Online Social Networks Event Detection: A survey

Mário Cordeiro[1] and João Gama[2]

[1] University of Porto
pro11001@fe.up.pt,
[2] University of Porto, INESC TEC Laboratory of Artificial Intelligence and Decision Support
jgama@fep.up.pt

Abstract. Today online social network services are challenging state-of-the-art social media mining algorithms and techniques due to its real-time nature, scale and amount of unstructured data generated. The continuous interactions between online social network participants generate streams of unbounded text content and evolutionary network structures within the social streams that make classical text mining and network analysis techniques obsolete and not suitable to deal with such new challenges. Performing event detection on online social networks is no exception, state-of-the-art algorithms rely on text mining techniques applied to pre-known datasets that are being processed with no restrictions on the computational complexity and required execution time per document analysis. Moreover, network analysis algorithms used to extract knowledge from users relations and interactions were not designed to handle evolutionary networks of such order of magnitude in terms of the number of nodes and edges. This specific problem of event detection becomes even more serious due to the real-time nature of online social networks. New or unforeseen events need to be identified and tracked on a real-time basis providing accurate results as quick as possible. It makes no sense to have an algorithm that provides detected event results a few hours after being announced by traditional newswire.

Keywords: Event Detection, Social Networks

1 Introduction

Today, online social networking services like Twitter [102], Facebook [99], Google+ [100], LinkedIn [101], among others, play an important role in the dissemination of information on a real-time basis [91].

Recent observation proves that some events and news emerge and spread first using those media channels rather than other traditional media like the online news sites, blogs or even television and radio breaking news [88, 50]. Natural disasters, celebrity news, products announcements, or mainstream event coverage show that people increasingly make use of those tools to be informed, discuss and exchange information [38]. Empirical studies [88, 50] show that the online social

networking service Twitter is often the first medium to break important natural events such as earthquakes often in a matter of seconds after they occur. Being Twitter the "what's-happening-right-now" tool [91] and given the nature of it's data — an real-time flow of text messages (tweets) coming from very different sources covering varied kinds of subjects in distinct languages and locations — makes the Twitter public stream an example of an interesting source of data for "real time" event detection based on text mining techniques. Note that "real time" means that events need to be discovered as early as possible after they start unraveling in the online social networking service stream. Such information about emerging events can be immensely valuable if it is discovered timely and made available.

When some broad major event happens, three factors are the main contributors to the rapidly spread of information materialized in exchanged messages between users of an online social network service. i) the ubiquity nature of today's social network services, that are available nowadays by any internet connected device like a personal computer or a smartphone; ii) the ease of use and agility of entering or forward information is also a key factor that lead some messages to be spread very fast on the network and go viral [40]; and iii) the lifespan of the messages is also an interesting feature of those online social network services. Posted messages tend to be exchanged, forwarded or commented following a time decay pattern, meaning that the information they contain has the importance peak when it is posted in the following hours or days [51]. This statement is coherent with the Barabasi [11] conclusion that the timing of many human activities, ranging from communication to entertainment and work patterns, follow non-Poisson statistics, characterized by bursts of rapidly occurring events separated by long periods of inactivity.

With the purpose of correlating the occurrence of events in the real world and the resulting activity in online social networks, Zhao et al. [108] and Sakaki et al. [88] introduced the concept of "social sensors" where the social text streams are seen as sensors of the real world. The assumption made is that each online social user (i.e.: a Twitter, Facebook, Google+ user) is regarded as a sensor and each message (i.e.: tweet, post, etc.) as sensory information. Zhao et al. [108] pointed two major substantial differences of the social text stream data over general text stream data: i) social text stream data contains rich social connections (between the information senders/authors and recipients/reviewers) and temporal attributes of each text piece; and ii) the content of text piece in the social text stream data is more context sensitive. Sakaki et al. [88] went beyond in its concept of "social sensors" by introducing an analogy to a physical sensor network. Some common characteristics of those "virtual" social sensors in comparison with real physical sensors are: i) some sensors are very active, others are not — the activity of each user is different as some users post more messages than others; ii) a sensor could be inoperable or malfunctioning sometimes — this means that a user can be offline at a given time ie.: sleeping, on vacation; or even offline (without internet connection); and iii) very noisy compared to

ordinal physical sensors — the output of the sensor is not normalized, there are many users that are posting messages that can be considered as spammers.

1.1 Event detection overview

The event detection problem is not a new research topic, Yang et al. [105] in 1998, investigated the use and extension of text retrieval and clustering techniques for event detection. The main task was to detect novel events from a temporally-ordered stream of news stories automatically. In an evaluation of the system using manually labeled events were obtained values for the F-score[3] of 82% in retrospective detection and 42% in on-line detection. Despite the fact of the size of the corpus with 15,836 documents used in the evaluation, the system performed quite well and showed that basic techniques such as document clustering can be highly effective to perform event detection.

Two years later Allan et al. [6] evaluated the UMASS reference system [4] in three of the five Topic Detection and Tracking (TDT) tasks: i) detection; ii) first story detection; and iii) and story link detection [3]. The core of this system used a vector model for representing stories, each story as a vector in term-space, and terms (or features) of each vector were single words, reduced to their root form by a dictionary-based stemmer. The study concluded that the results were acceptable for the three evaluated tasks but not as high quality as authors expected. Allan et al. [5] showed that performing first story detection based upon tracking technology has poor performance and to achieve high-quality first story detection the tracking effectiveness should be improved to a level that experiments showed not to be possible. Therefore Allan et al. [5] concluded that first story detection is either impossible or requires substantially different approaches.

Despite the fact that in the following 10 years, the period between years 2000 and 2010, the event detection problem was a relatively active research topic, it was in the latest 8 years, coinciding with the advent and massification of the online social networks phenomena and big data era that the problem gained more interest from the research community. Just targeting event detection specifically in the online social network service Twitter, Petrovic [73] pointed out and compared major scientific contribution from Hu et al. [39], Jurgens and Stevens [43], Sankaranarayanan et al. [89], Sakaki et al. [88], Popescu and Pennacchiotti [79], Cataldi et al. [20], Mathioudakis and Koudas [62], Phuvipadawat and Murata [76], Becker et al. [14], Weng et al. [98], Cordeiro [23], Li et al. [56], Li et al. [55], Agarwal et al. [1] and Ozdikis et al. [68]. This fact by itself is explanatory on the interest and relevance of the research topic. None of this listed publications managed to solve the problem of event detection in online social networks completely. Some of them assumed to solve the problem partially by defining constraints or limiting the scope of the problem. One year later, Atefeh and Khreich [9] published a survey that classifies the major techniques for Twitter

[3] `http://en.wikipedia.org/wiki/F1_score`

event detection according to the event type (specified or unspecified events), detection method (supervised or unsupervised learning), and detection task (new event detection or retrospective event detection). Due the fact that the research conducted by Petrovic [73] work was primarily focused on solving online new event detection of unspecified events using unsupervised methods, it did not compare his work with other references to advancements in other specific areas of the event detection. Atefeh and Khreich [9] survey considers work described by Petrovic [73] [89, 76, 74, 14, 98, 23, 79, 88] and additional advancements like Long et al. [58], Popescu et al. [80], Benson et al. [15], Lee and Sumiya [52], Becker et al. [12], Massoudi et al. [61], Metzler et al. [63] and Gu et al. [35]. This survey also discusses the common used features used in event detection tasks for each one of the listed methods. Imran et al. [41] in a survey under the subject of communication channels during mass convergence and emergency events, gave an overview of the challenges and existing computational methods to process social media messages that may lead to an effective response in mass emergency scenarios. This survey, not being specifically devoted to event detection, includes a full chapter where Retrospective and Online New Event Detection types are addressed.

Most of the techniques described by Petrovic [73], Atefeh and Khreich [9] and Imran et al. [41] lack evaluation or are evaluated empirically. Measuring the accuracy and performance of an event detection methods is hampered by the lack of standard corpora and results leading some authors to create and make publicly available their own datasets with events being annotated manually [73]. In other cases evaluation is made with some automation by comparing directly to a reference system as a baseline [98] by generating a list of detected event that serves as ground truth. The need for public benchmarks to evaluate the performance of different detection approaches and various features was also highlighted by Atefeh and Khreich [9].

1.2 Problem statement

Most of the described approaches to solving the event detection in text streams are not real-time and use batch algorithms. The good results obtained by reference systems used in the evaluation of the TDT task were obtained from reduced corpus datasets. Latter studies proved that they do not scale to larger amounts of data [72], in fact they were not even designed do deal with text streams. The performance, effectiveness and robustness of those reference systems was acceptable under the specified evaluation conditions at that time. Due the characteristics of today's online social network services data, unbounded massive unstructured text streams, these systems are nowadays considered as being obsolete. Apart from not being designed to handle big amounts of data, the data in online social network services is also dynamic, messages are arriving at high data rates, requiring the adaption of the computing models to process documents as they arrive. Finally today computation time is an issue, in most cases when using this kind of systems it is preferable to have an immediate and approximated solution rather than waiting too much for an exact solution [10].

Online social network text streams seem to be the ideal source to perform real-time event detection applying data mining techniques [88, 98, 74]. The main benefits of using those sources of data are their real-time or near real-time data availability, contextualization of the messages with additional information (temporal, geospatial, entity, etc. referenced in messages), possible data segmentation at several levels (by communities of user, by regions, by topic, etc.), access to static relations between users (friends, followers, user groups), possibility to build dynamic user relations built from exchanged messages flows, among others.

To perform data mining, every previously mentioned advantage of the online social network text data source reveals, in fact, to have a significant drawback and shortcoming. Performing real-time event detection using online social network services requires dealing and mining massive unstructured text data streams with messages arriving at high data rates. Given this, the approach to deal with this specific problem involves providing solutions that are able to mine continuous, high-volume, open-ended data streams as they arrive [17, 82]. Because text data source is not disjointed from the online social network topological properties, it is expected that information retrieved using metrics of networks analysis (nodes, connections and relations, distributions, clusters, and communities) could improve the quality of the solution of the algorithm. In Table 3, for each one of the techniques, is included the collection, corpus size, and temporal scope of the dataset used in the evaluation.

1.3 Scope and Organization

It makes no sense to talk about an event detection system without first specifying an defining exactly what is an event. Sect. 2 introduces the concepts of story, event and topic. Sect. 3 defines, under topic detection and tracking task, introduces the origins of event detection as Information Retrieval problem. New Event Detection (NED) and Retrospective Event Detection (RED) tasks are described in Sect. 3.1. Sect. 3.2 describes the differences of systems designed to detected specified and unspecified events. Pivot techniques are presented in Sect. 3.3. Sect. 4 presents a taxonomy of event detection systems. The taxonomy was made taking into account the type of event that the system tries to detect (specified or unspecified event), and the type of the detection (unsupervised, supervised or hybrid detection). An overview of the common detection methods is presented in Sect. 4.3. Sect. 5 includes a list of the datasets, their respective size, and temporal scope used to evaluate each one of the event detection techniques. Finally Sect. 6 presents the conclusions, future and trends of event detection systems.

2 Event Definition

Fiscus and Doddington [30] in the scope of the Topic Detection and Tracking project gave the following definitions of story, event and topic:

story is *"a topically cohesive segment of news that includes two or more declarative independent clauses about a single event."*;
event is *"something that happens at some specific time and place along with all necessary preconditions and unavoidable consequences"*;
topic is *"a seminal event or activity, along with all directly related events and activities."*.

Sakaki et al. [88] defines an event as an arbitrary classification of a space/time region that might have actively participating agents, passive factors, products, and a location in space/time like is being defined in the event ontology by Raimond and Abdallah [81]. The target events in this work are broad events that are visible through messages, posts, or status updates of active users in Twitter online social network service. These events have several properties: i) they are of large scale because many users experience the event, ii) they particularly influence people's daily life, being that the main reason why users are induced to mention it, and iii) they have both spatial and temporal regions, topically the importance of an event is correlated with the distance users have between themselves and the event and with the spent time since the occurrence.

