Online Information Review

Review on event detection techniques in social multimedia

Muskan Garg Mukesh Kumar

Review on event detection techniques in social multimedia

Muskan Garg and Mukesh Kumar
University Institute of Engineering and Technology, Panjab University, Chandigarh, India

Received 24 August 2015
Revised 24 August 2015
Accepted 9 November 2015

Abstract

Purpose – Social Media is one of the largest platforms to voluntarily communicate thoughts. With increase in multimedia data on social networking websites, information about human behaviour is increasing. This user-generated data are present on the internet in different modalities including text, images, audio, video, gesture, etc. The purpose of this paper is to consider multiple variables for event detection and analysis including weather data, temporal data, geo-location data, traffic data, weekday's data, etc.

Design/methodology/approach – In this paper, evolution of different approaches have been studied and explored for multivariate event analysis of uncertain social media data.

Findings – Based on burst of outbreak information from social media including natural disasters, contagious disease spread, etc. can be controlled. This can be path breaking input for instant emergency management resources. This has received much attention from academic researchers and practitioners to study the latent patterns for event detection from social media signals.

Originality/value – This paper provides useful insights into existing methodologies and recommendations for future attempts in this area of research. An overview of architecture of event analysis and statistical approaches are used to determine the events in social media which need attention.

Keywords Social media analysis, Event detection and prediction, Micro-blog latent pattern, Twitter data stream

Paper type Viewpoint

1. Introduction

Social media is the primary source of information of human behaviour due to its free, online and ease of availability. The content on social media websites is uncertain and user-generated. Social networking platforms can be business based (LinkedIn), location based (Foursquare), content sharing (PInterest, Blogs), photo sharing (Flickr, Instagram), microblogging (Twitter, Sina Weibo), video platforms (Youtube, Vimeo), etc. (Mainka *et al.*, 2014). Using this social media data to analyse different latent patterns have become challenging task. Social networking sites may or may not create directed graph of their users. For instance, Twitter creates directed graph which means there exists "follower-following relationship" in which all the profiles are public and X follows Y but Y may/may not follow X (Yardi and Boyd, 2010). However, Facebook do not support directed graph and provides privacy. Thus, Twitter being publically available is more of concern as compared to Facebook. The research field of social media analysis (SMA) has been growing rapidly and act as tool for user-driven access to uncertain information which is present on web. In recent growth, event detection and analysis has been introduced as hot research area in KDD (Wang, 2015). There are three different types of SMA which are associated with events, namely, event enrichment, event detection and event categorization (Liu *et al.*). Event enrichment deals with linking of multimedia to given topic, event detection deals with detection of events from multimedia and event categorization deals with categorization of social

Online Information Review
Vol. 40 No. 3, 2016
pp. 347-361

1468-4527
DOI 10.1108/OIR-08-2015-0281

media based on events. Out of these, event detection is major area of concern in this survey paper. The information being extracted from social media can be used for event detection, analysis, prediction, early warning, etc. Existing event detection approaches can be further broadly classified into feature-pivot (which words refer to event) and document-pivot (similarity between documents) approaches (Zhang *et al.*, 2015) as mentioned in Figure 1.

Events can be of different types for instance disastrous, traffic based, social gatherings, news, outbreak, etc. Many other field of analysis including topic popularity, trending topic and topic detection and tracking are related to event detection. Event is a physical entity whereas topic is considered as both logical and physical entity. Twitter data, being heterogeneous and large, contains various events at different scale (Becker *et al.*, 2011).

Contribution

After analysing different approaches for event detection, we have given remarks for better research work than the present scenario. We framed the flow from topic detection and trending (Allan, 2002) to event prediction (Zhang *et al.*, 2015), trending event detection (Kaleel and Abhari, 2015; Gao *et al.*, 2015), multimodal event detection (Alqhtani *et al.*, 2015; Poria *et al.*, 2015). Also, improved approaches for topic modelling, clustering algorithms and classification algorithms have been examined for different research work during last decade.

Organization

This paper has been organized in different sections. Section 2 provides background details about SMA in Twitter and enlists different types of applications which are associated with event detection from Twitter data. Section 3 discuss about various techniques and challenges based on different parameters for research in event detection during last decade. Section 4 briefly describes the historical perspective and evolution of different approaches for event detection in Twitter streams. Further, Section 5 provides general discussion and finally, Section 6 concludes the paper.

2. SMA in Twitter

Twitter is a kind of social media which acts as a source of information. In Twitter, users post things which they consider important and further shared by followers. It provides

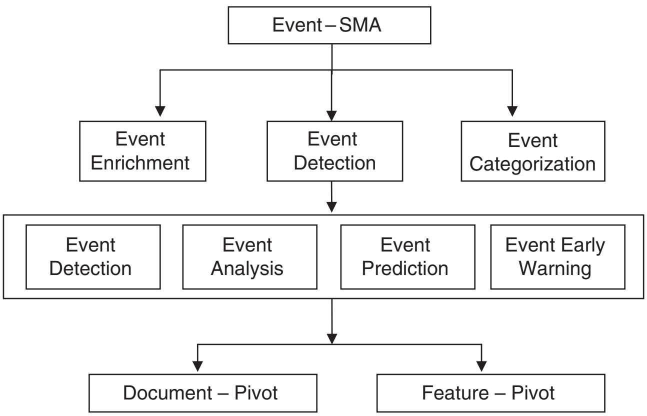

Figure 1. Overview of classification for event – social media analysis

open access to public tweets. Although, tweet limit is 140 characters but recently, Twitter has increased the upper limit for direct messages to 10,000 characters. This can contribute towards improved area of research for Twitter stream analysis. Twitter tweets may contain text and images which may contribute towards bimodal analysis (Alqhtani *et al.*, 2015). In addition to text and image, there exists other modal for analysis including audio, video, human gestures. Data can be extracted from social media of various modalities. Thus, this contributes towards multimodal SMA (Poria *et al.*, 2015).

