International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume IV, Issue VIII, August - 2017

Contextual Relationship based Attribute and Event Detection using Social Media Data Analysis

Ms. P. Saranya, MPhil Research Scholar,
Mrs. M. Preetha, MCA, MPhil., Assistant Professor,
PGP College of Arts and Science., Namakkal, Tamilnadu, India,

Abstract

The social networks are the medium used to share the views and opinions of the people. The social media interactions are carried out from anywhere and anytime. The location and time information are collected and maintained under the Location Based Social Networks (LBSN). The social media data analytics models are used to discover the user behaviors, product reviews and events. The health, pollution and social unrest events are discovered using the social media data analysis. The spatial and temporal parameters can also used in the social media data analysis.

The social network produces huge volume of data values. The feature learning and event discovery models are processed sequentially. The concurrent learning models are applied in the Multi Task Learning (MTL) framework. The Constrained Multi Task Feature Learning (CMTFL) scheme performs the feature learning and event detection simultaneously on each spatial points. The static features and dynamic features are used in the feature learning and event discovery process. The iterative group hard thresholding model is adapted for the event discovery process.

The feature learning and event discovery process is enhanced with user domain knowledge based data analysis models. The contextual information is employed to indicate the location and time details for the social network users. The Hybrid Multi Task Feature Learning (HMTFL) scheme is build with spatio temporal relationship based feature learning and event discovery process. The spatial relationships are analyzed to group the spatial points with proximity information. The temporal correlations are adapted to discover the events based on the time relationships. The user -venue relationship and venue user relationship based index model is used to improve the accuracy level of the event discovery process.

Index Terms: Social media data analysis, Multi Task Learning, Feature learning and event detection and Spatio temporal mining

1. Introduction

Social networks are phenomena of these days. Increasing number of users and personal character of data leads to the problem with security and unwilling loss of privacy. The advantages of using social networks are in free and fast communication with friends usually in a form of objects such as tweets, pictures, videos and texts Another feature is creation of networks – friends, colleagues, family members. The problem comes with the possibility to trace activities, relations and communication together with limited possibility to manage critical activities such as delete of data. The information managed by different kinds of social networks is very interesting mostly for its potential to form overall personal profile.

Social network analysis (SNA) is the process of investigating social structures through the use of networks and graph theory. It characterizes networked structures in terms of nodes and the ties, edges, or links connect them. Examples of social structures commonly visualized through social network analysis include social media networks, memes spread, friendship and acquaintance networks, collaboration graphs, kinship, disease transmission and sexual relationships. These networks are often visualized through sociograms in which nodes are represented as points and ties are represented as lines. Social network analysis has emerged as a key technique in modern sociology. It has also gained a significant following in anthropology, biology, demography, communication studies, economics, geography, history, information science, organizational studies, political science, social psychology, development studies and sociolinguistics and computer science and is now commonly available as a consumer tool.

Social network analysis is used extensively in a wide range of applications and disciplines. Some common network analysis applications include data aggregation and mining, network propagation modeling, network modeling and sampling, user attribute and behavior analysis, community maintained resource support, location-based interaction analysis, social sharing and filtering, recommender systems development and link prediction and entity resolution. In the private sector, businesses use social network analysis to support activities such as customer interaction and analysis, information system development analysis, marketing and business intelligence needs. Some public sector uses include development of leader engagement strategies, analysis of individual and group engagement and media use and community – based problem solving.

Social network analysis is also used in intelligence, counter-intelligence and law enforcement activities. This technique allows the analysts to map a clandestine or covert organization such as a espionage ring, an organized crime family or a street gang. The National Security Agency (NSA) uses its clandestine mass electronic surveillance programs to generate the data needed to perform this type of analysis on terrorist cells and other networks deemed relevant to national security. The NSA looks up to three nodes deep during this network analysis. After the initial mapping of the social network is complete, analysis is performed to determine the structure of the network and determine. The NSA has been performing social network analysis on call detail records (CDRs), also known as metadata.

Large textual corpora can be turned into networks and then analyzed with the method of social network analysis. In these networks, the nodes are Social Actors and the links are Actions. The ex traction of these networks can be automated, by using parsers. The resulting networks, which can contain thousands of nodes, are then analyzed by using tools from network theory to identify the key actors, the key communities or parties and general properties such as robustness or structural stability of the overall network, or centrality of certain nodes. This automates the approach introduced by Quantitative Narrative Analysis, whereby subject-verb-object triplets are identified with pairs of actors linked by an action, or pairs formed by actor-object.

