A Comparative Study on Various Approaches for Event Detection in Social Streams

M. Vijaya Maheswari
PhD Research Scholar, Department of Computer Science
Karpagam University
Coimbatore, Tamilnadu ,India

Dr T. Christopher
Assistant professor, Department of Computer Science
Govt. College of Arts & Science
Coimbatore, Tamilnadu ,India

***Abstract*— Social network sites (SNS) can be regarded as social sensors which can capture a number of events daily in the society. Ex: Face book, Twitter. SNS contain tremendous amount of text, image, audio or video content which can be leveraged for a wide range of business purposes. This paper explores a detailed comparative study about how to identify events in social streams. It also explained various event detection methods and algorithms.**

Keywords*— *Social network services; event; event detection; topic detection and tracking; stream

I. INTRODUCTION

Recent years social media sites (e.g., Twitter, Flickr and Facebook) for users to share their views, personal news and their interests have been significantly increased. In the form of social networking large amount of data, variety of real world events, photographs, videos are shared via internet. It creates more interest for researchers to work out in this area. Social network sites (SNS) can be regarded as social sensors which can capture a number of events daily in the society. Ex: Face book, Twitter .SNS contain tremendous amount of text, image, audio or video content which can be leveraged for a wide range of business purposes.

II. SOCIAL NETWORK SITE

Social network is defined by various persons. According to Barker "Individuals or groups linked by some common bond, shared social status, similar or shared functions, or geographic or cultural connection. Social networks form and discontinue on an ad hoc basis depending on specific need and interest". [R.Barker, The social work dictionary. NASW press Washington, DC, 2013] [1]. According to Boyd and Ellison, a "social network site" is characterized by three functions

(1) These web applications allow users to construct public or semipublic representation of themselves, usually known as user profiles, in a mediated environment.

(2) Such a site provides formal means for users to articulate their relationships with other users (e.g., friend lists), such that the formal articulation typically reflects existing social connections.

(3) Users may examine and "traverse" the articulated relationships in order to explore the space of user profiles (i.e., social graph) [2]. Social networks are categorized as static social networks & dynamic social networks [1].

III. EVENT DEFINITION

An event e is a real world phenomenon that occurred at some specific time t and is usually tied to a location l. [3]. According to social media, event is effects on actual people and how these are reflected in reactions. According to influential theories of emotions events are defined as the users experienced it and urged to externalize their reactions, e.g., tweet about that particular event. The reactions convey the users emotional state how it affects them. [3]

There are two ways to define an event.

- Feature-pivot: an event is a set of related keywords.
- Document -pivot: an event is a set of related tweets. [4]

IV. EVENT DETECTION

Event detection is to identify events in a continuous stream of documents. Topic detection and tracking (TDT) event detection task was studied in notable collective effort to discover and organize various kinds of news events in a continuous stream (e.g., newswire, radio broadcast) [3]. An event is something that occurs in a certain place in a certain time. [5] The online event detection is closely related to the problem of stream clustering and attempts to determine new topical trends in the text stream. The important and news worthy events are captured in the form of temporal bursts of closely related messages in social streams. It can be proposed under both the supervised and unsupervised scenarios. In the unsupervised case, it is assumed that no training data is available in order to direct the event detection process for the stream. In the supervised case, prior data about events is available in order to guide the event detection process.

The key challenges for event detection in social streams are as follows:

(a) The ability to use both the content and the graphical structure of the interactions for event detection process.

(b) The ability to use temporal information in the event detection process. For example, a new trend of closely related text documents from a structural and content point of view, which have not been encountered earlier may correspond to a new event in the stream.

(c) The ability to handle very large and massive volumes of text documents under the one-pass constraint of streaming scenarios.

V. COMPARATIVE STUDY IN RELATED WORKS

A. Duc.T.Nguyen & Jai E.Jung present a social software platform to detect a number of meaningful events from information diffusion patterns on such social network services. The applied method have advantage to detect new topic trend by clustering related message into corresponding categories using content-based methods, temporal information & propagating information of the messages among social community members in runtime.

It opens a challenge issue of detecting which conversation topic trends are discussing and how to cluster incoming messages into a relevant topic categories.

Many news channels & organizations use Twitter daily to post new short message with the purpose of sending information to a wide variety of people about a hot news.

A combination of information from multiple sources can show a completed image about one or a group of facts. Because of huge number of messages, it is difficult for people to follow all events or news, and also do not know what the significant information is.

It is very useful if the application can extract embedded information in its content that will help to reduce time of monitoring and managing news on social networks.

The event can be expressed at the rapidly increasing in a short time range of

- The number of messages described about a fact even if it is the personality opinions.
- The response transactions which includes replying, discussion or sharing action on an original message.
- The frequency occurring of terms including meaningful keywords, named-entities, or phrases.

These messages are collated in runtime on a data pipeline from Stream API (Application programming interface) of SNS sites; the number of in-coming messages at a certain time can be large so that it requires an effective method to process.