After extracting data from Twitter, proper pre-processing is important. For this, noise filtering (Liang *et al.*, 2015), spam detection, rumour propagation (Kwon *et al.*, 2013), stopword removal, stemming and lemmatization, etc. are used in different combinations. Many Twitter parameters can be used during event detection process as per requirement. For instance, values of location, coordinates, place, geo can be used for spatial parameters; "created_at", "timestamp_ms", "time_zone" can be used for temporal derivatives; "followers_count", "friends_count", "following" can be used for detection of networks among Twitter users; "retweeted" can be popularity detection; "verified" can be authorized accounts, etc. However, these are subjective measures and can be changed as per proposed methodology and perception of the academic researcher/practitioner. Moreover, on the basis of given parameters, other parameters can be derived. For instance, using temporal information from tweets, we can derive weather data using Wunderground API (Xu *et al.*, 2015). Thus, this may be used to perform multivariate event detection in SMA (Figure 2).

Challenges

Using Twitter, users post messages via different platforms including SMS message service, websites, multitude for clients for both computers and phones (Jackoway *et al.*, 2011). Twitter is believed as the ideal source of information. However, extracting useful information from Twitter is a challenging task. Due to the existing constraint on number of characters of tweet as 140 characters, there exists short forms, misspelled words, shortened URLs, etc. So, disambiguation and semantics are the major area of

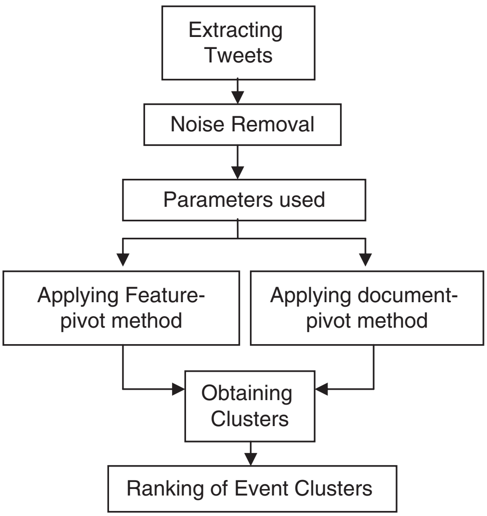

Figure 2.
Overview of event detection architecture

concern. Further, depending on tweet parameters, topic modelling, event detection, crowd-sourcing, noise removal, hotspot detection, clustering identical tweet, categorizing tweets, spam detection, rumour propagation, forecasting, community detection, link prediction, sentiment analysis, opinion mining, tweets credibility, topic popularity, topic ranking, summarization, user profiling, brands affinity, reputation detection, multi-lingual analysis, multimodal analysis are major areas of concern (Wang, 2015). Moreover, improving accuracy and looking for authentication of resulted data are other issues which need to be worked out.

3. Event detection

Event detection is that area of research which may be used for different application domains. These application domains may include instant outbreaks like earthquake, floods, bomb blast, quickly spreading communicable disease like swine flu, bird flu, etc., public gatherings like family functions, corporate gatherings, election campaigns, protests, conferences, ceremonies, clubbing, fest, etc. Information regarding instant outbreaks can be used to alarm emergency management. However, it should be noted that few researchers include tourist hotspots like shopping malls, lakes, gardens as events. Sometimes, logical discussions about a topic, for instance, net-neutrality, live news are also used as events. Rumour control strategies have proved to be the major factor for event detection (Kwon *et al.*, 2013). Recently, citizens wish to know which events are going on in the city as per their interest, events which are popularly discussed and attended, events which are happening regularly, ranking of events in city or in a particular range of geographical dimensions, events which they can attend in future, etc. These types of analysis are performed using multiple algorithms for different techniques.

Event detection has been initialized in 2002 by Allan with his research in topic detection and tracking. This has introduced different topics as base for event detection mechanism. Recently, in 2010, latent patterns were analysed for two major accidents of the town to understand how the information/news spread among local citizens (Yardi and Boyd, 2010). On the other hand, Mathioudakis and Koudas (2010) have studied trend detection from tweet streams by analysing bursty information. Bursty information is that information which is spread among follower-following network to a large extent within very small interval of time. This bursty information is grouped together and popularity is examined on the basis of density of tweets popularity. Becker *et al.* (2011) have researched more about tweets for events. They analysed if the tweet is event based or non-event based. Non-event-based tweets were removed. This can be further used as noise removal in context of event detection methodology. In Ferrari *et al.* (2011) latent urban patterns were detected from Twitter data using latent Dirichlet allocation (LDA) algorithm.