In computer-supported collaborative learning One of the most current methods of the application of SNA is to the study of computer-supported collaborative learning (CSCL). When applied to CSCL, SNA is used to help understand how learners collaborate in terms of amount, frequency and length, as well as the quality, topic and strategies of communication. Additionally, SNA can focus on specific aspects of the network connection, or the entire network as a whole. It uses graphical representations, written representations and data representations to help ex amine the connections within a CSCL network. When applying SNA to a CSCL environment the interactions of the participants are treated as a social network. The focus of the analysis is on the "connections" made among the participants – how they interact and communicate – as opposed to how each participant behaved on his or her own.

2. Related Work

There are several threads of related work of this paper. Query expansion in microblogs retrieval. Query expansion is a process that reformulates the seed query in order to improve the coverage and accuracy of information retrieval. To improve the performance of retrieval in Twitter, a new thread of work utilizes query expansion to dynamically expand keywords retrieve tweets [11] and discover events [2]. The expanded keywords are typically extracted by exploring their co-occurrence with the user-specified initial query in textual content, but information diffusion through social network has not been comprehensively explored. Event detection in Twitter. There exists a large amount of work on event detection in Twitter. Event detection methods utilize supervised or unsupervised framework to extract tweets subset related to potential events that can be formalized as spatial burstiness [13], [2], temporal burstiness [4], [12], or spatiotemporal burstiness [7], [3]. This thread of work has a different goal from our paper: it detects the emergence instead of the evolution of events, whereas our paper focuses on continuously tracking the evolutionary dynamics of a theme.

General theme tracking. A considerable body of work focuses on characterizing the general pattern of Twitter streams. The pattern is typically conceptualized as a mixture of "latent topics". For example, Blei *et al.* aligned the proportion priors and distributions of latent topics over time [10]. Yang *et al.* proposed an efficient Twitter stream summarization approach that can fit in a limited memory [1]. Hong *et al.* analyzed the inter-relationships of multiple social media streams by considering both local topics and shared topics [14]. Mei and Zhai modeled latent topics through a mixture language model, and discovered the transitions among them [5]. Because "latent topics" are typically extracted purely statistically based on data without human prior knowledge, they do not necessarily have real563 world meaning. Hence this thread of work is generally not appropriate to track targeted themes.

Targeted theme tracking. A thread of work focuses on tracking targeted themes, such as earthquakes [15]. The majority of research adopts classification framework to extract themerelated tweets based on contextual features only [6]. Hence, it is challenging to select an appropriate set of features. Li *et al.* proposed a generic framework for themerelated feature selection, whereas this approach is specially designed for scrawling two specific types of Twitter APIs and is not appropriate for the task of this paper [9]. A handful of methods have been proposed to take into account social relationships. Lin *et al.* implemented a probabilistic mixture model to characterize the temporal textual pattern and diffusion via friendship [10]. Ratkiewicz *et al.* applied a framework specifically designed to track the so-called "political astroturf" based on mentioning networks [8].

3. Multitask Learning on Social Networks

Microblogs such as Twitter and Weibo are experiencing an explosive level of growth. Millions of microblog users across the world broadcast their daily observations on an enormous variety of topics, such as crime, sports and politics. The spatial event forecasting from microblogs are focused for events such as civil unrest, disease outbreaks and crime hotspots. The approach searches for subtle patterns in specific cities that serve as indicators of ongoing or future events, where each pattern is a burst of context features relevant to a specific event. For instance, expressions of discontent about gas price increases could be a potential precursor to a protest about government policies.

Three technical challenges must be overcome when addressing this problem: 1) Dynamic features. The language used in microblogs is highly informal, ungrammatical and dynamic. Most existing methods treat fixed keywords as features, but expressions in tweets may dynamically evolve, rendering the use of fixed features and historical training data insufficient. For example, the most significant Twitter keyword for the Mexican protests in Aug 2012 was "#YoSoy132", alluding to the protests against the Mexican presidential election, but "#CNTE" had become the most popular term by the beginning of 2013 due to the growing resistance to Mexican education reform. Ideally, an event forecasting system must combine the judicious use of static features with an awareness of subtle changes involving dynamic features. 2) Geographic heterogeneity**.** Existing models usually build a single predictive model for all the different locations. Different cities have different characteristics, such as population, weather and administrative structures. As a result, it is difficult to impute basal levels of occurrence uniformly. Considering civil unrest as an example, finding 1000 tweets mentioning the keyword "protest" is not likely to be a strong indicator of an upcoming civil unrest event in a city with a population of a few million users but could be a strong signal in a much smaller city with a population of only 10,000. To consider the geographical heterogeneity, some works adapted to establish the corresponding model for each different location separately. But because each model only utilizes the data of its corresponding location, the data scarcity problem is a serious challenge that degrades the model performance and generalization. 3) Scalability. Spatiotemporal event forecasting in social media streams prefers real-time framework and has emphasis on computation efficiency. The efficiency is challenged by the huge scale of the data, including (1) High dimension features to characterize the rich text and network information; (2) large number of time points; and (3) heterogeneity in enormous geo-locations. This means that even a

medium-scale problem that contains 1000 keywords, 1000 dates and 1000 locations will involve at least 1 billion data points in the optimization computation. Some scalable forecasting methods are desired for this problem.