DATA CRAWLER

The data stream is temporal, it contains some messages which is fully embedded with rich information about facts and topic trends. Whenever a topic attracts attention of people, a lot of relevant messages will be posted in a short time range around its appearance time.

Topics life cycle is an approximate time range from a timestamp when people start to discuss to the time when people do not talk about it so much. It needs an application which can work as an on-line application to fetch and classify incoming message as quick as possible, so that data service and web-based techniques are suitable to implement it.

They discretize the data stream by using a sampling function with an interval Δ besides that depending on the performance of application system each sampling only collected a number N of messages which the system can process instantly. If messages come so quickly, it can be stored in a cache for next processes. Number of term distributed in a partition will be tracked as a series by time, where time concept is the partition index, the sequence is used to determine a term or term of pairs.

TOPIC TREND DETECTION

Around timestamp of a new topic, related messages will be fed into the system. The tweets contains similar keyword or phrase which has strong mean to topic's content. Besides that, no single keyword or phrase can describe all a fact, it have to be combined together in a complete sentence, hence the co-occurring frequency of each individual term pairs will also be increased.

B. Hassan Sayyadi developed a new event detection algorithm which creates a keyword graph and uses community detection methods analogous to those used for social network analysis to discover and describe events. In this community detection algorithm, nodes can fall into different communities as a word or phrase can be in keywords list of more than one event. Subset of keywords also taken into one community where the number of nodes is large.

New Event Detection (NED) models usually do a single pass incremental clustering algorithm. For a newly arrived document the similarity between the document and known events is computed and the maximum similarity will be selected. If the similarity is more than the predefined threshold, the document will be assigned to the corresponding event, otherwise it will be considered a new event.

PROPOSED ALGORITHM

Both news and events can be represented by keywords and there are several ways in which keywords can be extracted from articles. Nodes are the keywords and edges between the nodes are formed when those terms co-occur in a document.

BUILDING THE KEYGRAPH

A key graph is built by extracting a set of keywords. Then for each keyword k_i we calculate the term frequency (T F $_{i,j}$), document frequency (D F $_i$) and the inverse document frequency (I D F$_i$). Keywords with low document frequency are filtered, and a node (n_i) in the key graph is created for each remaining keyword. Then an edge $e_{i,j}$ between nodes n_i and n_j is be added if k_i and k_j co-occur in the same document. To reduce the noise in the data, each edge should satisfy two conditions.

- An edge is removed if the keywords associated with its nodes co-occur below some minimum threshold.
- The second condition relates to the conditional probability of the edge.

KEYWORD EXTRACTION

Three approaches to extract keywords for each document.

- The algorithm proposed by (Dunning 1993) for keyword extraction.
- Extracting noun phrases as keywords from each document.
- Extracting named entities & noun phrases.

For all three approaches all stop-words are removed and all extracted keywords are stemmed by a porter stemmer. The number of extracted events per day using the three approaches shows that extracting noun phrases and named entities as keywords for documents outperforms the other approaches. The approximated score is obtained by finding the shortest paths between pairs of nodes which are sampled randomly from all possible pairs of nodes.

C. Charu C. Aggarwal & Karthik Subbian present both the supervised and unsupervised case for the event detection problem. They proposed methods for clustering and event detection in social streams. Their result suggest that social streams can be used as a valuable resource to monitor and detect relevant and interesting events in the social stream.

SOCIAL STREAM

A social stream is a continuous and temporal sequence of objects S_1 …..S_r …, such that each object S_i corresponds to a content-based interaction between social entities, and contains explicit content information and linkage information between entities.

SOCIAL STREAM CLUSTERING

A social stream S_1…. S_r … is continuously partitioned into k current clusters C_1 … C_k, such that:

- Each object S_i belongs to at most one of the current clusters C_r.
- The objects are assigned to the different clusters with the use of a similarity function which captures both the content of the interchanged messages, and the dynamic social network structure implied by the different messages.

EVENT DETECTION WITH CLUSTERING

The clustering method can be used directly to perform event detection. The event detection algorithm uses a time horizon H as the input which is used for the event detection process. In order to perform the event detection, monitor the ratio for each cluster continuously over time, and trigger an alarm whenever this ratio exceeds the threshold of α. This suggests a significant change in the underlying social stream, is detected by a significant change in the ratios of stream objects being assigned to the different clusters.

SUPERVISED EVENT DETECTION

If one may want to detect known events which have been encountered earlier in the stream is known as supervised event detection.

To perform supervised event detection, we need to make some changes to the clustering portion of the algorithm. A major change is that it should not allow replacement of old clusters or creation of new clusters when a new incoming point does not naturally fit in any of the cluster. It will always assigned to its closest cluster.

The relative distribution of event specific stream objects to clusters is used as a signature which is specific to the event. These can be used in order to perform the detection in real time.

D. Hila Becker, Mor Naaman & Luis Gravano present a novel approach for identifying events and their associated social media documents, by combining multiple context features of the document. They experimented with ensemble algorithm to tune the clusters for the ensemble, assign weights, and evaluate the final result. Clusters to be homogeneous & include all members of each class in a single cluster. Ensemble should be considered during the selection & weight assignment process is the diversity of the individual clusters.