In Yang *et al.* (2012) researchers used the concept of hyperlink-induced topic search (HITS) algorithm for retweets used by users. Also, they used this algorithm for user-user network and tweets network differently and merged it. Finally, they found that the hybrid model successful and outperforms the original or single HITS algorithm. Significant improvement in event detection methodologies by using hashtags instead of words (Ozdikis *et al.*, 2012) and geo-location data for clustering of tweets (Li *et al.*, 2012) is observed. In Nguyen *et al.* (2013) outperforms LDA with hierarchical Dirichlet process (HDP) for sensor-based data. Further, (Sun *et al.*, 2013) new technique have been proposed for road-based travel recommendation using geo-tagged images from Flickr. Support vector machine (SVM) and Dijkastra are the major approaches which were followed for this work. Further, in 2014, ambiguous words of tweet were identified and evaluated for sentiment analysis. Also, different categorization approaches for sentiment analysis were compared (Bravo-Marquez *et al.*, 2014). On the other side, theme-based clustering of tweets was examined by using Wikipedia resource (Tripathy *et al.*, 2014).

In 2015, (Steiger *et al.*, 2015) a new system has been proposed for mining interesting tourist location and travel sequences from geo-tagged public images of Flickr. This is improved with clustering algorithm parallel density-based spatial clustering of application with noise (P-DBSCAN). However (Xu *et al.*, 2015), probabilistic latent semantic analysis (pLSA) have been used as topic modelling approach and outperformed LDA. Also, use of weather data for temporal and spatial-based parameters has been used in this research work. Similarly, shortest travelling route has been detected after using DBSCAN clustering algorithm (Memon *et al.*, 2015). Bimodal for event detection and multimodal (Poria *et al.*, 2015) for sentiment analysis have been proposed for Twitter-based SMA. Noise removal techniques have been proposed (Yamada *et al.*, 2015) to extract event information from text after considering it a signal. For clustering tweets of similar context, micro-blog clique (MC; Gao *et al.*, 2015) and locality sensitive hashing (LSH; Kaleel and Abhari, 2015) are newly proposed techniques which dominate over other clustering algorithms in this domain. Finally, new techniques were proposed for event prediction along with event detection (Zhang *et al.*, 2015). Newly proposed hybrid algorithms for event detection and event prediction is need of the hour for this domain. Academic researchers and practitioners are working on improved accuracy.

4. Evolution of different approaches

Tweets are random views which are tweeted by users. There are different types of approaches which are used for event detection and analysis. To perform data science, the task is distributed into data collection, data analysis and data representation. Although this work is concerned more about analysis but for complete process, data collection and data representation are equally important.

Data collection

Data collection is the primary source for any research. In order to carry out research from tweet, data are extracted from Twitter using different API. One of them is Tweepy API (Almatrafi *et al.*, 2015), which is used in Python for extracting tweets from Twitter stream. The data are extracted in JavaScript Object Notation format and can be stored in any format as per requirement. To access this API, the developer needs to have token key and secret key. Each tweet contains fields like:

{“created_at”, “id”, “id_str”, “text”, “source”, “truncated”, “name”, “screen_name”, “location”, “url”, “description”, “protected”, “verified”, “followers_count”, “friends_count”, “listed_count”, “favourites_count”, “statuses_count”, “created_at”, “utc_offset”, “time_zone”, “geo_enabled”, “lang”}, etc.

Wunderground API is another API which can be used to extract weather information for specific duration of particular place (Xu *et al.*, 2015). Preferably, this information is used to analyse the weather during which user tweet. For this, time and geo-location are inputs given and weather data is the output. Similarly, Google Map API is an API which is used to obtain satellite-based traffic data information of geo-location (Li *et al.*, 2012).

It is important to choose the appropriate social media for research. Twitter is the most widely used platform of social media among all others as shown in (Figure 3). This is because Twitter is directed and in public domain. Graph chart has been prepared as shown in Figure 4 by using Scopus search platform by giving input as the name of social media to be found in title of the article and year is mentioned. This helps to analyse in which platform maximum research is going on. Also, this will help to choose the appropriate platform for further enhancements in right domain. Similarly in Figure 3, it has been observed that year-wise number of publications for Twitter in event-related work is much more than other social media platforms. Although geo-location has received much attention from academic researchers and practitioners, still Foursquare and Flickr needs attention. Also, Facebook gives competition to Twitter in full-domain research, however, for event-related research, Twitter is leading. LinkedIn and Foursquare have least attention for event-based SMA. However, for geo-location-based event analysis, these two platforms may promise good academic research.

Data analysis

Sometimes, in order to remove irrelevant tweets, there exists the need to classify tweets for instance, to identify if the tweet is related to some rumour or not (Kwon *et al.*, 2013), to remove noise by identifying if the tweet is related to personal event, check if the

