A Brief Survey on Event Detection Methods for Geo-social Media Data

RICHARD WEN, Ryerson University

In this survey, 5 research papers from 2013 to 2017 were selected from the ACM digital library by using automated and manual criteria. These 5 papers were summarized to present commonly used raw data structures and processed data structures for geo-social media event detection. It was found that frequency-based event detection methods in the selected papers were commonly implemented with statistical measures to determine irregular frequencies in geo-social media data. These irregular location referenced frequencies were determined as events. In order for event detection methods to take advantage of geo-social media data, algorithms must ideally be efficient, scalable, and adaptable. For future work, incorporating video and sound data, integrating web-based knowledge resources, standardizing data structures, and working towards event prediction were suggested to improve and advance event detection methods for geo-social media data.

Additional Key Words and Phrases: survey, review, geo, location, social, media, event, detection

ACM Reference format:
Richard Wen. 2017. A Brief Survey on Event Detection Methods for Geo-social Media Data. N.A., N.A., Article 1 N.A. (April 2017), 9 pages.
https://doi.org/0000001.0000001

1 INTRODUCTION

The wide usage of social media platforms, such as Twitter [3] and Facebook [2], on mobile devices have enabled millions of people to exchange text, images, sound recordings, and videos from any location with wireless internet connection. The Global Positioning System (GPS) in mobile devices further enhances these data with coordinate references that provide information on the location of social media users. Massive amounts of social media data with locational references, referred to as geo-social media data, are then made possible with the integration of GPSes and mobile devices. The massive volumes of geo-social media data are created in real-time from millions of users everyday. This data is capable of providing information on real-world events such as traffic jams, festivals, disasters, and news in near real-time from around the world. Events can be known or predicted before news reports are released for situations in which time-sensitive information is important such as natural disasters or terrorist attacks. However, the data is subject to human errors, noise, changes, and lack of well-defined structures. This survey seeks to provide a review of research papers in which the objective is to create generalized methods for handling data and detecting real-world events from geo-social media sources.

This work fulfills the requirements of the Knowledge Discovery course instructed by Dr. Cherie Ding during Winter 2017 at Ryerson University. This work used the ACM paper template for journals, but was not intended to be submitted as a publication to the ACM digital library. **References (such as journal name, article name, volume number, and issue number) to this work are used for template purposes and should not be cited as a publication**.

XXXX-XXXX/2017/4-ART1 N.A. $15.00
https://doi.org/0000001.0000001

The goal of this survey is to provide a review that summarizes and synthesizes recent event detection methods for geo-social media data. The review is done by using both automated and manual criteria to filter research papers from the Association for Computing Machinery (ACM) digital library. These filtered papers are then selected to be summarized and compared for structures of data, methods, applications, and examples that apply to event detection for geo-social media data. Finally, a discussion of challenges, proposed improvements, and future directions is provided to address the implications of the papers.

The remaining sections are then organized as follows:

- Section 2 details the methods used for paper selection and summarizing selected papers
- Section 3 presents the results of the paper selection and summary of the selected papers
- Section 4 discusses the challenges, proposed improvements, and future directions relative to the selected papers
- Section 5 provides concluding remarks and implications

2 METHODS

This section explains the paper selection requirements and process. A combination of automated and manual criteria was used to discover detailed, relevant, and recent papers for geo-social event detection from an online digital library.

2.1 Search Criteria

This survey was based on a selection of papers from the ACM repository. An automated online search query was used to filter for relevant papers in the ACM digital library. An *advanced search* query was used to filter for published papers matching the following criteria:

(a) **Publication**: Published in an ACM proceeding
(b) **Year**: Published from 2013 to April 17, 2017
(c) **Keywords**: Contains the following words in any common field: *geo*, *location*, *social*, *media*, *event*

2.2 Non-search Criteria

Manual "hand-picked" criteria, referred to as the non-search criteria in this survey, were also used to filter for papers after applying the initial search criteria in Section 2.1. The *year* search criteria, criteria (b), was determined by querying backwards in decrements of a year from the April 17, 2017 until a total of 5 relevant papers with more than 7 pages (indicating full papers) were found. The relevancy and paper length requirement from this process was known as the non-search criteria. Relevant papers were required to have research objectives that sought to identify or predict meaningful events with geo-social media data. Relevant papers were also required to provide methods that were generalizable such that the techniques used were not limited to one application field. For example, the relevant papers were required to be applicable to multiple application fields such as healthcare, journalism, and disaster monitoring. The non-search criteria ensured that the selected papers were *detailed* (a), *recent* (b), and *relevant* (c) to the focus of this survey as defined in the following:

(a) **Detailed**: Paper contained detailed explanation of methods and results with the assumption that full papers (papers with more than 7 pages) had this characteristic
(b) **Recent**: Paper was published within the past 10 years
(c) **Relevant**: Paper had research objectives in which the goal was to detect or predict events using geo-social media data using generalizable methods

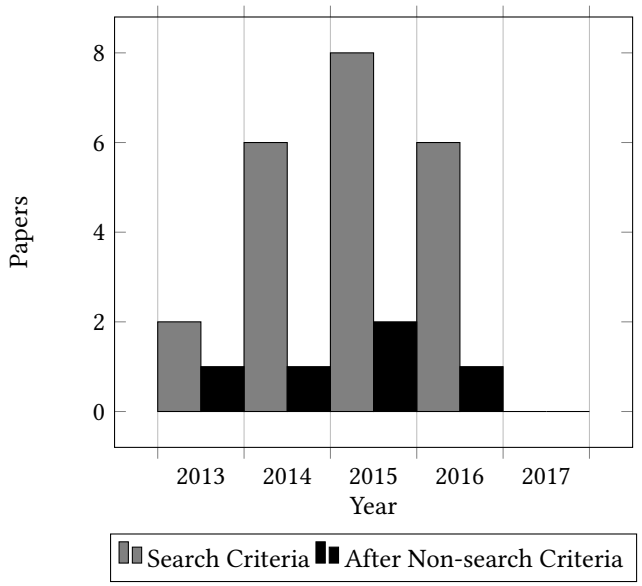

Fig. 1. ACM Published Proceeding Papers Found from 2013 to April 17, 2017.

2.3 Survey Procedure

A loosely-defined survey procedure was applied to review the selected papers filtered from the search criteria (Section 2.1) followed by the non-search criteria (Section 2.2). The survey procedure involved further reviewing the selected papers by:

(1) **Identifying** the methods used for detecting or predicting meaningful events
(2) **Identifying** the data structures appropriate for the methods in (1)
(3) **Identifying** the example applications or conducted experiments
(4) **Grouping** similar methods in (1), data structures in (2), and applications or experiments in (3) into summary sections
(5) **Comparing** the methods, data structures, applications, or experiments contained in each of the summary sections in (4)
(6) **Discussing** challenges, proposed improvements, and future directions relative to the summary sections in (4) after the comparisons in (5)

3 RESULTS

This section provides details on the paper selection results and survey procedure as described in Section 2. Papers from the years 2013 to 2016 were selected to be reviewed for geo-social event detection methods.

3.1 Selected Papers

The papers after applying the search criteria described in Section 2.1 returned 2, 6, 8, 6 and 0 papers for the years 2013, 2014, 2015, 2016, and 2017 respectively. A total of 22 papers from the search criteria were then manually filtered further using the non-search criteria described in Section 2.2. A summary of the papers used in this survey by year is shown in (Figure 1). The papers filtered

Table 1. Geo-social Media Data Structures for Selected Papers.

Structure	Example	Metadata	Source	Papers
Textual	*"Some characters"*	*time-stamp=2017-04-17 00:00:00* *coordinates=00.00,00.00*	Tweets	5
Image	*0 255 1 25* *0 105 1 15* *0 33 1 25*	*timestamp=2017-04-17 00:00:00* *coordinates=00.00,00.00*	User photos	2

from the non-search criteria, after initially applying the search criteria, were used in this survey. These 5 papers (ordered by year and then ordered alphabetically) were authored by:

(1) Shirai, Masaharu, and Hiro (2013) [5]
(2) Abdelhaq and Gertz (2014) [1]
(3) Wang and Kankanhalli (2015) [6]
(4) Wu et al. (2015) [7]
(5) Schubert, Weiler, and Kriegel (2016) [4]

3.2 Summary of Selected Papers

A summary of the selected papers in Section 3.1 was done to provide an overview of event detection methods for geo-social media data. This section provides a briefing of the various data structures, data processing transformations, event detection methods, and applications of geo-social event detection from the selected papers. Each paper used different processed data, but most papers used raw textual data from Twitter, and frequency-based event detection methods.