The Linguistic Data Consortium [57] defines the broad topic types denoting the category where an event falls into. As defined by the TDT5 [93] there are the following broad topic type categories: i) Elections; ii) Scandals/Hearings; iii) Legal/Criminal Cases; iv) Natural Disasters; v) Accidents; vi) Acts of Violence or War; vii) Science and Discovery News; viii) Financial News; ix) New Laws; x) Sports News; xi) Political and Diplomatic Meetings; xii) Celebrity and Human Interest News; and xiii) Miscellaneous News.

3 Earlier Event Detection and Discovery

The Topic Detection and Tracking project was started with the objective to improve technologies related to event-based information organization in 1998, see [3]. The project consisted of five distinct tasks: i) segmentation; ii) tracking; iii) detection; iv) first story detection; and v) linking. From the previous list of tasks the tracking, detection, and first story detection are the ones that are relevant for event detection.

tracking: the tracking task detect stories that discuss a previously known target topic. This task is very closely linked to the first story detection. A tracking system can be used to solve a first story detection by finding other on-topic stories in the rest of the corpus. A nearest-neighbour based first story detection system could be used to solve tracking;
detection: the detection task is concerned with the detection of new, previously unseen topics. This task is often also called on-line clustering, every newly received story is assigned to an existing cluster or to a new cluster depending if there is a new story of not;
first story detection: the first story detection is considered the most difficult of the five topic detection and tracking tasks [5]. The aim of the task is to

detect the very first story to discuss a previously unknown event. The first story detection can be considered a special case of detection by deciding when to start a new cluster in the on-line clustering problem.

Event detection on social media streams requires significantly different approaches than the ones used for traditional media. Social media data arrives at larger volumes and speed than traditional media. Moreover, most social media data is composed of short, noisy and unstructured content requiring significantly different techniques to solve similar machine learning or information retrieval problems [5]. Taking into account these considerations, Sect. 3.3 presents an overview of Document-Pivot and Feature-Pivot event detection techniques applied to traditional medial [98]. Document-pivot methods detect events by clustering documents based on the semantics distance between documents [105], feature-pivot methods studies the distributions of words and discovers events by grouping words together [46].

3.1 Detection Task

The task of discovering the "first story on a topic of interest" by continuously monitoring document streams is known in the literature as new event detection, first-story detection or novelty detection. Makkonen et al. [59] described first-story detection or novelty detection as an example of "query-less information retrieval" where events can be detected with no prior information available on a topic of interest. Events are evaluated using a binary decision on whether a document reports a new topic that has not been reported previously, or if should be merged with an existent event [103]. Depending on how data is processed, two categories of Event Detection systems were identified [105, 7].

Online New Event Detection (NED) Online New Event Detection refers to the task of identifying events from live streams of documents in real-time. Most new and retrospective event detection techniques rely on the use of well know clustering-based algorithms [16, 2]. Typically new event detection involves the continuous monitoring of Media feeds for discovering events in near real time, hereupon scenarios where the detection of real-world events like breaking news, natural disasters or other of general interest. Events are unknown apriori and in most cases use unspecified event detection. When monitoring specific NED like natural disasters or celebrities related, where specific apriori known information about the event can be used. In these cases, NED is performed using specified event detection.

Retrospective Event Detection (RED) Retrospective Event Detection refers to the process of identifying previously unidentified events from accumulated historical collections or documents that have arrived in the past. In Retrospective Event Detection, most methods are based on the retrieval of event relevant documents by performing queries over a collection of documents or by performing

TF-IDF analysis on the document corpus. Both techniques assume that event relevant documents contain the query terms. A variation of the previous approach is the use of query expansion techniques, meaning that some messages relevant to a specific event do not contain explicit event related information, but with the use of enhanced queries messages related to the event can be retrieved.

Table 1. Type, technique, detection method and detection task for each one of the references. Column Application refers to the target application of the work: a) Detecting General Interest Events; b) Identification of Novel Topics in Blogs; c) Detecting Controversial Events from Twitter; d) Calendar of Significant Events; e) Geo-Social Event Detection system; f) Detection of Natural Disaster Events; g) Query-based Event Retrieval; h) Query-based Structured Event Retrieval; i) Crime and Disaster related Events; j) Detection of Breaking News; k) Emergent topics; l) Trend Detection; m) Crisis-related Sub-event Detection; n) Event Photo Identification; o) Creating event-specific queries for Twitter

	Type of Event		Pivot Technique		Detection Method		Detection Task		Application
	Specified	Unspecified	Document	Feature	Supervised	Unsupervised	NED	RED	
Hu et al. [39]	x		x			x	x		a)
Jurgens and Stevens [43]	x		x			x	x		b)
Popescu and Pennacchiotti [79]	x			x	x		x		c)
Popescu et al. [80]	x			x	x		x		c)
Benson et al. [15]	x			x	x			x	d)
Lee and Sumiya [52]	x		x			x	x		e)
Sakaki et al. [88]	x			x	x		x		f)
Becker et al. [12]	x		x		x			x	g)
Becker et al. [13]	x			x				x	g)
Massoudi et al. [61]	x		x		x			x	g)
Metzler et al. [63]	x		x			x		x	h)
Gu et al. [35]	x		x			x		x	h)
Li et al. [56]	x			x	x		x		i)
Ozdikis et al. [68]	x		x			x	x		a)
Sankaranarayanan et al. [89]		x		x	x	x	x		j)
Cataldi et al. [20]		x	x		x	x	x		k)
Mathioudakis and Koudas [62]		x	x			x	x		l)
Phuvipadawat and Murata [76]		x	x			x	x		j)
Petrovic et al. [74]		x		x		x	x		a)
Becker et al. [14]		x		x	x	x	x		a)
Long et al. [58]		x		x		x	x		a)
Weng et al. [98]		x		x		x	x		a)
Cordeiro [23]		x		x		x	x		a)
Li et al. [55]		x		x		x	x		a)
Agarwal et al. [1]		x	x			x	x		a)
Sayyadi et al. [90]		x	x			x	x		a)
Zhao et al. [108]		x		x		x	x		a)
Pohl et al. [78]		x	x			x	x		m)
Chen and Roy [22]		x		x		x		x	n)
Ritter et al. [85]		x		x	x	x	x		d)
Robinson et al. [86]	x			x		x	x		f)
Corley et al. [24]		x	x			x	x		a)
Tanev et al. [94]	x		x			x	x		o)
Dou et al. [25]		x	x			x	x		a)

Table 2: Event detection approach

	Approach	Event Types	Scalable	Real-time	Query type	Statio-temporal	Sub-events
Hu et al. [39]	Online clustering of query profiles	open domain	yes	no	open	no	no
Jurgens and Stevens [43]	Temporal Random Indexing	open domain	yes	no	keywords	no	no
Popescu and Pennacchiotti [79]	Regression machine learning models based on Gradient Boosted Decision Trees	Controversial Events	no	no	open	no	no
Popescu et al. [80]	Regression machine learning models based on Gradient Boosted Decision Trees	Controversial Events	no	no	open	no	no
Benson et al. [15]	Factor Graph Model and Conditional Random Field	Concerts in New York City	no	no	keywords	yes	no
Lee and Sumiya [52]	K-means clustering method for detecting ROI, measuring statistical variations of a set of geo-tags	Local events such as local festivals	no	no	open	yes	no
Sakaki et al. [88]	support vector machine (SVM)	natural disaster event	yes	yes	keywords	yes	no
Becker et al. [12]	rule-based classifier	Planned Event	no	no	keywords	no	no
Becker et al. [13]	Precision / Recall Oriented Strategies	Planned Event	no	no	keywords	no	no
Massoudi et al. [61]	Query Expansion using the top k terms	Topic of interest	no	no	keywords	no	no
Metzler et al. [63]	Temporal Query Expansion based on temporal co-occurrence of terms	Topic of interest	no	no	keywords	no	no
Gu et al. [35]	Hierarquical clustering	Topic of interest	no	no	keywords	no	yes?

Table 2: Event detection approach

	Approach	Event Types	Scalable	Real-time	Query type	Statio-temporal	Sub-events
Li et al. [56]	Classification	Crime and Disaster related Events	no	no	spatial / temporal / keywords	yes	no
Ozdikis et al. [68]	Semantic Expansion of Hashtags via agglomerative clustering	open domain	no	no	spatial / keywords / users	no	no
Sankara-narayanan et al. [89]	Tweet Naive Bayes classifier and weighted term vector based online clustering	Breaking-News	yes	no	keywords	yes?	no
Cataldi et al. [20]	Keyword-based topic graph	Breaking-News	no	yes	keywords	no	no
Mathioudakis and Koudas [62]	Context extraction algorithms (PCA, SVD) and Keyword Co-Occurence Grouping	Breaking-News / Topic of interest	yes	yes	open	no	no
Phu-vipadawat and Murata [76]	Similarity based grouping via TF-IDF	Breaking-News	no	no	keywords	no	no
Petrovic et al. [74]	Detection of Events via Locally Sensitive Hashing	open domain	yes	yes	open	no	no
Becker et al. [14]	incremental, online clustering / classification via support vector machine (SVM)	open domain	yes	no	open	no	no
Long et al. [58]	top-down hierarchical divisive clustering on a co-occurence graph	open domain	no	no	open	no	no

Table 2: Event detection approach

	Approach	Event Types	Scalable	Real-time	Query type	Statio-temporal	Sub-events
Weng et al. [98]	Clustering of Wavelet-based Signals via graph partitioning	open domain	no	no	keywords	no	no
Cordeiro [23]	Wavelet-based Signals and Latent Dirichlet Allocation	open domain	yes	yes	open	no	no
Li et al. [55]	Symmetric Conditional Probability (SCP) for n-grams, bursty detection using binomial distribution, Clustering by k-Nearest Neighbor Graph	open domain	yes	no	open	no	no
Agarwal et al. [1]	Clustering in a Correlated Keyword Graph	open domain	yes	yes	open	no	no
Sayyadi et al. [90]	Community Detection on a Keyword Graph	open domain	no	no	open	no	no
Zhao et al. [108]	Content-Based Clustering where word in the text piece is quantified as the TF.IDF, adaptive time series, information flomodeling	open domain	no	no	open	no	no
Pohl et al. [78]	Two-phase clustering: 1. calculation of term-based centroids using geo-referenced data; 2. Assignment of best fitting data points using cosine distance measure	Crisis-related sub-event	no	no	Geo-Referenced Data	yes	yes
Chen and Roy [22]	Discrete Wavelet Transform (DWT), density-based clustering (DBSCAN)	periodic events / aperiodic events	yes	no	keywords	yes	no

Table 2: Event detection approach

	Approach	Event Types	Scalable	Real-time	Query type	Statio-temporal	Sub-events
Ritter et al. [85]	Named Entity Segmentation, Conditional Random Fields for learning and inference events, latent variable models to categorize events (LinkLDA)	open domain	no	no	keywords	no	no
Robinson et al. [86]	burst detector using binomial model	natural disaster event	no	yes	keywords	yes	no
Corley et al. [24]	Detection of Signal Consistency from Social Sensors, Topic Clustering via Pearson correlation coefficient, Autoregressive Integrated Moving Average	open domain	no	no	keywords	no	no
Tanev et al. [94]	query expansion methods	open domain	no	no	keywords	no	no
Dou et al. [25]	Topical themes using Latent Dirichlet Allocation (LDA), early event detection using cumulative sum control chart (CUSUM)	open domain	no	no	keywords	yes	no

3.2 Type of Event

Event detection can be classified into specified or unspecified event detection techniques [9, 29]. By using specific pre-known information and features about an event, traditional information retrieval and extraction techniques can be adapted to perform specified event detection (i.e.: filtering, query generation and expansion, clustering, and information aggregation). When no prior information or features are available about the event or even if we don't know a clue about the kind of event we want to detect, most traditional information retrieval and extraction techniques are useless. Unspecified event detection techniques address this issue on the basis that temporal signals constructed via document analysis can detect real work events. Monitoring bursts or trends in document streams, grouping features with identical trends, and classifying events into different categories are among some of the used tasks to perform unspecified event detection.

3.3 Pivot Techniques

Both Document-Pivot and Feature-Pivot techniques are being used in event detection applied to traditional media. The following sections describe how each of them works and how it is being used.