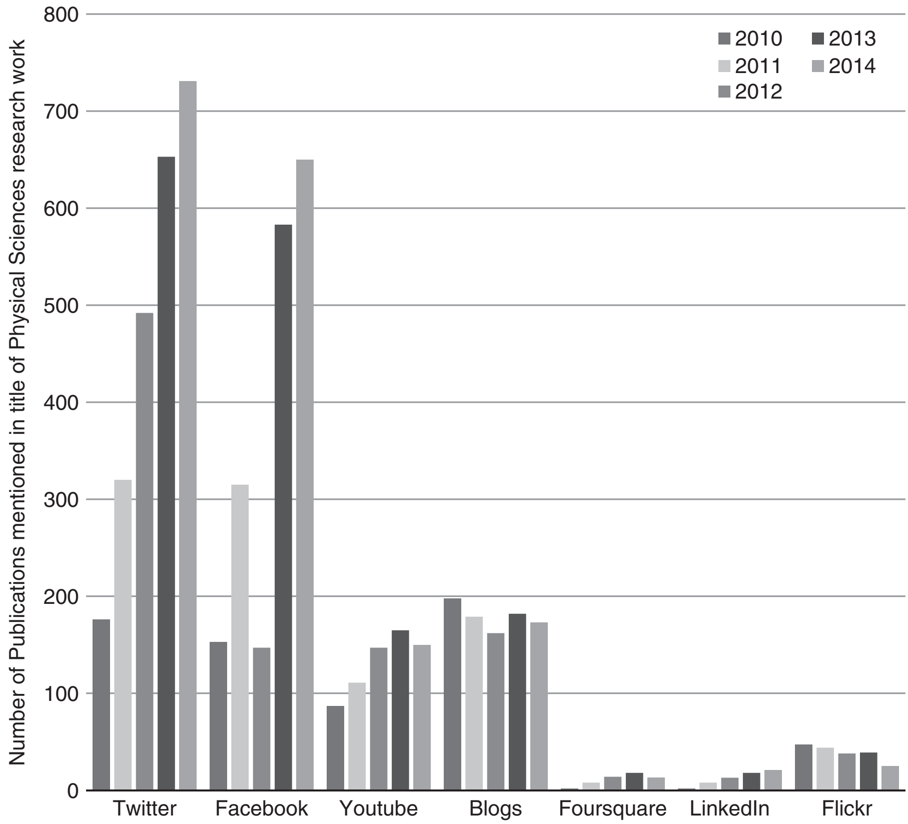

Figure 3.
Number of publications in different social media based on Scopus search

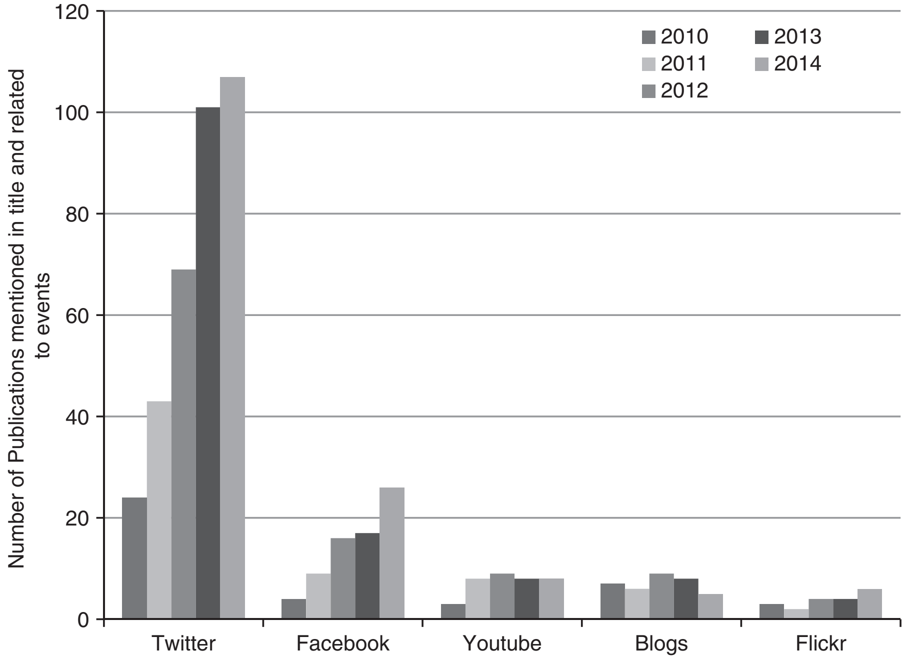

Figure 4.
Number of event-related publications in social media based on Scopus search

tweeting profile is verified or not, etc. For this, different categorization algorithms can be used which includes SVM, Naive Bayes, logistic regression, multilayer perceptron, etc. For sentiment analysis, perceptron and SVM outperform other comparative techniques (Bravo-Marquez *et al.*, 2014). However, SVM is widely used in many other applications including road-based travel recommendation using geo-tagged images (Sun *et al.*, 2013), multimodal sentiment analysis (Poria *et al.*, 2015) and extraction of event information (Yamada *et al.*, 2015). Also, Naive Bayes algorithms has been used recently for sentiment analysis (Almatrafi *et al.*, 2015; Weichselbraun *et al.*, 2014; Bravo-Marquez *et al.*, 2014) and classifying events and non-events from social data (Becker *et al.*, 2011).

After pre-processing of tweets, their clustering is the major area of concern. Initially, *k*-means algorithm have been proposed and used for clustering tweets (Kaleel and Abhari, 2015; Tripathy *et al.*, 2014). But, this was convex cluster-based algorithm. For non-convex linkage, single linkage-based algorithm was proposed and named as DBSCAN. Also, DBSCAN automatically allocates number of clusters which can be framed from give data. However, number of clusters need to be specified in k-means algorithm. Thus, DBSCAN is an improved algorithm which has been used recently to cluster tweets (Memon *et al.*, 2015). However, more improved partition-based DBSCAN algorithm named P-DBSCAN algorithm has been used for social data clustering (Steiger *et al.*, 2015). Hypergraph-based (MC; Gao *et al.*, 2015) and feature vector-based LSH (Kaleel and Abhari, 2015) are those techniques which are used for high-end clustering of similar/related tweets.