In order to concurrently address all these technical challenges, this work presents a novel computational approach in the form of a framework of multi task learning (MTL) that combines the strengths of methods that use static features and those that use dynamic features. In individually, for event forecasting, but the system tackles the challenges involved in unifying these contrasting approaches in a single framework. Learning multiple related tasks simultaneously effectively increases the sample size for each location, thus potentially also improving the forecasting performance, especially when the sample size for each task is small. One critical issue in multi-task learning is how to define and exploit the commonality among different tasks. Intuitively, events that occur around the same time may involve similar topics and therefore tweets from different cities may share many common keywords that are related to the event(s). The issues are focused on four multi-task feature learning (MTFL) formulations for event forecasting that differ in the specifics of how common features are extracted. Formulation of a multi-task learning framework for event forecasting. Here, event forecasting for multiple cities/states in the same country are treated as a multitask learning problem. Event forecasting models are build for different cities/states simultaneously by restricting all cities/states to select a common set of features with different weights exclusive to corresponding tasks. The system explores both penalized and constrained MTL formulations, applying 4 different strategies to control the common set of features selected. Concurrent modeling of static and dynamic terms. The existing models (LASSO and DQE) use different but complementary information: LASSO uses static terms, while DQE identifies dynamic terms. The MTL formulations make use of both types of information by integrating the strengths of LASSO into DQE.

Development of efficient algorithms is to both convex and non-convex optimization formulations are explored. For convex problems, the system employs proximal methods, such as FISTA have been shown to be efficient for solving sparse and multi-task learning problems. The iterative Group Hard Thresholding (IGHT) framework is guaranteed to converge to a local solution.

Two different Twitter datasets are used to the evaluation method: the Latin America civil unrest dataset and the United States influenza outbreaks dataset. The methods consistently outperformed the competing methods, namely LASSO, DQE, traditional multitask learning models and their variants. The sensitivity analyses is also performed to reveal the impact of the parameters on the performance methods. Multiple case studies are provided to demonstrate the utility of the method in practical applications.

4. Issues on Multi Task Learning Schemes

The social media data analysis operations are carried out to discover the events. The spatial event forecasting is a complex task in social media data analysis process. The dynamic features and geographic heterogeneity factors suffers the event forecasting operations. The spatial correlations, imbalanced samples and different population in different locations parameters are the spatial heterogeneity measures. The LASSO regression, dynamic query expansion and burst detection methods are applied for the spatial event detection process. The Multi Task Learning Framework (MTLF) is build to solve the dynamic patterns of features and geographic heterogeneity problems. Simultaneous event forecasting for all the locations is carried out under the multi task learning model. The data values for each location are passed and analyzed in the simultaneous learning process. The static features and dynamic features are used in the event forecasting model. The static features are derived from the predefined vocabulary prepared by domain experts. The dynamic features are fetched from the dynamic query expansion process. The iterative group hard thresholding based algorithms are used in the training and prediction operations. The Constrained Multi Task Feature Learning (CMTFL) algorithm is applied for the feature learning operations. The following issuses are identified from the existing system.

- Location relationships are not focused
- Human domain knowledge utilization is low
- Stand alone event discovery model
- Temporal relationships are not analyzed in the training process

5. Contextual Relationship based Attribute and Event Detection

The social network data analysis system integrates the location, time and textual data values for the event detection process. The event classification is performed with sub class details. The Multi Task Learning (MTL) scheme is used for the concurrent analysis process. The system is divided into four major modules. They are Social Network Data Analysis, Newsfeeds Analysis, multi task learning process and event discovery process.

Social network data analysis module is used to manage user profile and location details. Newsfeeds analysis is used to extract features from the textual data values. The multi task learning model identifies the features that are used for the event forecasting process. The event discovery module is build to identifies the events in the social media data values.

5.1. Social Network Data Analysis

The social network account profiles are extracted for the spatio temporal behavior analysis process. The user account creation and check in points are fetched for the analysis. User profile, check in details and newsfeeds are collected from social networks. Foursquare and Twitter social network data values are used in the system. User check-in venue and time details are maintained in location data values. The latitude and longitude values are used to indicate the geographical location information for the user checkin operations. User submitted messages are maintained under newsfeeds data collection.