Document-Pivot Techniques Document-pivot techniques try to detect events by clustering documents using their textual similarity, these techniques consider all documents to be relevant and assume that each of them contain events of interest [5]. The noisy characteristics of social networks, where relevant events are buried by in large amount of noisy data [95], allied with scale and speed processing restrictions [5] make document-pivot techniques not suitable to perform event detection in social media data. Nevertheless, because they were the primordial steps to modern event detection systems, they will be briefly presented here.

The main goal of the TDT research initiative was to provide core technology and tools that by monitoring multiple sources of traditional media are able to keep users updated about news and developments. A particular event detection goal was to discover new or previously unidentified events were each event refers to a specific thing that happens at a specific time and place [7]. Yang et al. [105, 106] described the three traditional event detection major phases as data prepossessing, data representation, and data organisation or clustering. Filtering out stop-words and applying words stemming and tokenization techniques are some of the steps done in the data prepossessing phase. Term vectors of bag of words are common use traditional data representations techniques used in event detection. Entries are non-zero if the corresponding term appear in the document and zero otherwise. Classical term frequency-inverse document frequency (tf-idf) is used to evaluate how important a word is in a corpus and also to retrieve the list of documents where the word is mentioned. This rudimentary event detection approach does not solve the problem, the term vector model size can grow indefinitely depending on the size of the corpus. Temporal order,

the semantics and syntactic features of the of words are discarded. Although this model can find similarities of documents it may not capture the similarity or dissimilarity of related or unrelated events. Exploring other data representation techniques such as semantical and contextual features was also done by Allan et al. [5, 6] where they presented an upper bound for full-text similarity. Alternative data representations such as the named entity vector [49] attempt to extract information answering the question who, what, when, and where [64]. Mixed models using term and named entity vectors were also proposed [49, 104]. Probabilistic representations including language models were applied by Lee et al. [53] and Ping Li et al. [77] proposed a probabilistic framework McRank that incorporates advanced probabilistic learning models. Traditional metrics like the Euclidean distance, Pearson's correlation coefficient, and cosine similarity were also used to measure the similarity between events. Other similarity measures like the Hellinger distance [19] and the clustering index [42] were also used.

Feature-Pivot Techniques Modeling an event in text streams as a bursty activity, with certain features rising sharply in frequency as the event emerges is the common approach for Feature-Pivot techniques. Kleinberg [46] show that events may be represented by a number of keywords showing bursts in appearance counts. Moreover, in his work, he developed a formal approach for modeling bursts in a way that they can be robustly and efficiently identified, provide an organizational framework for analyzing the underlying content. Several systems to detect emerging trends in textual data (Emerging Trend Detection systems) were described by Kontostathis et al. Kontostathis et al. [48]. The main goal of a trend detection task over textual data is to identify topic areas that were previously unseen or rapidly growing in importance within the corpus. Kontostathis et al. Kontostathis et al. [48] described, for each system, the components (including linguistic and statistical features), learning algorithms, training and test set generation, visualization, and evaluation. Bursty event detection has been also an active topic in recent years with contributions from Fung et al. [32], He et al. [37], He et al. [36], Wang et al. [97] and Goorha and Ungar [34].

Kleinberg [46] approach is based on modeling the stream using an infinite-state automaton, in which bursts appear naturally as state transitions. The output of the algorithm yields a nested representation of the set of bursts that imposes a hierarchical structure on the overall stream computed in a highly efficient way. Fung et al. [32] proposed a parameter free probabilistic approach, called feature-pivot clustering, that detect a set of bursty features for a burst event detected in a sequence of chronologically ordered documents. The feature-pivot clustering modeled word appearance as a binomial distribution, identified the bursty words according to a heuristic-based threshold, and grouped bursty features to find bursty events. Spectral analysis techniques using the discrete Fourier transformation (DFT) were used by He et al. [37] to categorise features for different event characteristics, i.e.: important or not important, and periodic or aperiodic events. Passing from the time domain to the frequency domain, using the DFT, allows the identification of bursts in signals by monitoring the

corresponding spike in the frequency domain. Aware that the DFT cannot identify the period of a burst event, He et al. [36] improved their previous works with Gaussian mixture models to identify the feature bursts and their associated periods. Important work in the domain of multiple coordinated text streams was done by Wang et al. [97]. They proposed a general probabilistic algorithm which can effectively discover correlated bursty patterns and their bursty periods across text streams even if the streams have completely different vocabularies (e.g., English vs. Chinese). An online approach for detecting events in news streams was presented by Snowsill et al. [92], this technique is based on statistical significant tests of n-gram word frequency within a time frame. The online detection was achieved, by reducing time and space constraints, when an incremental suffix tree data structure was applied. Social and mainstream media system monitoring tools are also available in Goorha and Ungar [34]. These tools are focused on the user discovery, query and visualisation process for lists of emerging trends previously collected by using some of the algorithms described in this section.

Like the document-pivot techniques, feature-pivot techniques do not deal well with noise resulting in poor event detection performance. Moreover, not all bursts are relevant events of interest, other ones may be missed due the fact that they happen without explicit burst occurrences.

4 Event Detection Taxonomy

The event detection taxonomy is presented in Table 1. A description of each one of the techniques is included in the present section. A division of each one of the techniques was made by taking into account the type of events they were designed (.i.e.: Specified or Unspecified Event Detection) and the respective detection method type (i.e.: supervised, unsupervised or hybrid in case it is a combination of both). A resume of the Approaches used in each one of the techniques is presented in Table 2.

4.1 Specified Event Detection

Specified event detection systems using either unsupervised, supervised and hybrid detection techniques are being described in this section.

Unsupervised Detection Hu et al. [39] proposed event detection of common user interests from huge volume of user-generated content by assuming that the degree of interest from common users in events is evidenced by a significant surge of event-related queries issued to search for documents (e.g., news articles, blog posts). Defining query profile as a set of documents matching a query at a given time and single streams of query profiles as the integration of a query profile and respective documents, events are therefore detected by applying incremental clustering to the stream of query profiles. A temporal query profile is a set of published documents at a given time matching the queries formulated by users at the same time. Based on the observations regarding the number of documents

retrieved, authors were able to associate a query profile to the occurrence of a specific event, correlate different query profiles in the context of the same event and establish their duration and evolution. Event detection uses a simple online clustering algorithm consisting of modules: event-related query identification, event assignment, and event archive.

Lee and Sumiya [52] developed a geo-social event detection system, which attempts to find out the occurrence of local events such as local festivals, by monitoring crowd behaviours indirectly via Twitter. To detect such unusual geo-social events, the proposed method depend on geographical regularities deduced from the usual behaviour patterns of crowds with geo-tagged microblogs. The decision whether or not there are any unusual events happening in the monitored geographical area is done by comparing these regularities with the estimated ones. The method performs event detection in the following steps: collecting geo-tagged tweets; configuration of region-of-interests (RoIs) is done using a clustering-based space partition method based on the geographical coordinates. The K-partitioned regions over a map, obtained via K-means clustering, are then regarded as RoIs; geographical regularity of each RoI crowd behaviours is estimated during a certain time period using following properties of a RoI: number of tweets, number of users, and moving users. Features are accumulated over historical data using 6-hour time intervals. Unusual events in the monitored geographical area are detected by comparing statistics from new tweets with the estimated behaviour.

Gu et al. [35] proposed ETree, an effective and efficient event modelling solution for social media network sites. ETree used three key components: an n-gram based content analysis technique for identifying and group large numbers of short messages into semantically coherent information blocks; an incremental and hierarchical modelling technique for identifying and constructing event theme structures at different granularities; and an enhanced temporal analysis technique for identifying inherent causalities between information blocks. The identification of core information blocks of an event is done using an n-gram based content analysis technique. Frequent word sequences (i.e., n-grams, or key phrases) among a large number of event-related messages are detected in a first stage. Each frequent sequence represents an initial information block. In the second stage, messages that are semantically coherent are merged into the corresponding information blocks. For each one of the remaining messages, messages that do not contain any of the frequent n-gram patterns, a similarity against each core information block is measured by calculating their TF-IDF weights using words that belongs to both. The weighted cosine similarity between each message and each information block allows the merging of messages into the information block with the highest similarity. Messages that belong to a specific "conversation thread" are also merged into the same information block. The construction of hierarchical theme structures is done by applying an incremental (top-down) hierarchical algorithm based on weighted cosine similarity in the previously identified information blocks. Each theme is represented as a tree

structure with information blocks as the leaf nodes and subtopics as the internal nodes.

Ozdikis et al. [68] proposed a document expansion based event detection method for Twitter using only hashtags. Their expansion was based on second-order relations, which is also known in NLP as distributional similarity. The event detection technique was based on clustering of hashtags by using the semantic similarities between hashtags. Items (i.e. tweets in this context) are clustered according to their similarity in vector space model using agglomerative text clustering. In their agglomerative clustering implementation, values in tweet vectors, i.e. weights of the corresponding terms for each tweet, are set as TF-IDF values. Cluster vectors are calculated by taking the arithmetic mean of values in tweet vectors in each dimension. The similarity of tweet vectors and cluster vectors is calculated by applying the cosine similarity. Tweets are only added to a cluster in case the similarity of the vectors being above a threshold defined empirically.

With respect on how event detection can work on corpora less structured than newswire releases, Jurgens and Stevens [43] proposed an automatic event detection that aims to identify novel, interesting topics as they are published in blogs. Authors proposed an adaptation of the Random Indexing algorithm [44, 87], Temporal Random Indexing, as a new way of detecting events in this media. The algorithm makes use of a temporally-annotated semantic space for tracking how words change semantics and demonstrate how these identified changes could be used to detect new events and their associated blog entries. Based on semantic slice of a single word, which covers all the time periods in which that word has been observed, the detection of events using Temporal Random Indexing is done in three steps: convert the corpus into month long semantic slices; semantic shift are calculated for each word for slices at consecutive timestamps and compared using the cosine similarity. Authors describe changes in angle as a change in a word's meaning, which can signify the presence of an event. Changes in magnitude showed not to be correlated with events; Finally, events are regarded as the selection of the topic words that undergo a significant semantic shift.

Metzler et al. [63] proposed the problem of structured retrieval of historical event information over microblog archives. Unlike all previous work, that retrieves individual microblog messages in response to an event query, they propose the retrieval of a ranked list of historical event summaries by distilling high quality event representations using a novel temporal query expansion technique. Taking a query as input, the proposed microblog event retrieval framework returns a ranked list of structured event representations. This is accomplished through two steps: the timespan retrieval, that identifies the timespans when the event happened; and the summarization step that retrieves a small set of microblog messages for each timespan. Temporal Query Expansion, Timespan Ranking and Timespan Summarization are used in the search task.

Robinson et al. [86] developed an earthquake detector for Australia and New Zealand by monitoring special keywords like "earthquake" and "#eqnz" in Twitter and available geolocation information. Based on the Emergency Situation

Awareness (ESA), the earthquake detector monitors Tweets and checks for specific earthquake related alerts. The system uses ESA burst detection methods based on a binomial model to generate an expected distribution of feature occurrences in a given time window. Then a test on the frequency of observed features in fixed-width time-windows against a statistical content model of historical word frequencies is done. In the cases where the historical model of word frequencies does not fit the observed data, an earthquake situation is identified.

Supervised Detection Controversial events provoke a public discussion in which audience members express opposing opinions, surprise or disbelief. Using social media as a starting point, Popescu and Pennacchiotti [79] addressed the detection of this kind of events, by proposing three alternative regression machine learning models based on Gradient Boosted Decision Trees [31]. Triplets consisting of a target entity, a given time period, and a set of tweets about the entity from the target period, were used. Authors call those triplets a snapshot with the detection task being done in three steps: separation of events and non-event snapshots using a supervised gradient boosted decision trees trained on a manually labeled data set; estimation of a controversy score to each snapshot using an ML regression model; ranking the snapshots according to the controversy score obtained in the previous step. In a successive work, Popescu et al. [80] used additional features with the same framework described earlier to extract events and their descriptions from Twitter. These new features inspired from the document aboutness system Paranjpe [71] allow the ranking of entities in a snapshot with respect to their relative importance to the snapshot.