Topic modelling is another major area of research in event detection. Latent semantic analysis is the technique of identifying a topic to which the document is related. Academic researchers and practitioners stated that it is very much possible that the document is related to multiple latent topics. Also, the probability of each topic may vary in the document. In pLSA expectation-maximization is used to find local maxima of log likelihood. pLSA is thus used for topic-based context-aware travel recommendation method exploiting geo-tagged photos (Xu *et al.*, 2015). However, a stochastic exploration with Gibbs sampling is used for a much better technique called LDA. Hence, LDA have been used by Ferrari *et al.* (2011) to extract hidden patterns from location-based Twitter data. However, the limitations of LSA, pLSA, LDA are that it is important to specify the number of parameters in LSA and pLSA and that all these models are static which means once trained, they cannot be modified to learn more. Thus, hierarchal Bayesian non-parametric model HDP has been used recently which removes both the limitations and outperforms LDA (Nguyen *et al.*, 2013).

Data representation

After analysis of Twitter streams, next step is to represent data. The analysis performed in existing approaches is observed using different performance metrics and displayed the relevant information via appropriate graphics. After clustering or topic modelling, the information is represented in graphical form. Term frequency – inverse document frequency is the representation of the data. Based on this, different analysis like spatio-temporal locality analysis (Sugitani *et al.*, 2013) and detection of trending events is performed (Kaleel and Abhari, 2015). Cosine similarity is used to measure the extent of similarity among different tweets. This is used in clustering of tweets using Wikipedia and other distance measure techniques (Tripathy *et al.*, 2014). Precision and recall are measured on the basis of true positive, true negative, false negative or false positive. However, graphical representation of results gives the clear picture of observation. In 2013, topic popularity has been observed by using different types of graphical representation like lifetime graphs, evolving graphs and cumulative evolving graphs (Ardon *et al.*, 2013). They used these graphs for data analysis and representation.

Summarization

This survey has been through a large number of research papers. The overall evaluation of input/output parameters, results obtained and critical analysis (remarks) have been mentioned for research work carried out in different research papers in Table I.

5. Discussion

During survey, it has been observed that latent patterns for social media data have been observed in context to event detection using LDA approach. However, later, HDP (Teh *et al.*, 2006) and pLSA has found to be better than LDA (Blei *et al.*, 2002). Thus, these approaches can be applied on the former one. For analysis, instead of single variable, multiple variables can be used like that of weather data, traffic data, etc. Another concerned area is noise removal which includes non-event tweets removal, rumour propagation control, spam control, signal-based filtration, the problems/assumptions/gaps which have been carried out by authors of the papers which have been discussed so far. Semantic clustering is done based on Wikipedia, SentiNet, ConceptNet, LSH, MC, dominant entity, entity linking, etc. Disambiguation is another major issue which needs to be considered. Different research practitioners use different methodologies for different targets. Based on text-feature extraction, clustering and

Table I. Summarization of analysis on social media data

Author year	Research area	Algorithm used	Input dataset	Results obtained	Findings
Mathioudakis	Trend detection	Queuing theory, group burst detection	Twitter tweets of 10 M tweets per day	Track popularity, identify origin of geo-location-based trends	Retweet parameter can be considered in bursty keyword detection for HITS algorithm. This may result in better analysis of tweets
Yardi *et al.*	SMA user characteristic	GUESS: a language and interface for graph exploration	11,017 and 1,602 tweets about church shooting and parking garage	Analysis is performed and outcomes are recorded.	Better poll observation can be done with large number of respondents which may vary the result of where people go for information about local events
Ferrari *et al.* (2011)	Extraction of patterns/routine	Latent Dirichlet allocation	3 GB of New York tweets	Meaningful results obtained	pLSA can be used instead of LDA which may give better results as stated in (Xu *et al.*, 2015)
Becker *et al.* (2011)	Classifying events and non-events	RW-event, Naive Bayes NB-text	26 lakh tweets posted during February 2010	RW-event outperforms NB-text	Proper noise removal may be required as this may reduce the overhead of classifying events
Yang *et al.* (2012)	find latent patterns from SMA	HITS	Tweets of 31 days of October 2011	Proposed HITS outperforms original HITS	Another parameter which can be considered other than retweet and user-user HITS is relation between events on same topic
Ozdikis *et al.* (2012)	Event detection	Twitter vector generation	Three day dataset of Turkish tweets	Higher accuracy obtained using hashtags than words	Better semantic expansion technique can be proposed using entity linking, Wikipedia, dominant entity, sentiNet, WordNet and ConceptNet
Li *et al.* (2012)	Event detection and analysis system	Classification model, API, Java, PHP, Lucene	CDE-related tweets	20% accuracy improved with classification model	It may not always possible that the Tweeter is tweeting from the place of event. They may vary. This assumption can be solved using better techniques
Ardon *et al.* (2013)	Topic popularity analysis	Lifetime, evolving and cumulative evolving graphs	Tweet7, Yahoo place finder service	Detected 16,492 events from 8,250 topics	In-degree and out-degree indicates popular/interesting and spammers/marketer profiles, respectively as stated in (Yardi and Boyd, 2010). This can be considered for advanced options

(*continued*)

Table I.