3.2.1 Data Structures. Geo-social media data were obtained in several different data structures. All 5 of the selected papers utilized textual data with locational reference such as sentences and keywords from social media posts for data filtering and model interpretation. The textual data in the selected papers were obtained as tweets, user posts containing 140 character texts, from Twitter [3] with time-stamps and locational references. For example, the textual data may contain characters *"some character text from a user post"* with metadata containing a time-stamp *"2017-04-17 00:00:00"* (year-month-day hours:minutes:seconds), and coordinates *"0.00, 0.00"* (longitude, latitude) relative to the time and location in which this textual data point was created. Image data from user photos were also used by Shirai et al (2013) [5] and Wang and Kankanhalli [6]. This image data contains pixel intensities arranged in a grid-like structure with time-stamps and coordinates metadata representing time and location references. Wang and Kankanhalli [6] also integrated external sources of data such as security camera feeds containing videos with locational reference with tweets to enhance the interpretation of model results. Table 1 provides a summary of the data structures used in the selected papers.

3.2.2 Data Processing. Various data processing methods were used in the selected papers to handle the loosely defined structures of geo-social media data. Schubert et al. (2016) [4] used grid-based tokenization where important words, known as tokens, were extracted, and coordinates were formatted as tokens, known as geo-tokens, using coordinate grids spaced at every even latitudinal and longitudinal degrees for a study area. For example, a token for *"processing"* is *"process"*, and a geo-token for the coordinates *"0.00, 0.00"* is *"!geo0!0!0"* [1]. The coordinate grids are further expanded to include administrative boundaries such as city or neighborhood boundaries.

[1]The number after *"!geo"* refers to the index of the coordinate grids.

Table 2. Processed Geo-social Media Data Structures for Selected Papers.

Processed	Example	Metadata	Source	Papers
Tokens	*"keyword #tag"*	*timestamp=2017-04-17 00:00*	Tweets	2
Geo-tokens	*"!geo0!0!0"*	*timestamp=2017-04-17 00:00*	Tweets	1
PST	*temp=2017-04-17 00:00* *loc=00.00,00.00* *label=human,parade* *prob=0.8*	*pointer=id*	Tweets User photos	1
Points	(00.00,00.00) (01.00,01.00)	*timestamp=2017-04-17 00:00* *timestamp=2017-04-17 00:01*	Tweets	1

Table 3. Geo-social Event Detection Methods for Selected Papers.

Method	Purpose	Data	Paper
Normalized EWMA	Event detection Real-time	Tokens Geo-tokens	Schubert et al. (2016)
KDE Rel. Score	Event detection Text annotation	Points	Wu et al. (2015)
Entropy Sliding Window	Event detection Word locality	Textual	Abdelhaq and Gertz (2014)
Common Stats Freq. Weights	Event detection Data integration	PST Tokens	Wang and Kankanhalli (2015)
Camera Orient.	Event detection Event boundary	Image Textual	Shirai et al. (2013)

Wang and Kankanhalli [6] used four elements of camera location, temporal information, textual label, and probability that the text labels belong to a topic [2] to form a uniform data structure called Probabilistic Spatio-temporal (PST) data. PST data also involved word tokenization. Wu et al. (2015) [7] used a set of coordinate points with textual data to represent mobility of social media users. For example, a set of coordinate points can be *"(00.00, 00.00) (01.00, 01.00) (02.00, 02.00)"* with each point encased in brackets. The data processing methods sought to unify the raw data structures from Twitter tweets in order to improve consistency and interpretation. Table 2 provides a summary of the processed data structures used in the selected papers.

3.2.3 Event Detection. The authors from the selected papers applied event detection methods using either processed data as described in Section 3.2.2 or raw data as described in Section 3.2.1. Schubert et al. (2016) [4] used an Exponentially Weighted Moving Average (EWMA) with normalization statistics to determine whether an event is detected from tokenized and geo-tokenized data. EWMA was based on the frequency of tokens and geo-tokens, where normalization was used to standardize the EWMA statistical values to measure how unusual an observed frequency is. An event was then detected when the normalized EWMA exceeded a user-set threshold. Wu et al. (2015) [7] used a Kernel Density Estimation (KDE) method to annotate coordinate points containing tweets and time-stamps. KDE was used to determine the density of keywords for localized tweets at particular points in order to assign relevancy scores to each point. KDE was used to avoid the need

[2] A topic is an abstract representation from textual data. For example, the textual data *"Heading to the concert"* may contain a topic named *"concert"* as a more abstract representation of the text.

Table 4. Application of Geo-social Event Detection Methods for Selected Papers.

