A Survey on Content based Semantic Relations in Tweets

Alby Thomas
Dept. of CSE
College Of Engineering, Poonjar

Sindhu L.
Dept. of CSE
College Of Engineering,Poonjar

ABSTRACT

Twitter is a popular micro blogging site, in which people share their ideas and views with others. Because of different writing conventions and character restriction there may be variation in the impact for the same event. Analyzing semantic relationship and analyze the variations have several use cases such as event detect etc. there are several technique are available for event detection and term similarity analysis using semantic information of tweets. This survey paper aims to highlight these techniques.

General Terms

Algorithm, Classification, Database, system,

Keywords

Event detection, clustering, term co-occurrence, semantics, event clusters, term similarity.

1. INTRODUCTION

Micro blogging is a way of communication in which user can share their opinions in short form such as Twitter, it provides a easy sharing of messages either publically or within the social networking platform. This social networking medium is not only for socializing with friends, but also for assemblage of information about the current world and whatever subject one might be interested in. People post short messages, interesting information from daily news, updates the current hot topics, and their views, opinions on a certain subjects, developments in their profession, and materials about their interests.

It is a real-time information network with its millions of users around the world and millions of tweets per day1. The information in Twitter is an abundant and analyzing of this information useful for information retrieval. This retrieved information is very useful for analyzing the real world trends such as analyzing public trends, detecting important events, tracking public opinion. But twitter is a free and unsupervised medium; there may be variation in the content of tweets, even if the same concept has to be referred. People can user their own way to express their idea about the same thing. Geographic, cultural and language difference may be cause variation in content. Restriction on the tweet content is an important problem; their character length should be less than 140, this led the users to write in a compact form. Additional information can be given in a tweet by using special characters such as abbreviations or symbols. Spelling mistakes are also important factor for variation in content. For these reasons, to extract semantic relationship among the tweets, it is necessary to apply information retrieval algorithm.

Twitter is a real time real social networking site, which allows users to publish short tweets about “what’s happening”. Real-life events are reported in Twitter. Detecting those events to have a better understanding of what users are really discussed. Event detection has long been a research topic. For event detection various methods are exploited, such as word co-occurance, term co-occurance. These methods also work on the information getting from semantic relationship analysis.

1.1 Semantic Social Network Analysis

Internet provides many ways to interacting between us informative huge social network structures. Social Network Analysis (SNA) is the process of understand the features and make use of the key features of social networks in order to manage their life cycle and predict their evolution. These networks are in which vertices represent users, and edges represent social relations such as closeness and co authorship links among users. Social network analysis is the study of social networks by understanding their social entities, the people and their relationships. The centrality focuses the most important actors of the network and three definitions have been proposed, the degree centrality is based on the average length of the number of adjacent edges. The closeness centrality is based on the average length of the number of edge connecting a node to others and reveals the capacity of a node to be reached. The betweenness centrality focus on the capacity of a node to be an intermediary between any two other nodes.

Twitter is an online social networking site, with nearly 600 millions of users and over 250 million messages per day. The textual content of twitter is plentiful and increasing day by day. We cannot validate the content of twitter or because of the limitation of character and the other writing conventions. Identifying entities, and moreover relating entities, via Tweets is a non-trivial problem. In order to help out the identification of entities within posts, we augment Tweets by mapping them to associated news articles. We then apply different methods to identify semantic relationships between entities within Tweets as well as news. Our main focus is directed at determining whether relations can be discovered from Twitter messages or whether some improvement is needed, rather than finding the optimal approach for relation learning. Furthermore, we validate the discovered relations over time in order to distinguish which entities are related at a given time, and thus provide the most interesting and relevant content at the right time. Sentiment analysis over Twitter offer organizations a fast and effective means to monitor the publics’ feelings towards their brand, business, directors etc.

1.2 Event Detection

Events can be defined as real-world occurrences that spread out over space and time. Social events are attended by people and presented in the multimedia content that is shared through social network websites. Eg: disasters, concerts, sports events, public merriments and protests amongst others. Many people use Twitter for knowing what happening in the real world, such as real world events. Twitter and social media trends have significantly changed in the recent past with millions of users use this platform for chat, exchange their ideas or share stories etc. therefore this platform has formed a rich place for news, events and information mining. Users would be interested in getting advice, opinions, facts, or updates on news or events. However, due to the enormous explode in information data mining in Twitter is a convoluted. Due to the traffic overflow of tweeter information's are multiple and huge in terms of the frequency. For instance this platform receives over 80 million tweets per day and we got billions of tweets per month. So, event prediction and detection requires the use of complex algorithms.