Author year	Research area	Algorithm used	Input dataset	Results obtained	Findings
Nguyen *et al.* (2013)	social signal-based data analysis	Hierarchical Dirichlet process	Socio-metric badges, reality mining dataset	HDP outperforms LDA up to 90%	Another application where this proposed methodology can be used is non-parametric event detection (Ferrari *et al.*, 2011)
Danyllo *et al.* (2013)	Credit analysis	Social network analysis	3,000 users from Twitter	504 matched out of 3,000	Another application for this methodology can be mapping interests of company employs who belong to same business unit may be similar
Sugitani *et al.* (2013)	locality analysis	IDF	30,149 tweets of Twitter data	Obtained 563 local events	Prediction of the location from non-geotagged tweets is possible as mentioned in (Li *et al.*, 2012) for better analysis
Sun *et al.* (2013)	Travel recommendation	SVM, Dijkastra, entropy filtering	45,950 images from Flickr. 1,223 images discarded	accuracy for BLR is 78.5% and SVM is 80.4%	Instead of static routing algorithms, dynamic routing algorithms can be used after considering real time traffic data using traffic API for tourist visits
Weichselbraun *et al.* (2014)	Identification of ambiguous terms and polarity classification	NB, Graph-based similarity and vector space similarity	Samples from Amazon.com and imdb.com	ConceptNet express better than WordNet	Wikipedia can be used for calculating graph-based similarity. This can be done by calculating scores after matching corresponding terms of tweets within Wikipedia paragraphs or further crawling the links present in Wikipedia paragraphs
Bravo-Marquez *et al.* (2014)	Sentiment recognition	NB, LR, MLP and SVM	STS, Sanders and SemEval	Perceptron and SVM achieve best performance	Further improvement can be observed by comparing these techniques with recurrent neural network, back propagation neural networks
Tripathy *et al.* (2014)	Theme-based clustering of tweets	WIKI-*k*-means, TF-IDF-*k*-means	From June 2009 to August 2009. 100,000 tweets were used for experiment	WIKI-*k*-means outperforms TF-IDF-*k*-means with 0.523 and 0.438 *F*-scores, respectively	Entity linking, dominant entity, sentiNet, WordNet, ConceptNet and other semantic techniques can be compared with the proposed methodology for optimal performance evaluation

(*continued*)

Author year	Research area	Algorithm used	Input dataset	Results obtained	Findings
Kaleel *et al.*	To detect the trending events	LSH, K-means, TF-IDF, prefix tree, NMI, entropy	Tweets 21 May 2011 queries with filtration gives 1,694 tweets	LSH performs 12.5% and 16.6% better results than k-means for purity and NMI, respectively	LSH can be further compared with micro-blog clique (Gao *et al.*, 2015) to cluster highly related tweets. Also, multimodal content can be considered to identify better methodology
Jain *et al.*	Bimodal event detection	TF-IDF HOG, GLCM, SVM, kNN	Twitter data 28 December 2014 for Indonesia Asia air flight 8501	Accuracy: only text – 0.89, only images – 0.86 and using both is 0.94	Better text-feature extraction techniques can be proposed with respect to audio and visual data
Poria *et al.* (2015)	Multimodal analysis	SVM, ELM	ISEAR, CK++, eNTERFACE	Obtained 87.95% accuracy	This can be used for real time analysis of uncertain social media data. This can be further improved by considering better features
Liang *et al.* (2015)	To remove noise from Twitter data	Signal-based filtering techniques	20 events from Wikipedia from February 2011 to February 2013	7-10% improvement	This signal-based noise removal from text can be fused with audio and visual noise removal techniques for improving multimodal analysis (Poria *et al.*, 2015)
Almatrafi *et al.* (2015)	Sentiment analysis	Twitter API, Naive Bayes, Python 2.7, NLTK 2.0	650,000 tweets of 5 days. Dataset: V.1.0	BJP tweets are popular than that of AAP	In future, this can be used for real time online analysis instead of storing the data and analysing so that appropriate actions can be taken
Yamada *et al.* (2015)	To extract event information	SVM, CRF, NLCS	Venue-based 23.63 million Twitter tweets November 2013	Precision: 69% complete matches	Data have been extracted using random filters. But better Twitter data extraction rule can be framed in (Li *et al.*, 2012) for relevant information extraction. This may reduce overhead of the proposed system
Gao *et al.* (2015)	Event detection	Harversine formula, ANMRR, MC	Brand-social-net	The proposed method outperforms CR and CLASS SVM	Considering other measures likes those of other parameters from detailed tweet, events can be predicted in future. However, MC is another strong technique for clustering

(*continued*)

Table I.

Table I.