With the focus on the identification of entertainment event Twitter messages, Benson et al. [15] formulated an approach to the problem as a structured graphical model which simultaneously analyzes individual messages, clusters them according to event, and induces a canonical value for each event property. This technique is able to construct entertainment event records for the city calendar section of NYC.com using a stream of Twitter messages with high precision and acceptable recall. At the message level, the model relies on a conditional random field (CRF) component to extract field values such as the location of the event and artist name. A factor-graph model was used to capture the interaction between each of these decisions. Variational inference techniques allow to make predictions on a large body of messages effectively and efficiently. A seed set of example records constitutes the only source of supervision; alignment between these seed records and individual messages is not observed, nor any message-level field annotation. The output of the model consists of an event-based clustering of messages, where each cluster is represented by a single multi-field record with a canonical value chosen for each field.

By considering a Twitter user as a sensor and tweets as sensory information, Sakaki et al. [88] employed a supervised classification technique to detect specific event types such as earthquakes, typhoons, and traffic jams. Positive events and negative events are classified according to an SVM trained on a manually labelled dataset. Three groups of features are used: statistical features, i.e.: the

number of words in a tweet message, and the position of the query word within a tweet; keyword features, the words in a tweet; word context features, the words before and after the query word. The analysis of the number of tweets over time for earthquakes and typhon data revealed an exponential distribution of events. Authors also mentioned that spikes occur on the number of tweets. Subsequently, a probabilistic spatio-temporal model for the target event that can find the center and the trajectory of the event location is produced. The estimation of the earthquake center and typhoon trajectory was done using Kalman filtering and particle filtering. Particle filters outperformed Kalman filter in both cases.

Massoudi et al. [61] presented a model for retrieving microblog posts that is enhanced with textual and microblog specific quality indicators and with a dynamic query expansion model. They used a generative language modeling approach based on query expansion and microblog "quality indicators" to retrieve individual microblog messages. Being the microblogs documents a special type of user-generated content due their limited size, Massoudi et al. [61], enumerated two interesting effects of its limited size: people use abbreviations or change spelling to fit their message in the allotted space, giving rise to a rather idiomatic language; redundancy-based IR methods may not be usable in a straightforward manner to provide effective access to very short documents. To address the first effect, they introduced credibility indicators for blog post search. To overcome the second effect a re-examination of the potential of local query expansion for searching microblog posts is done using a time-dependent expansion flavor that accounts for the dynamic nature of a topic.

Li et al. [56] proposed a domain-specific event detection method based on pre-specified rules called TEDAS. This system detects, analyses, and identifies relevant crime and disaster related events (CDEs) on Twitter. Based on the authors observation that similar types of CDEs share similar keywords, tweets are collected based on iteratively-refined rules (e.g.: keywords, hashtags). Due to the difficulty to manually define a good set of rules, authors adopted the bootstrapping idea to expand the tracking rule set automatically and iteratively. Next, tweets are classified via supervised learning based on content and Twitter-specific features (i.e.: URLs, hashtags, mentions) and CDE-specific features (i.e.: similarity to CDE tweets, time of day with high crime probability, high crime geographical zones). Location information is extracted using both GPS tagged and location information in tweet content. When no location information is present in the tweet, authors predict user's location as the location from his friends or tweets that minimizes the overall distances between locations in his tweets and from his friends. To rank tweets according to their level of importance, authors propose a learning-to-rank approach, which learns a function to assign a score to each tweet, integrating a variety of signals, such as author's credibility and the number of retweets. To predict a tweet's importance precisely, they explored signals from various aspects, including content, user and usage.

Hybrid Detection Using a set of automatic query building strategies, Becker et al. [12] presented a system for augmenting information about planned events with

Twitter messages. Simple query building strategies were used to achieve high precision results in the task of identifying Twitter messages related to a specific event. To improve recall, they employ term-frequency analysis and co-location techniques on the high-precision tweets to identify descriptive event terms and phrases, which are used recursively to define new queries. Additional queries using URL and hashtag statistics from the high-precision tweets for an event are also built. A rule-based classifier is used to select among this new set of queries, and then use the selected queries to retrieve additional event messages. Becker et al. [12] also developed centrality-based techniques for effective selection of quality event content that may, therefore, help improve applications such as event browsing and search. They address this problem with two concrete steps. First, by identifying each event and its associated Twitter messages using an online clustering technique that groups together topically similar Twitter messages. Second, for each identified event cluster, by providing a selection of messages that best represent the event. With the focus on the challenge of automatically identifying user-contributed content for events that are planned across different social media sites, Becker et al. [13] extended and incorporated into a more general approach their developed techniques of query formulation and centrality based approaches for retrieving content associated with an event on different social media sites.

4.2 Unspecified Event Detection

Unspecified event detection systems rely on either on unsupervised or hybrid detection techniques. The following sections describe examples of those two types of systems.

Unsupervised Detection TwitterMonitor, the trend detection system over the Twitter stream proposed by Mathioudakis and Koudas [62], was also designed to identify emerging topics in real-time. This system also provides meaningful analytics that synthesize an accurate description of each topic. The detection of bursty keywords was done using a data stream algorithm trends are obtained by grouping keywords into disjoint subsets, so all keywords in the same subset appear on the same topic of discussions. Keyword grouping employs a greedy strategy that produces groups in a small number of steps. The system employs context extraction algorithms (such as PCA and SVD) over the recent history of the trend and reports the keywords that are most correlated with it. To identify frequently mentioned entities in trends uses Grapevine's entity extractor [8].

Phuvipadawat and Murata [76] presented a methodology to collect, group, rank and track breaking news using Twitter tweets. Tasks are divided into two stages: story finding and story development: In the story finding, messages are fetched through the Twitter streaming API using pre-defined search queries to get near real-time public statuses. These pre-defined search queries can be messages containing for example, hashtags users often use to annotate breaking news

e.g.: #breakingnews and "breaking news" keyword. To accommodate the process of grouping similar messages, an index based on the content of messages is constructed using Apache Lucene. Messages that are similar to each other are grouped together to form a news story. Similarity between messages is compared using TF-IDF with an increased weight for proper noun terms, hashtags, and usernames. A general implementation of linear chain Conditional Random Field (CRF) sequence models, coupled with well-engineered feature extractors was used as the Named Entity Recognition (NER) technique to identify proper nouns. NER was trained on conventional news corpora; In story development, each news story is adjusted with appropriate ranking through a period of time. The final method ranks the clusters of news using a weighted combination of followers (reliability) and the number of re-tweeted messages (popularity) with a time adjustment for the freshness of the message; Phuvipadawat and Murata [76] emphasized that the key aspect to improving the similarity comparison for short-length messages was to put an emphasis on proper nouns.

Traditional first story detection approaches for news media like the one proposed by Allan et al. [5], which was based on the cosine similarity between documents to detect new events that never appeared in previous documents, revealed to be obsolete when used in a real-time event detection method over social data streams. Petrovic et al. [74] being aware of the limitations constraints of classical event detection methods, both in term of speed and efficiency, proposed a constant time and constant space approach to solve this problem. The proposed system [72] achieved over an order of magnitude speedup in processing time in comparison with the a state-of-the-art system on the first story detection task [6]. The author claimed comparable performance event detection on a collection of 160 million tweets. Modern event detection systems face important challenges when dealing with the high-volume, unbounded nature of today social networks data streams. Using an adapted a variant of the Locality Sensitive Hashing methods [33], was able to detect never seen events when a new bucket is created after hashing a new document to calculate its approximate nearest neighbor. In following work, Petrovic et al. [74] evaluated the use of paraphrases [66] and cross stream event detection by combining Twitter data with Wikipedia spikes [67] and Twitter data with traditional newswires sources [75]. In direct comparison with the UMass system [6], Petrovic et al. [74] also concludes that his approximate technique sometimes outperforms the exact technique. The reason for outperforming the exact system lies in the combination of using LSH and the variance reduction strategy.

Long et al. [58] proposed a unified workflow of event detection, tracking and summarization on microblog data composed by three main steps: in the first events from daily microblog posts are detected using clustering of topical words, afterwards related events are tracked by formulating the event tracking task as a bipartite graph matching problem, and finally tracked event chains are summarized for user to better understand what are happening. Summaries are presented using the top-k most relevant posts considering their relevance to the event as well as their topical coverage and abilities to reflecting event evolution over time.

Topical words are extracted from messages using word frequency, word occurrence in hashtags, and word entropy. The separation of topical words into event clusters is done using top-down hierarchical divisive clustering on a co-occurrence graph. Authors state that, using any clustering method, their proposed feature selection outperforms, document frequency only and document frequency with entropy. It also stated that top-down hierarchical divisive clustering outperforms both k-means and traditional hierarchical clustering no matter what k to use.

Weng et al. [98] proposed the EDCoW, an event detection algorithm that clusters wavelet-based signals built from the analysis of the text stream in Twitter. The algorithm builds signals for individual words by applying wavelet analysis to the frequency-based raw signals of the words. Then filters away the trivial words by looking at their corresponding signal auto-correlations. Remaining words are then clustered to form events with a modularity-based graph partitioning technique. In a direct comparison with Discrete Fourier Transformation (DFT) approaches [37, 36] that converts the signals from the time domain into the frequency domain, Weng et al. [98] use wavelet transformation to analyses signals in both time and frequency domain. Unlike the sine and cosine used in the DFT, which are localized in frequency but extend infinitely in time, the wavelet transformation allows the identification of the exact time and the duration of a bursty event within a signal. Weng et al. [98] argue why the use of wavelet transformation is, in general, a better choice for event detection, giving as one example an event detection systems using a similar technique on Flickr data [22]). Event detection is performed in four separate steps: Construction of signals for individual words using wavelet analysis. Signal construction is based on time-dependent of document frequency-inverse document frequency (DF-IDF), where DF counts the counts the number of documents containing a specific word, while IDF accommodates word frequency up to the current time step; The detection of events done by grouping a set of words with similar patterns of burst. To achieve this, the similarities between words need to be computed first, by building a symmetric sparse word cross-correlation matrix. This step is called computation of cross-correlation; Applying a modularity-based graph partitioning in the cross-correlation matrix will allow to group co-occurrences of words at the same time. Weng et al. [98] formulated the event detection problem as a graph partitioning problem, i.e. to cut the graph into subgraphs, where each subgraph corresponds to an event, which contains a set of words with high cross-correlation. Finally, the quantification of event significance compute a significance value for each event by summing all the cross-correlation values between signals associated with an event and discounting the significance when the event is associated with too many words.

A lightweight method for event detection using wavelet signal analysis of hashtag occurrences in the Twitter public stream was presented by Cordeiro [23]. In his work hashtags were used to build signals, instead of individual words [98]. The author considered that an abrupt increase in the use of a given hashtag at a given time is a good indicator of the occurrence of an event. Hashtags signals were constructed by counting distinct hashtag mentions grouped in intervals of

5 minutes. Each hashtag represented a separate time series. The context of the hashtag was kept by concatenating all the text included in documents with mentions to a specific hashtag. Four separate tasks were performed to detect events: representation of each one of the hashtag signals in a time-frequency representation using a continuous wavelet transformations (CWT); Signal pre-processing using Kolmogorov-Zurbenko Adaptive Filters to remove noise; Wavelet peak and local maxima detection using the continuous wavelet transformation; Finally, event summarization was done by applying LDA [18] topic inference to retrieve a list of topics that describes the event.