1.3 Term Similarity Analysis

Natural language sentence contains both semantic and grammatical term, that is, a term can be synonym, antonym, or hyponym of another term. In human or natural language cases we can use dictionary, Thesauri or encyclopedia for find these relationships. But in the case of internet we can use online resources such as WordNet and Wikipedia instead of dictionaries etc.. These online recourses are not mature enough for online social networking platforms, especially for Twitter. The language of Twitter is different than the one in WordNet or other resources. Because it is a social networking platform the contents are unauthenticated that is there is no authority to check the contents of tweets. The next problem is the limitation of characters of tweets, to fulfill this condition the users write posts in their own way, that is it may contain spelling mistakes or they use words in their own style. To overcome these problems by using a statistic-based technique to identify the similarity of terms.

There are several relationship metrics depending on which stastical pattern in term distribution to look for. In this survey we consider first and second order relation metrics. First order relationship metric is used to identify the frequently co-occur term pairs. Second order relationship metric is used to identify the term co-occurrences.

The ultimate aim of this survey is to study different technique for event detection methods and term similarity analysis.

2. RELATED WORKS

Here we are discussing different techniques of event detection and term similarity analysis of Twitter.

2.1 Event detection techniques

Maximilian Walther.et.al[1] describe a novel algorithm for geo-spatial event detection on twitter. In this work the authors first analyzing all the tweets from specific location in a specific time for identify the place with high occurring of events. In the next step a spatio-temporal clustering is applying in the posts with a machine learning component to detect whether any real world events are constitute or not. Finally detected events are displayed in a map. This system shows high accuracy in detecting this temporally and spatially clustering tweets are real world events are not. The candidate tweets are not all the events in twitter, these events picked by the authers. The systems uses MongoDB database because it contains spatio- real time processing tweets.

Takeshi sakaki.et.al[2] construct an earthquake reporting system in Japan. In this system detect the real time event such as earthquake by observing the tweets. In this system use a classifier of tweets based on features to detect the target event. Then use a probalistic spatio-temporal model for find the centre and trajectory of the event location. The authors use kalaman filtering and paricle filtering for the estimation of event location. This system detect earthquake promptly and send emails to users which is faster than announcement. This system usem use SVM classifier for learning purpose. For event location estimation it uses kalaman filter and paricle filter. This system takes only single instance of the target event exists.

Jianshu Weng.el.at[3] propose a method to detect real time events from text streams such as twitter. In work the authors introduce an event detection algorithm EDCoW (Event Detection with Clustering of Wavelet-based Signals). This system use wavelet analysis on the frequency based raw signals of the words for creating signal of individual words. In the next step applying autocorrelation on this resulting signals to filter these signals for avoid the trivial words. The final step is the detection of events by clustering the remaining words using a modularity-based graph partitioning technique. This method is scalable. It treat each word independably so it fail to identify semantic words. And also it does not exploit relationship among users.

Sasa petrovic.et.al[4] present first story detection (event detection) algorithm based on Locality-Sensitive Hashing(LSH). LSH is random technique, that reduce the needed to find a nearest neighbors in vector space, and also it save the space by reduce the amount of stories in memory constant. The limitation of applying pure LSH to the Twitter stream is the poor performance and variation in the expected results. To improve the performance of this algorithm the authors introduce a modification of this technique. The modification of LSH not store all previous data (stories) in the main memory nor compare the new document to all the documents returned by LSH. The goal of this proposed system is to automatically detect significant events, preferably with a minimal number of an non- important events. Here the authors use threads of tweets for analysis.

Hamed Abdelhaq.et.al[5] present a system "Even Tweet" ,to detect localized events by exploiting the location information of georeferenced messages. In this work the authors introduce a novel frame work to detect localized events in real time from a twitter stream and to track the evolution of events over time. To identify meaningful candidate for event descriptions using the extracted information about the spatio-temporal characteristics of keywords. Then clustered these keywords to get the information about the localized events. The authors use a scoring scheme to determine the most important events in a time frame. The proposed system focuses on detecting localized real time events by using continuous analysis of most recent tweets within a time sliding window. This system also track the evolution of events over time by using scoring scheme, which gives a score for each event. This score is acting as an indicator of its significance over time. Even Tweet detecting localized events in real time by continuous analysis.

Xun Wang.et.al[6] propose a model for bursty word extraction in Twitter by using a mixture of Gaussian model. And also propose a model for new topic detection by using a novel time-dependent HDP. In this proposed system there are three step for event detection. In the first step select the bursty word candidate from huge tweets stream by using a Gaussian mixture. In the second step using an evolutionary clustering concepts , which focuses on detecting dynamics of a given topic. For this purpose authors develop a novel time dependent HDP(td-HDP) for new event detection. This model based on the assumption that the data of Twitter forms a Markovian chain. The final step is the detection of location of event by using CRF algorithm. This model detect the location and time of an event accurately and promptly. The limitation is the analysis restricted in a particular time frame.

