computer science. He is working for years in the Information Visualization and Data Analysis Research Group at the University of Konstanz. His main research interests include visual text event detection, visual analytics of heterogeneous datasets in conjunction with text and document visualizations.

Andreas Stoffel received his PhD degree in computer science in 2013. He is currently working in the Information Visualization and Data Analysis Research Group at the University of Konstanz. His main research interests include visual analytics, document visualization and automatic document analysis methods.

Dominik Jäckle is PhD student at the Data Analysis and Visualization Group at the University of Konstanz. His main research interests include the development and application of new visualization techniques and algorithms for the exploration of vast amounts of data.

Bum Chul Kwon received his PhD degree specializing in information visualization and human computer interaction at Purdue University in 2013. He is currently working in the Information Visualization and Data Analysis Research Group at the University of Konstanz. His main research interests include information visualization, visual analytics and human-based computation methods.

Andreas Weiler is PhD student at the Database and Information Systems Research Group at the University of Konstanz. His main research focuses on processing, analyzing, and visualizing of Social Media Data Streams for Event Identification and Tracking.

Daniel Keim is a full professor and the head of the Information Visualization and Data Analysis Research Group in the University of Konstanz's Computer Science Department. Keim received a habilitation in computer science from the University of Munich. He has been program cochair of the IEEE Information Visualization Conference, the IEEE Conference on Visual Analytics Science and Technology (VAST), and the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. He's on the steering committees of IEEE VAST and the Eurographics Conference on Visualization.