Li et al. [55] proposed Twevent, a segment-based event detection system for tweets. Authors define a tweet segment as one or more consecutive words (or phrase) in a tweet message. Based on the fact that tweet segments contained in a large number of tweets are likely to be named entities (e.g. Steve Jobs) or some semantically meaningful unit (e.g. Argentina vs. Nigeria), authors refer that a tweet segment often contains much more specific information than any of the unigrams contained in the segment. Where other techniques rely on bursts of terms or topics (unigrams) to detect events, this particular system first detects bursty tweet segments as event segments. Tweets are split into non-overlapping and consecutive segments, this tweet segmentation problem is formulated as an optimization problem with an objective function based on the stickiness of a segment or a tweet by using the generalized Symmetric Conditional Probability (SCP) for n-grams with n greater or equal to 2, supported by statistical information derived from Microsoft Web N-Gram service and Wikipedia. Bursty segments are identified by modeling the frequency of a segment as a Gaussian distribution based on predefined fixed time-window. By considering their frequency distribution and their content similarity, the grouping of event-related segments as candidate events was done using k-Nearest Neighbor graph and a cosine based similarity measure. Each one of the event clusters is regarded as candidate events detected in that time window. Wikipedia is exploited to identify the realistic events and to derive the most newsworthy segments to describe the identified events.

Agarwal et al. [1] model the problem of discovering events that are unravelling in microblog message streams as a problem of discovering dense clusters in highly dynamic graphs. Authors state that the identification of a set of temporally correlated keywords is the starting point to identify an emerging topic. Moreover, they go further and define temporally correlated keywords as keywords that show burstiness at the same time and are spatially correlated, and more specifically keywords that co-occur in temporally correlated messages from the same user. To capture these characteristics, a dynamic graph model that uses the moving window paradigm and is constructed using the most recent messages present in the message stream, was used. An edge between two nodes — representing two keywords — indicates that messages from a user within the recent sliding window involve the respective keywords. A Correlated Keyword Graph (CKG) captures the properties of microblog contents by representing all the keywords, after removing stop words, appearing in the messages in the current window

as nodes in an undirected graph. Emerging events are, therefore, identified by discovering clusters in CKG. Clusters of interest are obtained via majority quasi cliques (MQCs). Being the discovering majority quasi cliques an NP-complete problem even for static graphs, authors proposed the use of short cycle property (SCP) of MQCs to make event discovery a tractable and local problem. Because Correlated Keyword Graph is dynamic and not static, efficient algorithms for maintaining the clusters locally even under numerous additions and deletions of nodes and edges were also proposed.

Under the premises that documents that describe the same event contain similar sets of keywords, and graph of keywords for a document collection contain clusters of individual events, Sayyadi et al. [90] proposed an event detection approach that overlays a graph over the documents, based on word co-occurrences. Authors assume that keywords co-occur between documents when there is some topical relationship between them and use a community detection method over the graph to detect and describe events. The method uses two steps: Building of a KeyGraph, by first extracting a set of keywords from documents, then for each keyword calculating the term frequency (TF), document frequency (DF) and the inverse document frequency (IDF). Using keywords with higher occurrences nodes are created in the KeyGraph for keyword. Edges between nodes (keywords) are added if the two co-occur in the same document; Community Detection in KeyGraph, community detection is done removing edges in the graph till communities get isolated. Authors consider that by removing the edges with a high betweenness centrality score, every connected component of the KeyGraph represents a hypothesis about an event, the keywords forming a bag of words summary of the event; Document Clustering, community of keywords are seen as synthetic documents. Original documents are clustered using cosine similarity distance to the keywords synthetic documents. Documents that truly represent events are obtained by filtering keywords synthetic documents with high variance.

Zhao et al. [108] proposed the detection of events by exploring not only the features of the textual content but also the temporal, and social dimensions present in social text streams. Authors define an event as the information flow between a group of social actors on a specific topic over a certain time period. Social text streams are modeled as multi-graphs, where nodes represent social actors, and each edge represents the information flow between two actors. The content and temporal associations within the flow of information are embedded in the corresponding edge. Events are detected by combining text-based clustering, temporal segmentation, and information flow-based graph cuts of the dual graph of the social networks. The proposed method begins with social text streams being represented as a graph of text pieces connected by content-based document similarity. The weight of each word in the text piece is quantified as the TF-IDF, with the content-based similarity being defined as the cosine similarity of the vector representation of each text pieces. Using a graph cut algorithm [60], text pieces are then clustered into a set of topics. The resulting graph then is partitioned into a sequence of graphs based on the intensity along the temporal dimension using the adaptive time series model proposed by Lemire

[54]. Each graph in the temporal dimension, for a given topic, represents a communication peak (intensive discussion) that corresponds to a specific aspect or a smaller event. After that, each graph in a specific time window with respect to a specific topic is converted into its dual graph and the dual graph is further partitioned into a set of smaller graphs based on the dynamic time warping [45] based information flow pattern similarity between social actor pairs using graph cut algorithm [60]. Finally, the output of each event will be represented as a graph of social actors connected via a set of emails or blog comments during a specific time period about a specific topic.

Pohl et al. [78] proposed crisis-related sub-event detection using social media data obtained from Flickr and Youtube. Considering the Geo-referenced data an important source of information for crisis management, authors decided to apply a two-phase clustering approach to identify crisis-related sub-events. The method relies on longitude and latitude coordinates of existing data items for sub-event detection. In a pre-processing step each item is therefore represented in two parts: the coordinates, represented by longitude and latitude values; and the terms, extracted from textual metadata fields belonging to a specific item. Term frequency-inverse document frequency (tf-idf) values are also computed. The two-phase clustering consists of the calculation of term-based centroids with a Self-Organizing Map (SOM) Kohonen [47] using the geo-referenced data. In the second phase, the assignment of best fitting data points to the calculated centroids using reassignment and the cosine distance measure is done.

Chen and Roy [22] presents a method to perform event detection from Flickr photos by exploiting the tags supplied by user's annotations. As not every photo represents an event, authors use feature-pivot approaches to detect event-related tags before detecting events of photos. The methods is done in three steps: In Event Tag Detection, the temporal and locational distributions of tag usage are analyzed in order to discover event-related tags using the Scale-structure Identification (SI) approach Rattenbury et al. [83]. A wavelet transform is employed to suppress noise; In Event Generation, by examining the characteristics of the distribution patterns, authors are able to distinguish between aperiodic-event-related and periodic-event-related tags. Event-related tags are clustered such that each cluster, representing an event, consists of tags with similar temporal and locational distribution patterns as well as with similarly associated photos. A density-based clustering method was used (DBSCAN) Ester et al. [28]; In Event Photo Identification, for each tag cluster, photos corresponding to the represented event are extracted.

Corley et al. [24] proposed a conceptual framework for interpreting social media as a sensor network. The system quantifies a baseline from the social sensor measurements. Those baselines provide the expected value at a particular point in time of the volume of social media features fitting some criterion. Using a brute-force approach, they detect aberrations (Events) in the sensor data when an observed value is significantly different from the expected baseline. Signals are built considering the varying time-dependent measures of frequency such as user retweets, term and hashtag usage, and user-specific posts. Measures like

the signal magnitude, which is the value of the centered moving average of an indicated time period of that signal, and the social signal noise, defined as the range of counts bounded by the values of two standard deviations above and below the signal magnitude, are used to calculated the signal aberration (or event) is an instance when the social signal exceeds signal noise boundaries. To produce baseline signals for related topics, topic clustering through using the dot product similarity metric between authors and their hashtag usage, over the course of a specified time period, is used.

Tanev et al. [94] described an Information Retrieval approach to link news about events to Twitter messages. The authors also explored several methods for creating event-specific queries for Twitter. They also claim that methods based on utilization of word co-occurrence clustering, domain-specific keywords and named entity recognition have shown good performance. Basic detection of known bi-grams in the input news article is performed using an index of word uni-grams and bi-grams previously calculated. Because each word uni-gram and bi-gram is accompanied by its frequency and the frequency of the co-occurrences with the other uni/bi-grams, the same index is also used to calculate IDF for each term and suggest classes of terms which are used to formulate the queries to Twitter based on the co-occurrence information. Other techniques like word co-occurrences, named entities, domain-specific keywords were used to improved the detection method.

Dou et al. [25] proposed an interactive visual analytics system, LeadLine, that automatically identify meaningful events in news and social media data and support exploration of the events. To characterize events, topic modeling, event detection, and named entity recognition techniques were used to automatically extract information regarding event details. First, text data such as news stories and microblog messages are organized based on topical themes using LDA [18]. An Early Event Detection algorithm is used to identify the temporal scale for events by determining the length and "burstyness" of events.

Supervised Detection No supervised detection techniques to detect unspecified events were included. No pure supervised event detection systems to detect unspecified events were found in the literature. This fact may be related to the fact that supervised techniques with prior training on ground truth datasets, could not detect unforeseen events in that dataset. Supervised Detection techniques are always used in conjunction with unsupervised techniques (Hybrid Detection) that are being described in the following section.

Hybrid Detection Sankaranarayanan et al. [89] proposed a news processing system, called TwitterStand, that primarily demonstrates how to use a microblog service (i.e. Twitter) to automatically obtain breaking news from the tweets posted by Twitter users. Since the geographic location of the user as well as the geographic terms comprising the tweets play an important role in clustering tweets and establishing clusters' geographic foci, providing users a map interface for reading this news. This system discards tweets that clearly cannot be

news by using a naive Bayes classifier previously trained on a training corpus of tweets that have already been marked as either news or junk. A clustering algorithm based on weighted term vector according to TF-IDF and cosine similarity was used to form clusters of news. The leader-follower clustering [26] algorithm needed to be modified in order to work in an online fashion.

Cataldi et al. [20] use burstiness of terms in a time interval to detect when an event is happening. They proposed a topic detection technique that retrieves in real-time the most emergent topics expressed by the Twitter community. The process begins with the extraction and formalisation of the user-generated content expressed by the tweets as vectors of terms with their relative frequencies; author's authority is calculated by the Page Rank algorithm [69] applied to a directed graph of the active authors based on their social relationships; for each term, its life cycle is modeled according to an aging theory [21] that leverages the user's authority in order to study its usage in a specified time interval; a set of emerging terms is selected by ranking the keywords depending on their life status (defined by an energy value). Supervised term selection relies on a user-specified threshold parameter while the unsupervised term selection relies on an unsupervised ranking model with the cut-off being adaptively computed; finally a navigable topic graph is created which links the extracted emerging terms with their relative co-occurrent terms in order to obtain a set of emerging topics.

Becker et al. [14] explored approaches for analyzing the stream of Twitter messages to distinguish between messages about real-world events and non-event messages. Their approach relies on a rich family of aggregate statistics of topically similar message clusters. Using an incremental, online clustering technique that does not require a priori knowledge of the number of clusters, a task of grouping together topically similar tweets is done. To identify event clusters in the stream, a variety of revealing features is computed using statistics of the cluster messages. Authors used a combination of temporal, social, topical, and Twitter-centric features that must be updated periodically once that they constantly evolve over time. Temporal Features characterize the volume of frequent cluster terms (i.e., terms that frequently appear in the set of messages associated with a cluster) over time. These features capture any deviation from expected message volume for any frequent cluster term or a set of frequent cluster terms. Social Features capture the interaction of users in a cluster's messages. These interactions might be different between events, Twitter-centric activities, and other non-event messages. User interactions on Twitter include retweets, replies, and mentions. Topical Features describe the topical coherence of a cluster, based on a hypothesis that event clusters tend to revolve around a central topic, whereas non-event clusters do not. Twitter-centric features target commonly occurring patterns in non-event clusters with Twitter-centric behavior, including tag usage, and presence of multi-word hashtags. Subsequently, classification via support vector machine (SVM) using the cluster features representation and a previously labeled training set of clusters is done in order to

decide whether or not the cluster, and its associated messages, contains event information (i.e.: distinguish between event and non-event clusters).

Ritter et al. [85] proposed TwiCal, an open-domain event-extraction and categorization system for Twitter. The system extract event phrases, named entities, and calendar dates from Twitter by focusing on certain types of words and phrases. Named entities are extracted using a named entity tagger trained on 800 randomly selected tweets, while the event mentions are extracted using a specific Twitter-tuned part-of-speech tagger[84]. The extracted events are classified retrospectively into event types using a latent variable model (LinkLDA [27]) which infers an appropriate set of event types to match the data (via collapsed Gibbs Sampling using a streaming approach [107]), and then classifies events into types by leveraging large amounts of unlabeled data. The approaches used were based on latent variable models inspired on modeling selectional preferences, and unsupervised information extraction.