5.2. Newsfeeds Analysis

User messages are analyzed in newsfeeds analysis. The news feeds are maintained as textual data elements. Feature selection is performed on textual data values. Noisy data are eliminated from the messages. The Term Frequency (TF) and Inverse Document Frequency (IDF) are used in the term level analysis process. The term frequency and inverse document frequency values are used in the term weight estimation process. The term weights are used in the document weight estimation process. The entire tweet weight is utilized in the similarity relationship estimation process. Term level relationships are analyzed in the similarity estimation process.

5.3. Multi Task Learning Process

The multi task learning models are used to extract the features that are used for the event forecasting process. The indexing process is called to arrange the user and location relationships. The indexing process is carried out with two levels. They are user-venue relationship based model and Venue – user relationship models. The indexing process arranges all the data items that are submitted by the users. The Constrained Multi Task Feature Learning (CMTFL) model uses the location and news fed relationships for the feature extraction process. The Hybrid Multi Task Feature Learning (HMTFL) scheme combines the spatio temporal correlations with news feed features. The feature learning operations are simultaneously carried out with each location levels. The location and time proximity is used in the HMTFL scheme. All the feature learning results are represented in tree structured data models.

5.4. Event Discovery Process

The event discovery process is build to discover the social activities that are reflected in the social media data values. The event discovery operations are carried out for each location context. The event discovery process is carried out with two models. They are constrained event discovery model and hybrid event discovery model. The constrained event discovery model uses the features that are extracted using the Constrained Multi Task Feature Learning (CMTFL) scheme. The CMTFL scheme mainly focuses on the newsfeed relationships. The hybrid event discovery process uses the features that are extracted from the Hybrid Multi Task Feature Learning (HMTFL) scheme. The location and time relationships are combined with the news feeds features. The combination of events is

also discovered from the hybrid event discovery process.

6. Conclusion and Future Work

The social networks are the key medium to people to express their views and ideas. The social network data analysis methods are employed to discover the events. The feature learning methods are used for the data analysis and learning operations. The Multi Task Learning (MTL) schemes support simultaneous learning operations. The concurrent learning operations are carried out for each location environment. The Constrained Multi Task Feature Learning (CMTFL) scheme is used to fetch the features that are required for the training process. The Hybrid Multi Task Feature Learning (HMTFL) scheme utilizes the spatio temporal parameters and the social media data values. The location and time relationships are utilized in the hybrid feature learning and event discovery process. The hybrid feature learning and event discovery scheme can be enhanced with privacy preserved learning and event discovery models. The stream based data analysis mechanism can be integrated with the social media data analysis process.

References

1. X. Yang, A. Ghoting, Y. Ruan, and S. Parthasarathy. A framework for summarizing and analyzing Twitter feeds. In KDD, pages 370–378. ACM, 2012.
2. L. Zhao, F. Chen, J. Dai, T. Hua, C.-T. Lu, and N. Ramakrishnan. Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling. PloS one, 9(10):e110206, 2014.
3. L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan. Spatiotemporal event forecasting in social media. In SDM, pages 963– 971. SIAM, 2015.
4. C. C. Aggarwal and K. Subbian. Event detection in social streams. In SDM, volume 12, pages 624–635. SIAM, 2012.
5. Liang Zhao, "Dynamic Theme Tracking in Twitter", IEEE International Conference on Big Data (Big Data) 2015
6. J. Lin, R. Snow, and W. Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In KDD, pages 422–429. ACM, 2011.
7. T. Lappas, M. R. Vieira, D. Gunopulos, and V. J. Tsotras. On the spatiotemporal burstiness of terms. VLDB, 5(9):836–847, 2012.
8. J. Ratkiewicz, M. Conover, M. Meiss, B. Gonc¸alves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In ICWSM, pages 297–304, 2011.
9. R. Li, S. Wang, and K. C.-C. Chang. Towards social data platform: automatic topic-focused monitor for Twitter stream. VLDB, 6(14):1966–1977, 2013.
10. C. X. Lin, Q. Mei, J. Han, Y. Jiang, and M. Danilevsky. The joint inference of topic diffusion and evolution in social communities. In ICDM, pages 378–387. IEEE, 2011.
11. K. Massoudi, M. Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. In Advances in Information Retrieval, pages 362–367. Springer, 2011.
12. N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz, et al. 'beating the news' with embers: forecasting civil unrest using open source indicators. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1799–1808. ACM, 2014.
13. L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan. Multi-task learning for spatio-temporal even In KDD, pages 1503–1512. ACM, 2015.
14. L. Hong, B. Dom, S. Gurumurthy, and K. Tsioutsiouliklis. A time-dependent topic model for multiple text streams. In KDD, pages 832–840. ACM, 2011.
15. Ben Verhoeven, Walter Daelemans and Barbara Plank. Twisty: A multilingual twitter stylometry corpus for gender and personality profiling. In Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC, pages 1632–1637, 2016.