Author year	Research area	Algorithm used	Input dataset	Results obtained	Findings
					similar tweets and can be affective for better research
Xu *et al.* (2015)	Topic-based context-aware travel recommendation	DBSCAN, PLSA, EM, Wunderground API	Geotagged photos in Flickr	For 37 topics, the prediction is 45%, number of similar users among 25 is 53% prediction	Stated that pLSA is better than LDA. So, pLSA is used. However, the accuracy obtained is 45%. This can be improved further by using Hierarchal Dirichlet process (Liu *et al.*) which may perform better than LDA
Memon *et al.* (2015)	Travel recommendation	DBSCAN, different performance measures	1,376,886 photographs of Flickr	Short and long visit can be predicted using popularity based and collaborative filtering, respectively	Instead of using geotagged photos, the social media data from other sources based on multiple modalities can be used to obtain better efficiency. For this multimodal social media analysis techniques can be used as described in (Poria *et al.*, 2015)
Alqhtani *et al.* (2015)	Event detection	TF-IDF HOG, GLCM, SVM, kNN	Twitter data stream 28 December 2014 Indonesia Asia air flight were considered	Accuracy with only text is 0.89, only images is 0.86 and using both is 0.94	Better text-feature extraction techniques can be proposed with respect to audio and visual data
Zhang *et al.* (2015)	Event detection and popularity prediction in microblogging	TSUB, event prediction proposed approach	31 M Twitter posts July 2011 and June 2012 Sina Weibo data January 2012 and June 2012	Both the proposed approaches are better than baseline methods	Topic and events can be considered as separate issues. On one topic, different events can run in future as per specified geo-location
Steiger *et al.* (2015)	Mining tourist location and travel sequences	MAP and DCG, P-DBSCAN	Flickr API, 736,383 geo-tagged photos. Wunderground API	popular ranking = 33% Personalized ranking = 29%	Local follower-following relations among ordinary profiles may have similar kind of opinion about tourist spot. Opinion mining is possible for travel recommendations to identical groups which belong to same location and are connected

topic modelling can be done for bimodal and multimodal in event detection. One of the wide areas used for event detection techniques is semantic analysis which examines the meaning of different tweets and tweets are clustered. Another event detection technique used is topic modelling. In this, tweets are analysed as to which topic they are mostly talking about. Different topics are assigned probabilities as per which topic is discussed most. Both of these techniques can be used as hybrid methodology. Apart from these two, named-entity recognition is another approach for event detection to find similarities. These techniques can be further used for real time analytics in big data platform for social media data. This can be enhanced for different applications multimodal using big data. Further, multiple modalities can be considered including text like microblogging, blogs; images like interest based (PInterest), social images (Flickr), video, gesture recognition, etc. These are some issues which demand more focus and can lead to strong research. It has been believed that research in event recommendation systems is a wide area of research. Event recommendation system may belong to recommending the happenings around in future or which events should be organized in future as per interest of users based on temporal and spatial parameters.

6. Conclusion

In this survey paper, we have been through different approaches of event detection and recommends best practices which can be used for path breaking results. Information regarding instant outbreaks and others can be used to alarm emergency management. However, dataset and input factors vary from one research methodology to another. This paper contains different clustering, categorization and topic modelling approaches used for event detection. Also, we discuss the multivariate research on event detection for multimodal social media. Event recommender and real time big data analytic on social media data in context of event analysis research is recommended in future. On the basis of data acquisition, data analysis and data representation, different event detection approaches are explored. Also, overview of event detection architecture has been introduced and comparative analysis for statistical data of research publications for different social media platforms have been analysed in this paper.

References

Allan, J. (Ed.) (2002), *Topic Detection and Tracking: Event-Based Information Organization*, Kluwer Academic Publisher.

Almatrafi, O., Parack, S. and Chavan, B. (2015), "Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014", *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication. ACM*, p. 41.

Alqhtani, S.M., Luo, S. and Regan, B. (2015), "Fusing text and image for event detection in Twitter", *The International Journal of Multimedia & its Applications (IJMA)*, Vol. 7 No. 1.

Ardon, S., Bagchi, A., Mahanti, A., Ruhela, A., Seth, A., Tripathy, R.M. and Triukose, S. (2013), "Spatio-temporal and events based analysis of topic popularity in twitter", *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM, October*, pp. 219-228.

Becker, H., Naaman, M. and Gravano, L. (2011), "Beyond trending topics: real-world event identification on Twitter", *ICWSM*, Vol. 11, pp. 438-441.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2002), "Latent Dirichlet allocation", *The Journal of machine Learning Research*, Vol. 3, pp. 993-1022.

Bravo-Marquez, F., Mendoza, M. and Poblete, B. (2014), "Meta-level sentiment models for big social data analysis", *Knowledge-Based Systems*, Vol. 69, pp. 86-99.

Danyllo, W.A., Alisson, V.B., Alexandre, N.D., Moacir, L.M.J., Jansepetrus, B.P. and Oliveira, R.F. (2013), "Identifying relevant users and groups in the context of credit analysis based on data from Twitter", *3rd International Conference on Cloud and Green Computing (CGC), IEEE, September*, pp. 587-592.

Ferrari, L., Rosi, A., Mamei, M. and Zambonelli, F. (2011), "Extracting urban patterns from location-based social networks", *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, ACM, November*, pp. 9-16.

Gao, Y., Zhao, S., Yang, Y. and Chua, T.S. (2015), "Multimedia social event detection in microblog", *MultiMedia Modeling*, Springer International Publishing, pp. 269-281.

Jackoway, A., Samet, H. and Sankaranarayanan, J. (2011), "Identification of live news events using Twitter", *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, ACM*, pp. 25-32.

Kaleel, S.B. and Abhari, A. (2015), "Cluster-discovery of Twitter messages for event detection and trending", *Journal of Computational Science*, Vol. 6, pp. 47-57.

Kwon, S., Cha, M., Jung, K., Chen, W. and Wang, Y. (2013), "Prominent features of rumor propagation in online social media", *Data Mining (ICDM), 2013 IEEE 13th International Conference, IEEE*, pp. 1103-1108.

Li, R., Lei, K.H., Khadiwala, R. and Chang, K.C.C. (2012), "Tedas: a twitter-based event detection and analysis system", *IEEE 28th International Conference on Data Engineering (ICDE), IEEE, April*, pp. 1273-1276.