Application	Method	Data	Paper
Regional events	Normalized EWMA	Tokens Geo-tokens	Schubert et al. (2016)
User profiling	KDE Rel. Score	Points	Wu et al. (2015)
Local events	Entropy Sliding Window	Textual	Abdelhaq and Gertz (2014)
Surveilled events	Common Stats Freq. Weights	PST Tokens	Wang and Kankanhalli (2015)
Tourism Points of interest	Camera Orient.	Image Textual	Shirai et al. (2013)

to set arbitrary distances that represented whether a tweet is near a point or not. An event was then detected by interpreting the top k keywords based on the highest relevancy scores. Abdelhaq and Gertz (2014) [1] used entropy measures and sliding spatial windows for determining whether tweet keywords were localized and unusually frequent in a limited spatial extent. Sliding spatial windows compare recent time-stamped tweets to detect unusually high occurrence of keywords. Wang and Kankanhalli (2015) [6] used PST data to improve the textual interpretation of security cameras by capturing tweets around an area of the security camera locations. Events were detected by statistically measuring the frequency tweets from the PST data around each security camera and connecting the tweets with the security footage for improved inference. Shirai et al. (2013) [5] used image data with camera orientation metadata from mobile phones to capture locational events where large amount of photos were taken. Inward and outward photos (measured from the camera orientation) were used to determine areas and points of interest with boundaries. Frequency-based methods were common among all the selected papers used for geo-social media event detection. Table 3 provides a summary of the event detection methods used in the selected papers.

3.2.4 Applications and Experiments. The papers applied the event detection methods described in Section 3.2.3 to practical use cases and experiments. Schubert et al. (2016) [4] applied event detection methods to discover significant regional events such as New Year's Eve and earthquakes. Wu et al. (2015) [7] applied event detection methods to case studies that geographically profiled social media users at certain locations. Abdelhaq and Gertz (2014) [1] experimented with the event detection of localized events such as fashion shows and sports games in New York, United States. Wang and Kankanhalli (2015) [6] used Twitter tweets and city camera data to detect events with large populations such as parades and music festivals at camera locations. Shirai et al. (2013) [5] experimented with mobile phone photos to detect areas and points of interest and event locations where many images were taken by social media users. The event detection methods performed well for the practical applications, except in the case of earthquakes in [4], where seismic monitoring equipment provides more accurate detection of earthquake events. Table 4 provides a summary of the applications for the event detection methods used in the selected papers.

4 DISCUSSION

This section discusses the challenges, proposed improvements, and future directions for the summary of selected papers in Section 3.2.

4.1 Challenges

The challenges of event detection for geo-social media data was based on the issues of large data volumes, lack of data structure, consistency, and constantly updating data. Geo-social media data was updated by millions of users in real-time during each study in the research papers. For example, Schubert et al. (2016) [4] had to process over 5 million tweets per day, and 3000 to 5000 tweets per minute. This constant flow of data presented challenges in scalability and efficient algorithms that can process and model data in real-time, receiving and analyzing newly produced data from social media users. Schubert et al. (2016) [4] proposed parallelizable and efficient hash tables to update and compare detected events with newly arriving geo-social media data that can be possible events. This hash table implementation allowed an hour of tweets to be processed in 9 seconds on a single-core Central Processing Unit (CPU). Geo-social media data was also inconsistency in the volumes of data located in different parts of the world, and noisy data as a result of human errors, bots, or spam. A large portion of the Twitter data exists in countries in which the platform is more popular, such as the United States or Brazil [4]. The locational reference of geo-social media data can also be inaccurate, where erroneous data such as false coordinates can be given by spammers or bots. Thus, it is important to incorporate normalization methods to account for geographic variations, and data cleaning for erroneous data points. A final challenge was that geo-social media data had a lack of data structure. Different social media platforms may have different data structures, which requires a data processing procedure to unify the data such as the PST data structure in Wang and Kankanhalli (2015) [6]. The main challenges of event detection for geo-social media data emerge from the following data characteristics:

(a) **High Velocity**: Geo-social media data is produced in near real-time by millions of users. Event detection methods must process incoming data updates to take full advantage of the data.
(b) **Large Volume**: Geo-social media data comes in large volumes from all over the world. Event detection methods must be scalable to account for massive volumes of data points.
(c) **Inconsistency**: Geo-social media data are not distributed evenly throughout the world, are not standardized across data sources, and may contain errorneous locational or textual data from human errors, spammers, or bots. Event detection methods must account for uneven data distributions and noisy data.