Long R.et.al[7] propose a traditional clustering approach by integrating some specific features to the characteristics of Twitter. These features are based on events. Events are extracted from daily messages on the basis of word frequency, word occurrence in hash tag and word entropy. A top down divisive approach clustering approach is employed on a co-occurrence graph which connecting messages in which topical words co-occur to divide topical words into event topical into event clusters. To create event chains use a maximum-weighted bipartite graph. A Jaccard coefficient is used for similarity measures between clusters. Finally to detect the top event by using cosine similarity augmented with a time interval between messages. Then the event summaries are plotted on a time line to link the event chain clusters. This approach has high precision value.

Ozer ozdikis.et.al[9] present an event detection method in Twitter, based on the clustering of hashtags and introduce an enhancement technique by using the semantic similarities between the hash tags. To achieve this goal the authors introduce a system that contains two methods. First method for generating tweet vector and the second one for evaluating their effect on clustering event detection performance. To identify the paradigmatic relationships and similarities by analysing the context of hash tags and their co-occurrence statistics with other words. The authors apply a lexico-semantic expansion on tweet contents before clustering the tweets based on their similarity. Agglomerative text clustering technique is used for event detection. In this method clustering is performed on the basis of similarity in vector space model. In this clustering values in tweet vectors that is, weights of the corresponding terms for each vectors such as TF-IDF values.

Ozer Ozdikis.et.al[10]propose an event detection methods in a micro-blogging platform, such as Twitter. The enhancement technique the authors propose the method based on lexico-semantic expansion of tweet contents while applying document similarity and clustering algorithms. Considering the length limitations and idiosyncratic spelling in Twitter environment, it is possible to take advantage of word similarities and to enrich texts with similar words. The semantic expansion technique we implement is based on syntagmatic and paradigmatic relationships between words, extracted from their co-occurrence statistics. The semantic expansion technique we implement is based on syntagmatic and paradigmatic relationships between words, extracted from their co-occurrence statistics. As our technique does not depend on an existing ontology or a lexicon database such as WordNet, it should be applicable for any language.

Table 1: comparison of event detection methods

Author	**Publication Year**	**Method**	**Database**	**Advantages/Disadvantages**
Maximilian Walther and Michael Kaisser[1]	2013	Sptio-temporal clustering Manually perform machine learning	MongoDB	High precision and recall value Applicable only for small scale and localized events in a specific time period.
Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo[2]	2010	Probalistic-spatio temporal word(event location) Kalaman filtering andparticle filtering	Online tweets about earthquake	Faster than JMA Applicable only for single instance of the target events exists.
Jianshu Weng, Yuxia Yao, Erwin Leonardi, and Bu-Sung Lee[3]	2011	EDCoW Modularity based graph partitioning.	1000 Singapore-based *Twitter* users	Scalable Applicable only to small data set.
Victor Lavrenko,	2010	Locality Sensitive Hashing (LSH)	TDT5 db	Constant space, processing time and good results

Miles Osborn,Saˇsa Petroviˊ [4]				Loss in performance in the case of large scale datas
Hamed Abdelhaq, Christian Sengstock, and Michael Gertz[5]	2013	Cosine similarity(clustering algorithm) Spatial signature Temporal resolution	Twitter API	Detecting localized events and evolution of events. Consider events within a limited time period.
Xun Wang, Feida Zhu, Jing Jiang2, Sujian Li1[6]	2013	CRF algorithm Gaussian mixture Novel time dependent HDP	Tweets of Singapore in one year	Accurate results. Restricted in a particular time frames.
long, r., h.wang, y. chen, o. jin, and y. yu.[7]	2011	Hierarchical divisive clustering	Twitter API	High performance Restricted to small scale data
Ozer Ozdikis, Pinar Senkul, Halit Oguztuzun[8]	2012	Agglomerative clustering based on word based cooccurance	Twitter API	High accuracy in performance Restricted to small collection of tweets
Ozer Ozdikis, Pinar Senkul, Halit Oguztuzun[9]	2012	Agglomerative clustering based on hash taging	Twitter API	High accuracy in performance Restricted to small collection of tweets

2.2 Term similarity techniques

Taneeya Satyapanich.et.al[10] develop a system UMBC-STS system for Paraphrase and Semantic Similarity in Twitter. In this system, create similarity vectors from two-skip trigrams of preprocessed tweets and measure their semantic similarity. Then it computed the statistical value as maximum and average of each pair and use two regression models; logistic regression and support vector regression. This method decide whether two tweets are paraphrases or not, we use a measurement based on semantic similarity values. If two tweets are semantically similar, they are judged as paraphrases, otherwise they are not. Best result of this system performance is run achieved an F1 score of 0.599 and was ranked eleventh out of eighteen teams.