Table 3: Collection type, corpus size and temporal scope of datasets

Reference	Collection	Corpus size	Temporal scope
Hu et al. [39]	Technorati popular queries	4075 * 15 queries	from 2006-11-08 1AM to 2008- 03-31 10PM (17 months)
Jurgens and Stevens [43]	Blog articles harvested by BlogLines	15,725,511 blog entries	one year (2006)
Popescu and Pennacchiotti [79]	Twitter streaming api	738,045 Twitter snapshots	July 2009 - February 2010
Popescu et al. [80]	Twitter streaming api		
Benson et al. [15]	Twitter streaming api	4.7 Million tweets (5,800 messages)	three weekends
Lee and Sumiya [52]	Twitter Search API	21,623,947 geo-tagged tweets from 366,556 distinct users	one and a half months (2010/06/04–2010/07/20)
Sakaki et al. [88]	Twitter Search API	49,314 tweets	one month; 2009 Aug. 10 01:00 - 2009 Oct. 12 18:42
Becker et al. [12]	Twitter Search API	NA	NA
Becker et al. [13]	Last.fm events, EventBrite, LinkedIn events, and Facebook events	NA	May 13, 2011 and June 11, 2011
Massoudi et al. [61]	Twitter Search API	110,038,694 tweets	Nov ‘09–Apr ‘10
Metzler et al. [63]	Twitter streaming api	46,611,766 English tweets	July 16, 2010 and Jan 1st, 2011
Gu et al. [35]	Twitter Search API	3.5 million tweets	5 month pediod
Li et al. [56]	Twitter Search API	1 million of CDE tweets	two months
Ozdikis et al. [68]	Twitter Search API	388K tweets	March 16, 2012 and March 19, 2012
Sankaranarayanan et al. [89]	Twitter streaming api / Twitter Search API	NA	NA
Cataldi et al. [20]	Twitter streaming api	3 million tweets	13th and 28th of April 2010
Mathioudakis and Koudas [62]	Twitter streaming api	1.2 million per day	NA
Phuvipadawat and Murata [76]	Twitter streaming api	NA	February 2010

Table 3: Collection type, corpus size and temporal scope of datasets

Reference	Collection	Corpus size	Temporal scope
Petrovic et al. [74]	Twitter streaming api	163.5 million timestamped tweets	six months (April 1st 2009 to October 14th 2009)
Becker et al. [14]	Twitter streaming api	2,600,000 Twitter messages	February 2010
Long et al. [58]	Sina Microblog API	22 million microblog posts	December 23th, 2010 to March 8th, 2011
Weng et al. [98]	Twitter Search API	4,331,937 tweets	April 13, 2011 till May 13, 2011
Cordeiro [23]	Twitter streaming api	13.651.464 tweets	00:00 of the 10th of November and 23:59 of 18th of November of 2011
Li et al. [55]	Wikipedia / Twitter streaming api	3, 246, 821 articles from wikipedia (30 Jan, 2010) / 4, 331, 937 tweets	April 13, 2011 till May 13, 2011
Agarwal et al. [1]	Twitter streaming api	1.3 million	18 hours on 29th Feb 2012
Sayyadi et al. [90]	Live Labs's Social Streams platform	18,000 posts	two months (May and June 2009)
Zhao et al. [108]	Enron Email dataset / Dailykos blog dataset	619,446 messages / 249543 blog entries	1998 to year 2002 / October 12, 2003 to October 28, 2006
Pohl et al. [78]	Youtube / Flickr	4 datasets (2.039.442, 31.222, 178.274 and 455.700 videos and images)	4 datasets (04-19 May, 22 July, 06-10 Aug and 23-29 Aug)
Chen and Roy [22]	Flickr	7, 405, 135 photos, annotated with 44, 139, 261 tags	two-year-period starting at Jan 01, 2006, until Dec 31, 2007
Ritter et al. [85]	Twitter streaming api	100 million tweets	November 3rd 2011
Robinson et al. [86]	Twitter streaming api	870 million	September 2011–January 2013
Corley et al. [24]	Twitter streaming api	NA (8.73 TB)	13-June 2011 through 11- March 2013
Tanev et al. [94]	Twitter Search API	NA	NA
Dou et al. [25]	Twitter Search API / CNN news	100,000 tweets / 3,130 news articles	Aug 19 to Nov 01 2011 / Aug 15, 2011 to Nov 5, 2011

4.3 Detection Methods

Distinct methods to perform event detection are described in the following sections.

Clustering Clustering is the most used technique in event detection systems. Different clustering techniques are described in literature, from classical clustering, passing by incremental clustering, hierarchical clustering or graph partitioning techniques, authors see the separation of documents in similar clusters as a valid method to detect events.

Although they require a prior knowledge of the number of clusters, partition clustering techniques such as K-Means, K-median, K-medoid were used by [52]. Clustering of hashtags based on the similarity of documents vectors and cluster vectors using cosine similarity was proposed by [68]. Frequent word sentences (n-grams) using weighted cosine similarity were also used by [35]. The clustering of wavelet signals was proposed in [23] to signals constructed by hashtags occurrences, and using Co-occurrence of words in [98].

With the necessity of grouping continuously arriving text documents, incremental threshold-based clustering approaches need to be used. Examples of this approach are [39] were incremental clustering to the stream of query profiles is proposed and the Locally Sensitive Hashing method proposed by [72] where documents are clustered after applying a dimensional reduction technique. The major drawbacks of these methods are the fragmentation issues and the correct setting for the threshold value.

Graph-based clustering algorithms were also used. Hierarchical divisive clustering techniques used on a co-occurrence graph, that connects messages according to word co-occurrences, to divide topical words into event clusters [1] [58] [94]. Modularity-based graph partitioning techniques are used to form events by splitting the graph into subgraphs each one corresponding to an event [90]. PageRank was used as an alternative to the costly of finding the largest eigenvalue of the modularity matrix [20]. In general hierarchical clustering algorithms do not scale because they require the full similarity matrix. [55] proposed a k-Nearest Neighbour graph of non-overlapping and consecutive document segments based on the generalised Symmetric Conditional Probability (SCP) for n-grams.

Classification Classification algorithms, commonly used for the detection of specified events, rely mainly on supervised learning approaches. Classification algorithms include naive Bayes [14] [89], support vector machines (SVM) [14] [88] and gradient boosted decision trees [79] [80]. Classifiers are typically trained on a small set of documents collected over a few months or weeks and labeled according event or non-event ([14] [89]), earthquake or non-earthquake ([88]) and controversial or non-controversial event ([79] [80]). Usually labeling involves human annotators with domain knowledge and is done manually. Previous filtering of irrelevant messages to increase accuracy is also done, e.g.: [88] filter documents that contains special words like “earthquake”.

Dimension Reduction Dimension Reduction techniques are used in most cases to speed up the event detection methods. They are commonly used in streaming scenarios were very high volumes of documents arrive at very high speeds. Normally they are used in conjunction with other techniques (i.e.: clustering, classification, etc.). Petrovic et al. [72] proposed a first story detection algorithm using Locality-sensitive hashing (LSH). This method performs a probabilistic dimension reduction of high-dimensional data. The basic idea is to hash the input items so that similar items are mapped to the same buckets with high probability (the number of buckets being much smaller than the universe of possible input items). With this improvement the scaling problem of the traditional approaches to FSD, where each new story is compared to all, was overcome by a system that works in the streaming model and takes constant time to process each new document, while also using constant space [65]. The proposed system follows the streaming model of computation where items arrive continuously in a chronological order and are processed, each new one, in bounded space and time.

Early successful approaches such as Latent Semantic Analysis use the Singular Value Decomposition (SVD) to reduce the number of dimensions. Principle Component Analysis (PCA) and SVD event detection techniques were addressed by [62]. Although SVD resulted in significant improvements in information retrieval, their poor performance makes them impractical for use in large corpora. Moreover, the SVD and other forms of PCA must have the entire corpus present at once, which makes it difficult to update space as new words and contexts are added. This is particularly problematic for event detection, as the corpus is expected to grow continuously as new events occur. Random Indexing offers an alternative method for reducing the dimensionality of the semantic space by using a random projection of the full co-occurrence matrix onto a lower dimensional space and was used by [43] as the underlying technique for event detection.

Wavelets analysis Weng et al. [98] attempted to solve the event detection in the Twitter online social network by proposing the detection of generic events using signal analysis. The algorithm, called Event Detection with Clustering of Wavelet-based Signals, builds signals for individual words by applying wavelet analysis to the frequency-based raw signals of the words. It then filters away the trivial words by looking at their corresponding signal auto-correlations. The remaining words are then clustered to form events with a modularity-based graph partitioning technique. The algorithm didn't follow the streaming model defined by [65] and is not expected to scale to unbounded text streams. Additionally the authors applied a Latent Dirichlet Allocation algorithm to extract topics from the detected events.

Burstiness analysis Analysis of burstiness is also a frequent technique. Analysis if burstiness of special keywords was proposed by [86] while burstiness of topics extracted via LDA was proposed by [25]. Pan and Mitra [70] proposes two event detection approaches using generative models. In the first approach

they combined Latent Dirichlet Allocation (LDA) model [18] with temporal segmentation and spatial clustering and afterwards adapted an image segmentation model, Spatial Latent Dirichlet Allocation (SLDA) [96], for spatial-temporal event detection on text.

Other Hybrid detection approaches are used in techniques composed by more than one step. Supervised classification or detection techniques are commonly used to identify relevant or important documents before performing the unsupervised step (e.g.: clustering) [89]. Other techniques use a factor graph model that simultaneously detects information of events using supervised CRF and then clusters them according to the event type [15]. A temporal query expansion method was proposed by [63] while Generative Language models were proposed by [61].

5 Datasets

Event detection research is hampered by the lack of standard corpora that could be used to evaluate and benchmark systems. Most researchers that work on event detection, often create their ad-hoc corpora to perform the evaluation. Event labelling is typically done by manual inspection or using external sources/systems to mark events, very often are not publicly available, and usually present problems that pass for being: i) tied to a specific domain application or data source; ii) they only cover high-volume events ignoring low-volume events; iii) they do not cover broad range of event types. Table 3 presents all the datasets and respective properties used for the evaluation of each one of the techniques. Through a quick analysis, it can be observed the heterogeneity in terms of source, size and temporal scope of each one of the used corpus.

6 Conclusions

This chapter presents a survey of techniques proposed for event detection in online social networks. The survey also presents an overview of the challenges that event detections techniques face when dealing with today's Online Social Networks data. While some techniques were designed for the detection of specified events (i.e. natural disasters), others were designed to detect events without prior information of the event itself (i.e. unspecified events).

Event detection techniques are classified according to the type of target event into specified or unspecified event detection. Depending on the target application the way data is being analyzed, the techniques are also classified into Online New Event Detection (NED) or in Retrospective Event Detection (RED). Depending on the underlying detection method involved in the event detection, a classification in supervised, unsupervised or hybrid approaches was also done. Depending if the detection method operates at the document or feature domain, the techniques could also be classified in two main categories: Document-pivot

techniques and Feature-pivot techniques. A resume of the main detection methods was also provided, clustering methods are the most used in unsupervised detection systems for unspecified event detection. Classification methods are in the basis of most of the supervised methods for specified event detection. Dimension reduction approaches, specially LSH, is used when processing of high volumes of data arriving at very high speeds. Systems based on burstiness analysis are commonly used to monitor trends and changes in behaviour that may indicate the presence of events. In the present survey is also shown that there is a very high variety of applications and sources of data. It was shown that most of the techniques use different datasets and evaluation methods, which makes their direct comparison almost impossible. Some of them also have different event detection objectives and meet specific detection requirements.

Although the extensive literature presented an high degree of maturity of some methods, the event detection problem is still one of the most actives in the research community. The continuous growth and evolving of the Online Social Networks Services is challenging state-of-the-art methods in terms of volume, speed and data diversity. Recent trends in research using approximation methods show that equivalent results were obtained when compared to exact methods in a much more efficient way.

Acknowledgements

This work was supported by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT), and by European Commission through the project MAESTRA (Grant number ICT-2013-612944).