4.2 Improvements and Future Directions

The geo-social media event detection methods described in Section 3.2.3 used a variety of effective techniques based on the practical applications described in Section 3.2.4. However, the methods did not consider data structures aside from textual and image based geo-social media data. Other media-based data structures such as videos or sound recordings may be possible sources of data that can lead to the improvement in the discovery of potential geo-social events. Events were based on keyword and abstract topics provided by the algorithms, which require human knowledge and interpretation. This may present bias or misinterpretations if the interpreting user is not knowledgeable of events or does not have ground truth data to compare to. Natural language processing and learning methods, together with information available on the internet (such as through Universal Resource Locator (URL) links), may potentially provide better insight and improved interpretation for users that are inexperienced or not knowledgeable of events. A standard data structure for geo-social media data would provide easier comparisons, data integration, and development of general processing frameworks for social media researchers from a variety of different backgrounds. The majority of the selected papers focused on event detection from geo-social media data. Event prediction is another task that could potentially be useful to observe if events, or insight into events, can be

predicted in a time-sensitive manner. This can lead to potentially preventing, instead of mitigating, negative events such as disasters. Thus, future work in the following would advance event detection methods for geo-social media data:

(a) **Video and Sound Data**: Large amounts of video and sound recordings are available from social media platforms. These can be used to improve geo-social media event discovery.
(b) **Web-based Knowledge Resources**: Websites provide a wealth of information that can be used to provide better insight into keywords and topics from event detection methods. This would aid in cases where the interpreting user is not knowledgeable in the events detected.
(c) **Standardized Data**: Standardized geo-social media data processing in event detection can enable easier integration of different social media sources, and allow better base-line comparisons for ground-truth data. Standard ground-truth datasets can then be used to compare across different event detection methods.
(d) **Event Prediction**: Event detection methods focus on identifying irregularities in geo-social media data for events. Predicting these irregularities can be potentially useful for gaining insight into future events and possibly preventing negative outcomes.

5 CONCLUSION

In this survey, research papers were selected from the ACM digital library by using automated and manual criteria. The automated criteria (search criteria) included search queries that limited papers to full paper proceedings, have publication dates from 2013 until 2017, and contain keywords relevant to event detection for geo-social media data. The manual criteria (non-search criteria) included arbitrary qualitative criteria that further limited papers to have greater than 7 pages to indicate detailed writings, have publication dates within the past 10 years to indicate recent writings, and have research objectives in which the goal was to detect or predict events with generalizable methods using geo-social media data to indicate relevancy. Out of the 22 papers found from the search criteria filtering, 5 papers were selected from the non-search criteria filtering to be reviewed for this survey. These 5 papers were summarized to present commonly used raw data structures (textual and image based), and processed data structures (tokens and points) for geo-social media event detection. Frequency-based event detection methods in the selected papers were commonly implemented with statistical measures to determine irregular frequencies in geo-social media data. These irregular location referenced frequencies were determined as events. The event detection methods in the paper were used for a variety of potential practical applications such as locational profiling, tourism, journalism, and disaster response. Event detection methods for geo-social media data need to ideally be efficient, scalable, and adaptable to take advantage of the large amounts of real-time incoming data. For future work, incorporating video and sound data, integrating web-based knowledge resources, standardizing data structures, and working towards event prediction are suggested to improve and advance event detection methods for geo-social media data.

ACKNOWLEDGMENTS

I would like to thank Dr. Cherie Ding for providing the assignment outline, resources, and directions for this survey.

REFERENCES

[1] Hamed Abdelhaq and Michael Gertz. 2014. On the Locality of Keywords in Twitter Streams. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on GeoStreaming (IWGS '14)*. ACM, New York, NY, USA, 12–20. https://doi.org/10.1145/2676552.2676554
[2] Facebook. 2017. Facebook. https://www.facebook.com. (2017). Accessed: 2017-04-24.

[3] Twitter Inc. 2017. Twitter. https://twitter.com. (2017). Accessed: 2017-04-24.

[4] Erich Schubert, Michael Weiler, and Hans-Peter Kriegel. 2016. SPOTHOT: Scalable Detection of Geo-spatial Events in Large Textual Streams. In *Proceedings of the 28th International Conference on Scientific and Statistical Database Management (SSDBM '16)*. ACM, New York, NY, USA, Article 8, 12 pages. https://doi.org/10.1145/2949689.2949699

[5] Motohiro Shirai, Masaharu Hirota, and Hiroshi Ishikawa. 2013. A Method of Area of Interest and Shooting Spot Detection Using Geo-tagged Photographs. In *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place (COMP '13)*. ACM, New York, NY, USA, Article 34, 8 pages. https://doi.org/10.1145/2534848.2534854

[6] Yuhui Wang and Mohan S. Kankanhalli. 2015. Tweeting Cameras for Event Detection. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1231–1241. https://doi.org/10.1145/2736277.2741634

[7] Fei Wu, Zhenhui Li, Wang-Chien Lee, Hongjian Wang, and Zhuojie Huang. 2015. Semantic Annotation of Mobility Data Using Social Media. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1253–1263. https://doi.org/10.1145/2736277.2741675