Liang, Y., Caverlee, J. and Cao, C. (2015), "A noise-filtering approach for spatio-temporal event detection in social media", *Advances in Information Retrieval*, Springer International Publishing, pp. 233-244.

Liu, X., Wang, M. and Huet, B. "Event analysis in social multimedia: a survey", *Frontiers of Computer Science*, pp. 1-14.

Mainka, A., Hartmann, S., Stock, W.G. and Peters, I. (2014), "Government and social media: a case study of 31 informational world cities", *47th Hawaii International Conference on System Sciences (HICSS), IEEE, January*, pp. 1715-1724.

Mathioudakis, M. and Koudas, N. (2010), "Twittermonitor: trend detection over the twitter stream", *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, ACM, June*, pp. 1155-1158.

Memon, I., Chen, L., Majid, A., Lv, M., Hussain, I. and Chen, G. (2015), "Travel recommendation using geo-tagged photos in social media for tourist", *Wireless Personal Communications*, Vol. 80 No. 4, pp. 1347-1362.

Nguyen, T., Phung, D., Gupta, S. and Venkatesh, S. (2013), "Extraction of latent patterns and contexts from social honest signals using hierarchical Dirichlet processes", *IEEE International Conference on Pervasive Computing and Communications (PerCom), IEEE, March*, pp. 47-55.

Ozdikis, O., Senkul, P. and Oguztuzun, H. (2012), "Semantic expansion of hashtags for enhanced event detection in Twitter", *Proceedings of the 1st International Workshop on Online Social Systems, VLDB 2012 WOSS, 31 August, Istanbul.*

Poria, S., Cambria, E., Hussain, A. and Huang, G.B. (2015), "Towards an intelligent framework for multimodal affective data analysis", *Neural Networks*, Vol. 63, pp. 104-116.

Steiger, E., Albuquerque, J.P. and Zipf, A. (2015), "An advanced systematic literature review on spatiotemporal analyses of Twitter data", *Transactions in GIS*, Vol. 19 No. 6, pp. 809-834.

Sugitani, T., Shirakawa, M., Hara, T. and Nishio, S. (2013), "Detecting local events by analyzing spatiotemporal locality of tweets", *27th International Conference on Advanced Information Networking and Applications Workshops (WAINA), IEEE, March*, pp. 191-196.

Sun, Y., Fan, H., Bakillah, M. and Zipf, A. (2013), "Road-based travel recommendation using geo-tagged images", *Computers, Environment and Urban Systems*.

Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M. (2006), "Hierarchical Dirichlet processes", *Journal of the American Statistical Association*, Vol. 101 No. 476.

Tripathy, R.M., Sharma, S., Joshi, S., Mehta, S. and Bagchi, A. (2014), "Theme based clustering of tweets", *Proceedings of the 1st IKDD Conference on Data Sciences, ACM, March*, pp. 1-5.

Wang, W. (2015), "Data science for social good – 2014 KDD highlights", *Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin Texas USA, April*.

Weichselbraun, A., Gindl, S. and Scharl, A. (2014), "Enriching semantic knowledge bases for opinion mining in big data applications", *Knowledge-Based Systems*, Vol. 69, pp. 78-85.

Xu, Z., Chen, L. and Chen, G. (2015), "Topic based context-aware travel recommendation method exploiting geotagged photos", *Neurocomputing*, Vol. 155, pp. 99-107.

Yamada, W., Torii, D., Kikuchi, H., Inamura, H., Ochiai, K. and Ohta, K. (2015), "Extracting local event information from micro-blogs for trip planning", *8th International Conference on Mobile Computing and Ubiquitous Networking (ICMU), IEEE, January*, pp. 7-12.

Yang, M.C., Lee, J.T., Lee, S.W. and Rim, H.C. (2012), "Finding interesting posts in twitter based on retweet graph analysis", *Proceedings of the 35th International ACM Sigir Conference on Research and Development in Information Retrieval, ACM, August*, pp. 1073-1074.

Yardi, S. and Boyd, D. (2010), "Tweeting from the town square: measuring geographic local networks", *ICWSM, May*.

Zhang, X., Chen, X., Chen, Y., Wang, S., Li, Z. and Xia, J. (2015), "Event detection and popularity prediction in microblogging", *Neurocomputing*, Vol. 149, pp. 1469-1480.

Further reading

Atefeh, F. and Khreich, W. (2015), "A survey of techniques for event detection in Twitter", *Computational Intelligence*, Vol. 31 No. 1, pp. 132-164.

Chen, M., Mao, S. and Liu, Y. (2014), "Big data: a survey", *Mobile Networks and Applications*, Vol. 19 No. 2, pp. 171-209.

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999), "Data clustering: a review", *ACM Computing Surveys (CSUR)*, Vol. 31 No. 3, pp. 264-323.

Majid, A., Chen, L., Mirza, H.T., Hussain, I. and Chen, G. (2015), "A system for mining interesting tourist locations and travel sequences from public geo-tagged photos", *Data & Knowledge Engineering*, Vol. 95, pp. 66-86.

Nurwidyantoro, A. and Winarko, E. (2013), "Event detection in social media: a survey", *International Conference on ICT for Smart Society (ICISS), IEEE, June*, pp. 1-5.

Corresponding author
Muskan Garg can be contacted at: muskanphd@gmail.com