References

1. Agarwal, M.K., Ramamritham, K., Bhide, M.: Real time discovery of dense clusters in highly dynamic graphs: Identifying real world events in highly dynamic environments. Proceedings of the VLDB Endowment 5(10), 980–991 (2012), http://arxiv.org/abs/1207.0138
2. Aggarwal, C.C., Zhai, C.: A Survey of Text Clustering Algorithms. In: Mining Text Data, pp. 77–128 (2012)
3. Allan, J.: Topic detection and tracking: event-based information organization, The Kluwer international series on information retrieval, vol. 12. Springer (2002), http://portal.acm.org/citation.cfm?id=772260
4. Allan, J., Jin, H., Rajman, M., Wayne, C., Gildea, D., Lavrenko, V., Hoberman, R., Caputo, D.: Topic-based novelty detection 1999 summer workshop at clsp final report (1999), http://old-site.clsp.jhu.edu/ws99/projects/tdt/final_report/report.pdf, [Online; accessed 02-November-2013]
5. Allan, J., Lavrenko, V., Jin, H.: First Story Detection In TDT Is Hard. In: CIKM '00 Proceedings of the ninth international conference on Information and knowledge management. pp. 374–381. ACM (2000)
6. Allan, J., Lavrenko, V., Malin, D., Swan, R.: Detections, Bounds, and Timelines: UMass and TDT-3. Information Retrieval pp. 167–174 (2000), http://maroo.cs.umass.edu/pdf/IR-201.pdf

7. 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 - SIGIR '98. pp. 37–45. ACM Press, New York, New York, USA (1998), http://portal.acm.org/citation.cfm?doid=290941.290954
8. Angel, A., Koudas, N., Sarkas, N., Srivastava, D.: What's on the grapevine? Proceedings of the 35th SIGMOD international conference on Management of data - SIGMOD '09 p. 1047 (2009), http://portal.acm.org/citation.cfm?doid=1559845.1559977
9. Atefeh, F., Khreich, W.: A Survey of Techniques for Event Detection in Twitter. Computational Intelligence 0(0), n/a–n/a (2013), http://doi.wiley.com/10.1111/coin.12017
10. Bampis, E., Jansen, K., Kenyon, C.: Efficient Approximation and Online Algorithms. Springer (2010), http://www.amazon.com/Efficient-Approximation-Online-Algorithms-Combinatorial/dp/3540322124
11. Barabasi, A.L.: The origin of bursts and heavy tails in human dynamics. Nature 435, 207 (2005), http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0505371
12. Becker, H., Chen, F., Iter, D.: Automatic identification and presentation of Twitter content for planned events. Proceedings of the Fifth ... pp. 655–656 (2011), http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2743/3198
13. Becker, H., Iter, D., Naaman, M., Gravano, L.: Identifying content for planned events across social media sites. In: Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12. p. 533 (2012), http://dl.acm.org/citation.cfm?doid=2124295.2124360
14. Becker, H., Naaman, M., Gravano, L.: Beyond Trending Topics: Real-World Event Identification on Twitter. In: ICWSM. pp. 438–441. Technical Report cucs-012-11, Columbia University, The AAAI Press (2011), http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2745/3207
15. Benson, E., Haghighi, A., Barzilay, R.: Event Discovery in Social Media Feeds. Artificial Intelligence 3(2-3), 389–398 (Jun 2011), http://aria42.com/pubs/events.pdfhttp://dl.acm.org/citation.cfm?id=2002472.2002522
16. Berkhin, P.: A survey of clustering data mining techniques. Grouping multidimensional data pp. 25–71 (2006), http://link.springer.com/chapter/10.1007/3-540-28349-8_2
17. Bifet, A., Kirkby, R.: Data stream mining: a practical approach. Tech. rep., The University of Waikato (Aug 2009)
18. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3(4-5), 993–1022 (2003), http://www.crossref.org/jmlr_DOI.html
19. Brants, T., Chen, F.: A System for new event detection. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval SIGIR 03 2002, 330 (2003), http://portal.acm.org/citation.cfm?doid=860435.860495
20. Cataldi, M., Torino, U., Caro, L.D., Schifanella, C.: Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation. Search pp. 1–10 (2010), http://dl.acm.org/citation.cfm?id=1814245.1814249

21. Chen, C.C., Chen, Y.t., Sun, Y., Chen, M.C.: Life Cycle Modeling of News Events Using Aging. 14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings 2837, pp 47–59 (2003), http://link.springer.com/chapter/10.1007/978-3-540-39857-8_7
22. Chen, L., Roy, a.: Event detection from flickr data through wavelet-based spatial analysis. Proceedings of the 18th ACM conference on Information and knowledge management pp. 523–532 (2009), http://dl.acm.org/citation.cfm?id=1646021$\backslash$npapers2://publication/uuid/8EC6E15D-D958-4A0A-88E5-8D62631BF7C5
23. Cordeiro, M.: Twitter event detection: combining wavelet analysis and topic inference summarization. In: the Doctoral Symposium on Informatics Engineering - DSIE'12 (2012), http://paginas.fe.up.pt/~prodei/dsie12/papers/paper_14.pdf
24. Corley, C.D., Dowling, C., Rose, S.J., McKenzie, T.: SociAL Sensor Analytics: Measuring phenomenology at scale. In: 2013 IEEE International Conference on Intelligence and Security Informatics. pp. 61–66. IEEE (Jun 2013), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6578787
25. 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, VAST 2012 - Proceedings. pp. 93–102 (2012)
26. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification (2nd Edition) (Oct 2000), http://dl.acm.org/citation.cfm?id=954544
27. Erosheva, E., Fienberg, S., Lafferty, J.: Mixed-membership models of scientific publications. Proceedings of the National Academy of Sciences of the United States of America 101 Suppl 1, 5220–5227 (2004)
28. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Second International Conference on Knowledge Discovery and Data Mining. pp. 226–231 (1996), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.2930
29. Farzindar, A.: Social Network Integration in Document Summarization. In: Innovative Document Summarization Techniques: Revolutionizing Knowledge Understanding. IGI-Global (2014)
30. Fiscus, J.G., Doddington, G.R.: Topic detection and tracking evaluation overview. Topic detection and tracking pp. 17–31 (2002), http://www.springerlink.com/index/T652P42711XW6421.pdf
31. Friedman, J.H.: Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics 29(5), 1189–1232 (2001), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.9093
32. Fung, G.G.P.C., Yu, J.X.J., Yu, P.P.S., Lu, H.: Parameter free bursty events detection in text streams. Proceedings of the 31st international conference on Very large data bases - VLDB '05 1, 181–192 (2005), http://dl.acm.org/citation.cfm?id=1083616http://www.scopus.com/inward/record.url?eid=2-s2.0-33745624002\&partnerID=tZOtx3y1
33. Gionis, A., Indyk, P., Motwani, R.: Similarity Search in High Dimensions via Hashing. In: VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases. pp. 518–529 (1999), http://portal.acm.org/citation.cfm?id=671516
34. Goorha, S., Ungar, L.: Discovery of significant emerging trends. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). p. 57 (2010), http://dl.acm.org/citation.cfm?doid=1835804.1835815

35. Gu, H., Xie, X., Lv, Q., Ruan, Y., Shang, L.: ETree: Effective and Efficient Event Modeling for Real-Time Online Social Media Networks. In: 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. pp. 300–307. IEEE (Aug 2011), http://dl.acm.org/citation.cfm?id=2052138.2052366
36. He, Q., Chang, K., Lim, E.P.: Analyzing feature trajectories for event detection. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07. p. 207 (2007), http://portal.acm.org/citation.cfm?doid=1277741.1277779$\backslash$nhttp://doi.acm.org/10.1145/1277741.1277779
37. He, Q., Chang, K., Lim, E., Zhang, J.: Bursty Feature Representation for Clustering Text Streams. In: SDM. pp. 491–496 (2007), https://www.siam.org/proceedings/datamining/2007/dm07_050he.pdf
38. Hounshell, B.: The Revolution Will Be Tweeted. Foreign Policy (187), 20–21 (Aug 2011)
39. Hu, M., Sun, A., Lim, E.P.: Event detection with common user interests. In: Proceeding of the 10th ACM workshop on Web information and data management - WIDM '08. p. 1. ACM Press, New York, New York, USA (Oct 2008), http://dl.acm.org/citation.cfm?id=1458502.1458504
40. Hussein, D., Alaa, G., Hamad, A.: Towards usage-centered design patterns for social networking systems. In: Park, J., Yang, L., Lee, C. (eds.) Future Information Technology, Communications in Computer and Information Science, vol. 185, pp. 80–89. Springer Berlin Heidelberg (2011), http://dx.doi.org/10.1007/978-3-642-22309-9_10
41. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing Social Media Messages in Mass Emergency: A Survey (Jul 2014), http://arxiv.org/abs/1407.7071
42. Jo, T., Lee, M.: The evaluation measure of text clustering for the variable number of clusters. In: Liu, D., Fei, S., Hou, Z., Zhang, H., Sun, C. (eds.) Advances in Neural Networks - ISNN 2007, 4th International Symposium on Neural Networks, ISNN 2007, Nanjing, China, June 3-7, 2007, Proceedings, Part II. Lecture Notes in Computer Science, vol. 4492, pp. 871–879. Springer (2007), http://dx.doi.org/10.1007/978-3-540-72393-6_104
43. Jurgens, D., Stevens, K.: Event Detection in Blogs using Temporal Random Indexing. Proceedings of the Workshop on Events in Emerging Text Types pp. 9–16 (2009), http://dl.acm.org/citation.cfm?id=1859650.1859652
44. Kanerva, P., Kristofersson, J., Holst, A.: Random indexing of text samples for latent semantic analysis. In: Proceedings of the 22nd Annual Conference of the Cognitive Science Society. vol. 1036, pp. 16429–16429 (2000), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.6523\&rep=rep1\&type=pdf
45. Keogh, E.: Exact indexing of dynamic time warping. Proceedings of the 28th international conference on Very Large Data Bases VLDB '02 pp. 406–417 (Aug 2002), http://dl.acm.org/citation.cfm?id=1287369.1287405
46. Kleinberg, J.: Bursty and hierarchical structure in streams. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining KDD 02 7(4), 91 (2002), http://portal.acm.org/citation.cfm?doid=775047.775061
47. Kohonen, T.: The self-organizing map. Proceedings of the IEEE 78, 1464–1480 (1990)

48. Kontostathis, A., Galitsky, L.M., Pottenger, W.M., Roy, S., Phelps, D.J.: A survey of emerging trend detection in textual data mining (2004), http://link.springer.com/chapter/10.1007/978-1-4757-4305-0_9
49. Kumaran, G., Allan, J.: Text classification and named entities for new event detection. Proceedings of the 27th annual international conference on Research and development in information retrieval SIGIR 04 pp. 297–304 (2004), http://portal.acm.org/citation.cfm?doid=1008992.1009044
50. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter , a Social Network or a News Media ? Categories and Subject Descriptors. Most 112(2), 591–600 (2010), http://portal.acm.org/citation.cfm?doid=1772690.1772751
51. Lardinois, F.: Readwritesocial: The short lifespan of a tweet: Retweets only happen within the first hour (2010), http://readwrite.com/2010/09/29/the_short_lifespan_of_a_tweet_retweets_only_happen, [Online; accessed 02-April-2013]
52. Lee, R., Sumiya, K.: Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks pp. 1–10 (2010), http://doi.acm.org/10.1145/1867699.1867701
53. Leek, T., Schwartz, R., Sista, S.: Probabilistic Approaches to Topic Detection and Tracking. In: Topic detection and tracking, pp. 67—-83 (2002), http://portal.acm.org/citation.cfm?id=772260.772265
54. Lemire, D.: A better alternative to piecewise linear time series segmentation. In: SIAM Data Mining 2007 (2007)
55. Li, C., Sun, A., Datta, A.: Twevent: Segment-based Event Detection from Tweets. In: Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12. p. 155. ACM Press, New York, New York, USA (Oct 2012), http://dl.acm.org/citation.cfm?id=2396761.2396785
56. Li, R., Lei, K.H., Khadiwala, R., Chang, K.C.C.: TEDAS: A Twitter-based Event Detection and Analysis System. In: Kementsietsidis, A., Salles, M.A.V. (eds.) 2012 IEEE 28th International Conference on Data Engineering. pp. 1273–1276. IEEE (Apr 2012), http://dblp.uni-trier.de/db/conf/icde/icde2012.html\#LiLKC12http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6228186
57. Linguistic Data Consortium: TDT 2004: Annotation Manual – version 1.2, August 4, 2004 (2004), http://projects.ldc.upenn.edu/TDT5/Annotation/TDT2004V1.2.pdf
58. Long, R., Wang, H., Chen, Y., Jin, O., Yu, Y.: Towards effective event detection, tracking and summarization on microblog data. WAIM'11 Proceedings of the 12th international conference on Web-age information managemen (800), 652–663 (Sep 2011), http://dl.acm.org/citation.cfm?id=2035562.2035636
59. Makkonen, J., Ahonen-myka, H., Salmenkivi, M.: Topic Detection and Tracking with Spatio-Temporal Evidence. In: ECIR'03 - 25th European Conference on Information Retrieval. p. 15 (2003), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.8469
60. Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 888–905 (2000), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=868688
61. Massoudi, K., Tsagkias, M., de Rijke, M., Weerkamp, W.: Incorporating query expansion and quality indicators in searching microblog posts pp. 362–367 (Apr 2011), http://dl.acm.org/citation.cfm?id=1996889.1996936