Reza Bosagh Zadeh.et.al[11] present a suite of algorithms for Dimension Independent Similarity Computation (DISCO)to compute all pair wise similarities between very high-dimensional sparse vectors. The results are independent of dimension and geared toward the MapReduce framework. This algorithm uses Cosine, Dice, Overlap, and the Jaccard similarity measures. The MapReduce and Streaming computation models can be generalized to other frameworks because it shuffle size and memory, the sampling strategy we use. This sampling strategy to be useful to computing anumber between 0 and 1 by taking the ratio of an unknown number to some known number.

Anlei Dong.et.al[12] propose a method to use the micro-blogging data stream to detect fresh URLs and also use micro-blogging data to compute novel and effective features for ranking fresh URLs such as recency sensitive queries. This approach is based on preserving the quality of data presented to the general web searcher by only using micro-blog data as evidence for discovering and ranking URLs. The results demonstrate the power of leveraging widespread user behavior for recency sensitive queries. Although other sources of user behavior information exist, Twitter is one of the only sources which is both public and widely adopted. This makes Twitter a valuable source of real time user behavior for institutions lacking access to more sensitive log data.

Table 2: comparison of term similarity analysis

Autor	Publication Year	Method	Database	Advantages/Disadvantages
Taneeya Satyapanich, Hang Gao and Tim Finin[10]	2015	LSA word similarity measure SVM	Twitter API	High accurate semantic similarity measurement.
Reza Bosagh Zadeh, Ashish Goel [11]	2010	Dimension Independent Similarity Computation(DISC)	Online tweets about earthquake	Higher similarity pairs are estimated with provably better accuracy.
Anlei Dong Ruiqiang Zhang Pranam Kolari Jing Bai.et.al[12]	2010	RankSVM, RankNet RankBoost, GBrank	Twitter API	Detect fresh URLs. Failed to detect spam

3. CONCLUSION

From this survey paper discussing about different event detection techniques. Each of these methods has different criteria in event detection and find term similarity. It can be conclude that for event detection agglomerative clustering applied on tweets collected in a given period of time may produce better results. In this algorithm we use tweet vectors that are generated TF-IDF value of the term in each tweet. This method detect events with high accuracy than other models, which are discussed here. And also term similarity scores are more better to extract the semantic relation among tweets instead of using word co-occurrence.

4. REFERENCES

[1] Maximilian Walther and Michael Kaisser, , ‘Geo-spatial Event Detection in the Twitter Stream’, Springer Verlag Berlin Heidelberg 2013

[2] Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo, ‘Earthquake Shakes Twitter Users:Real-time Event Detection by Social Sensors’, April 26-30, 2010, Raleigh, North Carolina. 2010

[3] Jianshu Weng, Yuxia Yao, Erwin Leonardi, and Bu-Sung Lee, ‘Event Detection in Twitter’, HP Laboratories HPL-2011-98, 2010

[4] Victor Lavrenko, Miles Osborn,Saˇsa Petrovi, ‘Streaming First Story Detection with application to Twitter’, 2010

[5] Gertz , ‘EvenTweet: Online Localized Event Hamed Abdelhaq, Christian Sengstock, and Michael Gertz ‘EvenTweet: Online Localized Event Detection from Twitter’,Proceedings of the VLDB Endowment, Vol. 6, No. 12 2013

[6] Xun Wang, Feida Zhu, Jing Jiang, Sujian Li, ‘Real Time Event Detection in Twitter’, Springer-Verlag Berlin Heidelberg , 2011

[7] long, r., h.wang, y. chen, o. jin, and y. yu, ‘Towards effective event detection, tracking and summarizatioon microblog data’, Web-Age Information Management, Vol. 6897 of Lecture Notes in Computer Science, 2011

[8] Ozer Ozdikis, Pinar Senkul, Halit Oguztuzun, ‘Semantic Expansion of Hashtags for Enhanced Event Detection in Twitter’, VLDB 2012 WOSS, August 31, 2012, Istanbul, Turkey 2012

[9] Ozer Ozdikis, Pinar Senkul, Halit Oguztuzun, ‘Semantic Expansion of Tweet Contents for Enhanced Event Detection in Twitter’, 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2012

[10] Taneeya Satyapanich, Hang Gao and Tim Finin, ‘Ebiquity: Paraphrase and Semantic Similarity in Twitter using Skipgram’, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 51–55, 2015

[11] Reza Bosagh Zadeh, Ashish Goel , ‘Dimension Independent Similarity Computation’, Journal of Machine Learning Research, 2012

[12] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, ‘Time is of the Essence: Improving Recency Ranking Using Twitter Data’, International World Wide Web Conference Committee(IW3C2),2010