Consider a set of social media documents where each document is associated with an event. Our goal is to partition this set of documents into clusters such that each cluster corresponds to all documents that are associated with one event.

CLUSTER ENSEMBLES

Ensemble clustering is a clustering approach that combines multiple partitions of a document set. The advantage of using an ensemble approach is in the ability to combine different similarity metrics into the clustering process by learning a weighted similarity normalization technique.

ENSEMBLE SELECTION

The first step in any cluster ensemble algorithm is to select techniques for partitioning the data. It also referred to as clusters, produce mappings from documents to clusters. Each of these techniques should have a unique view of the data, or use a different underlying model to generate the data partition.

ENSEMBLE PREDICTION

In ensemble prediction step, carefully select the clustering threshold for each technique, as well as a confidence weight associated with each technique. Given a set of documents, use each technique to generate a clustering partition of this set.

E. George Valkanas & Dimitrios Gunopulos focused on the problem of automatically identifying events in user-driven, fast paced & voluminous setting. They propose a novel way to address the issue using notions from emotional theories, combined with spatiotemporal information and employ online event detection mechanisms to solve it at large scale in a distributed fashion.

Identifying real life events from social media data is not easy. Some of the challenges are:

- The large adoption means that we must process in real time voluminous amount of data.
- The content is usually short, noisy and diverse in terms of location, languages and topics.
- User location is also a scarce commodity leading to several techniques for location extraction.

The problem of detecting events in a short form messages, focusing on twitter. The main goal is to devise techniques that work regardless of the category the events belong to. We take a novel approach and employ techniques grounded on influential theories of emotions, such as cognitive and affective.

Tweets will not be a flat description of the event, but will also convey the user’s emotional state, partially disclosing how it affected them. An event can then be modeled as a time- and place- related phenomenon, which triggered a significant change in the emotional state of a group of people and goal is to automatically capture such sudden changes.

CONCLUSION

This paper presents a detailed comparative study about how to identify events in social streams. It also explained various event detection methods and algorithms. It represents the outline of the algorithms but not with the experimental results with the user driven events. This paper can be further developed by comparing the event detection algorithms and methods with the same set of user driven events in social streams and their performance can be analyzed.

REFERENCES

[1] Atul S Choudhary , Sunil S Mhamane, "A COMPARATIVE STUDY OF VARIOUS FRAMEWORKS FOR COMMUNITY DETECTION IN DYNAMIC SOCIAL NETWORK",International Journal of Engineering Research & Technology (IJERT), Volume 3 , Issue 4, 2014.

[2] Philip W L Fong , Mohd Anwar , Zhen Zhao, "A Privacy Preservation Model for Facebook- Style Social Network Systems".

[3] George Valkanas, Dimitrios Gunopulos," Event Detection from Social Media Data", IEEE Computer Society Technical Committee on Data Engineering, August 2013.

[4] "Social Listener : An Event Detection Framework Using Social Media".

[5] H.Becker, M.Naaman ,L.Gravano, Event identification in social media, in: Proceedings of the 12th International Workshop on the Web and Databases (WebDB 2009), Providence , Rhode Island, USA, June 28 , 2009.

[6] Charu C.Aggarwal, Karthik Subbian , Event dection in social streams, in:Proceedings of the Twelfth SIAM International Conference on Data Mining, Anaheim, California, USA, April 26-28, 2012, SIAM / Omnipress,2012,pp.624-635.

[7] Duc T. Nguyen, Jai E.Jung , Privacy- preserving Discovery of Topic based Events from Social Sensor Signals :An Experimental Study on Twitter, National Research foundation of Korea, No. 2011-0017156 , pp.749-757, 2014.

[8] H.Sayyadi, M.Hurst, A.Maykov, Event detection and tracking in social streams, Proceedings of the third International Conference on Weblogs and Social Media(ICWSM 2009), San Jose, California, USA,May 17-20, AAAI press, 2009.

[9] Chenyun Dai,Fang-Yu Rao,Traian Marius Truta,Elisa Bertino Privacy-Preserving Assessment of Social Network Data Trustworthiness.

[10] Olivera Grljevic, Zita Bosnjak, Renata Mekovee, Privacy Preservation in Social Network Analysis, Central European Conference on Information and Intelligent Systems, 2012, pp.314-493.

[11] C.C.Aggarwal, Social network data analytics, Springer, 2011.

[12] Junghoon Chae, Dennis Thom, Harald Bosch, Yun Jang, Ross Maciejewski, Spatiotemporal Social Media Analytics for Abnormal Event Detection and Examination using Seasonal-Trend Decomposition, IEEE Conference on Visual Analytics Science and Technology 2012.

[13] J.J.Jung,Cross-lingual Query Expansion in Multilingual Folksonomies: a Case Study on Flickr, Knowledge – Based Systems, Vol.42,pp:60-67,2013.