62. Mathioudakis, M., Koudas, N.: TwitterMonitor: Trend Detection over the Twitter Stream. In: Proceedings of the 2010 international conference on Management of data - SIGMOD '10. p. 1155. SIGMOD '10, ACM, ACM Press, New York, New York, USA (2010), http://portal.acm.org/citation.cfm?id=1807306http://portal.acm.org/citation.cfm?doid=1807167.1807306
63. Metzler, D., Cai, C., Hovy, E.: Structured event retrieval over microblog archives. Proceedings of the 2012 Conference of the North ... pp. 646–655 (2012), http://www.aclweb.org/anthology/N12-1083
64. Mohd, M.: Named entity patterns across news domains. BCS IRSG Symposium: Future Directions in Information ... (Fdia), 1–6 (2007), http://www.mendeley.com/research/named-entity-patterns-across-news-domains/
65. Muthukrishnan, S.: Data Streams: Algorithms and Applications. Foundations and Trends in Theoretical Computer Science 1(2), 117–236 (2005), http://www.nowpublishers.com/product.aspx?product=TCS\&doi=0400000002
66. Osborne, M., Lavrenko, V., Petrovic, S., Osborne, M., Lavrenko, V.: Using paraphrases for improving first story detection in news and Twitter. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies. pp. 338–346. The Association for Computational Linguistics (2012), http://www.aclweb.org/anthology/N12-1034
67. Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., Ounis, I.: Bieber no more: First Story Detection using Twitter and Wikipedia. In: Proceedings of TAIA'12 (2012)
68. Ozdikis, O., Senkul, P., Oguztuzun, H.: Semantic Expansion of Hashtags for Enhanced Event Detection in Twitter. The First International Workshop on Online Social Systems (WOSS 2012) (2012)
69. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. World Wide Web Internet And Web Information Systems 54, 1–17 (1998), http://ilpubs.stanford.edu:8090/422
70. Pan, C.C., Mitra, P.: Event Detection with Spatial Latent Dirichlet Allocation. Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries 20, 349–358 (2011), http://portal.acm.org/citation.cfm?id=1315460
71. Paranjpe, D.: Learning document aboutness from implicit user feedback and document structure. In: Proceeding of the 18th ACM conference on Information and knowledge management. p. 365 (2009), http://dl.acm.org/citation.cfm?id=1645953.1646002
72. Petrovic, S., Osborne, M., Lavrenko, V.: Streaming first story detection with application to twitter. Proceedings of NAACL (2010), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.170.9438\&rep=rep1\&type=pdf
73. Petrovic, S.: Real-time Event Detection in Massive Streams. Phd thesis, University of Edinburgh (2012), http://homepages.inf.ed.ac.uk/s0894589/petrovic-thesis.pdf
74. Petrovic, S., Osborne, M., Lavrenko, V.: Streaming first story detection with application to twitter. In: HLT-NAACL. pp. 181–189. The Association for Computational Linguistics (2010)
75. Petrovic, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I., Shrimpton, L.: Can Twitter Replace Newswire for Breaking News? In: 7th International AAAI Conference on Web and Social Media (ICWSM) (2013)

76. Phuvipadawat, S., Murata, T.: Breaking News Detection and Tracking in Twitter. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. WI-IAT '10, vol. 3, pp. 120–123. IEEE (Aug 2010), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5616930
77. Ping Li, Christopher J. C. Burges, Q.W., Li, P., Burges, C.J.C., Wu, Q.: McRank: Learning to Rank Using Multiple Classification and Gradient Boosting. In: Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007 (Jan 2007), http://www.researchgate.net/publication/221619438_McRank_Learning_to_Rank_Using_Multiple_Classification_and_Gradient_Boostinghttp://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.143.6630
78. Pohl, D., Bouchachia, A., Hellwagner, H.: Automatic identification of crisis-related sub-events using clustering. In: Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012. vol. 2, pp. 333–338 (2012)
79. Popescu, A.m., Pennacchiotti, M.: Detecting controversial events from twitter. In: Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10. p. 1873. CIKM '10, ACM Press, New York, New York, USA (2010), http://dl.acm.org/citation.cfm?id=1871751http://portal.acm.org/citation.cfm?doid=1871437.1871751
80. Popescu, A.M., Pennacchiotti, M., Paranjpe, D.: Extracting events and event descriptions from Twitter. In: Proceedings of the 20th international conference companion on World wide web - WWW '11. p. 105. ACM Press, New York, New York, USA (Mar 2011), http://dl.acm.org/citation.cfm?id=1963192.1963246
81. Raimond, Y., Abdallah, S.: The event ontology (2007), http://motools.sf.net/event
82. Rajaraman, A., Ullman, J.D.: Mining of massive datasets. Cambridge University Press, Cambridge (2012), http://www.amazon.de/Mining-Massive-Datasets-Anand-Rajaraman/dp/1107015359/ref=sr_1_1?ie=UTF8&qid=1350890245&sr=8-1
83. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from flickr tags. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 07 pages, 103 (2007)
84. Ritter, A., Clark, S., Mausam, Etzioni, O.: Named Entity Recognition in Tweets: An Experimental Study. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 1524–1534 (2011), http://www.aclweb.org/anthology/D11-1141
85. Ritter, A., Etzioni, O., Clark, S.: Open domain event extraction from twitter. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12 p. 1104 (2012), http://dl.acm.org/citation.cfm?id=2339530.2339704
86. Robinson, B., Power, R., Cameron, M.: A sensitive twitter earthquake detector. In: WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web. pp. 999–1002 (2013), http://www.scopus.com/inward/record.url?eid=2-s2.0-84893039051\&partnerID=tZOtx3y1
87. Sahlgren, M.: Vector-based Semantic Analysis: Representing Word Meaning Based on Random Labels. ESSLI Workshop on Semantic Knowledge Ac-

quistion and Categorization (2002), http://www.sics.se/~mange/papers/VBSA_Esslli.ps

88. 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. pp. 851–860. ACM, ACM (2010), http://dl.acm.org/citation.cfm?id=1772690.1772777
89. Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J.: TwitterStand: News in Tweets. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09. GIS '09, vol. 156, p. 42. ACM Press, New York, New York, USA (2009), http://portal.acm.org/citation.cfm?id=1653781http://portal.acm.org/citation.cfm?doid=1653771.1653781
90. Sayyadi, H., Hurst, M., Maykov, A., Livelabs, M.: Event Detection and Tracking in Social Streams. In Proceedings of International Conference on Weblogs and Social Media (ICWSM) pp. 311–314 (2009), http://www.aaai.org/ocs/index.php/ICWSM/09/paper/viewFile/170/493
91. Schonfeld, E.: Techcrunch: Mining the thought stream (2009), http://techcrunch.com/2009/02/15/mining-the-thought-stream, [Online; accessed 09-July-2013]
92. Snowsill, T., Nicart, F., Stefani, M., De Bie, T., Cristianini, N.: Finding surprising patterns in textual data streams. 2010 2nd International Workshop on Cognitive Information Processing pp. 405–410 (Jun 2010), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5604085
93. Strassel, S.: Topic Detection & Traking (TDT-5). http://www.ldc.upenn.edu/Projects/TDT2004 (2004), http://www.ldc.upenn.edu/Projects/TDT2004
94. Tanev, H., Ehrmann, M., Piskorski, J., Zavarella, V.: Enhancing Event Descriptions through Twitter Mining. In: Sixth International AAAI Conference on Weblogs and Social Media. pp. 587–590 (2012), http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4631$\backslash$nhttp://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4631/5065
95. Timothy Baldwin, Paul Cook, M.L.A.M.L.W., Baldwin, T., Cook, P., Lui, M., Mackinlay, A., Wang, L.: How Noisy Social Media Text, How Diffrnt Social Media Sources? In: Proc. IJCNLP 2013. pp. 356–364 (2013), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.385.1683
96. Wang, X., Grimson, E.: Spatial Latent Dirichlet Allocation. Advances in Neural Information Processing Systems 20, 1–8 (2007), http://people.csail.mit.edu/xgwang/papers/STLDA.pdf
97. Wang, X., Zhai, C., Hu, X., Sproat, R.: Mining correlated bursty topic patterns from coordinated text streams. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 784–793 (2007), http://dl.acm.org/citation.cfm?id=1281276$\backslash$npapers2://publication/uuid/A6A05DF5-1873-4DC4-BAB3-F73712691FCA
98. Weng, J., Yao, Y., Leonardi, E., Lee, F., Lee, B.s.: Event Detection in Twitter. Development (98), 401–408 (2011), http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2767/3299
99. Wikipedia: Facebook — Wikipedia, the free encyclopedia (2013), http://en.wikipedia.org/w/index.php?title=Facebook&oldid=548760277, [Online; accessed 07-April-2013]

100. Wikipedia: Google+ — Wikipedia, the free encyclopedia (2013), http://en.wikipedia.org/w/index.php?title=Google\%2B&oldid=548920007, [Online; accessed 07-April-2013]
101. Wikipedia: Linkedin — Wikipedia, the free encyclopedia (2013), http://en.wikipedia.org/w/index.php?title=LinkedIn&oldid=549175950, [Online; accessed 07-April-2013]
102. Wikipedia: Twitter — Wikipedia, the free encyclopedia (2013), http://en.wikipedia.org/w/index.php?title=Twitter&oldid=549164139, [Online; accessed 07-April-2013]
103. Yang, C.C., Shi, X., Wei, C.P.: Discovering event evolution graphs from news corpora. IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans 39, 850–863 (2009)
104. Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B., Liu, X.: Learning approaches for detecting and tracking news events (1999)
105. Yang, Y., Pierce, T.T., Carbonell, J.G.: A Study of Retrospective and On-Line Event Detection. In: SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia. pp. 28–36. ACM, New York, New York, USA (1998), http://portal.acm.org/citation.cfm?doid=290941.290953
106. Yang, Y., Zhang, J., Carbonell, J., Jin, C.: Topic-conditioned novelty detection. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02. pp. 688–693 (2002), http://dl.acm.org/citation.cfm?id=775047.775150
107. Yao, L., Mimno, D., McCallum, A.: Efficient methods for topic model inference on streaming document collections. Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining (2009) (4), 937 (2009), http://portal.acm.org/citation.cfm?doid=1557019.1557121
108. Zhao, Q., Chen, B., Mitra, P.: Temporal and Information Flow Based Event Detection from Social Text Streams. In: Proceedings of the 22nd national conference on Artificial intelligence - AAAI'07. vol. 2, pp. 1501–1506. AAAI Press (2007), http://www.aaai.org/Papers/AAAI/2007/AAAI07-238.pdf