Mu Tong

1004452

Literature Review of Event Detection in Social Media

Introduction

Analyzing the data on online social networks draws attention in the last few years according to the greatly increasing development of social medias [5], such as Twitter, Facebook, and YouTube. People use these social medias not only to socialize, but also to do other things [5], which is of great value [1]. For example, News media could use social media like Twitter to spread news, because people are willing to spread something important [5]; some companies can analyze the opinions from their customers to get ideas about their products or services [2], which is called "sentiment analysis" technique; Markets can analyze their selling data and customer profiles to identify the market trend [3] or provide personal advertisements [4], which is called "personalized recommendation" [1]; there are still a lot of applications and the list goes on.

Event detection is one of these applications, and it is highly noticeable due to its difficulties and social impact [1]. Petrovic et al. [7] defined event detection is to determine the first story which describes "a particular event that happened at specific time and place"; However, social media is different from traditional newswire due to its characteristic of no limitation about time and space [5]; therefore, Dou, et al. [6] defined event detection in social media as "an occurrence causing change in the volume of text data that discusses the associated topic at a specific time. This

occurrence is characterized by topic and time, and often associated with entities such as people and location". Briefly speaking, "event detection is the problem of automatically identifying significant incidents by analyzing social media data" [1], such as a baseball game or a big traffic crash.

There are five reasons why event detection in social media is more challenging [1]: volume and velocity, real-time event detection, noise and veracity, feature engineering, and evaluation. Briefly speaking, there are plenty of short and noisy information in the content, rapidly changing topics, and great volume data [5], so different methods could be used to handle this kind of problem.

Methods

Two types of basic algorithms could be used in event detection in social media: clustering-based event detection, anomaly-based event detection [1]. In addition, we also consider multiscale event detection.

1. Clustering based Event Detection

The task of event detection in social media is often using clustering, which is usually solved by supervised learning or unsupervised learning [1]. As for the supervised learning, supervised machine learning algorithms such as Support Vector Machine (SVM) will make decisions for event clusters. On the other hand, it requires a scoring function for unsupervised learning [1], and clusters will be identified as "event" or "non-event" based on the scoring function [1]. The main difference between supervised and unsupervised learning is that a labelled training set is required for supervised learning to train the classifier [1].

Some researchers consider many aspects to cluster the tweets, including temporal, spatial, semantic, frequency of word and user information [8] – [17].

Walther et al. [12] described an easy supervised learning algorithm for geo-spatial event detection on Tweets by monitoring all posts in a given spatial region and identified the places with the most important activity on the map. The clusters extract textual and non-textual features to train the classifiers including Decision Tree and Naïve Bayes classifier. The result showed that using both textual and non-textual features would have better precision and recall than only using textual features.

Becker et al. [8] introduced a supervised classification that considered several categories of features including temporal, social, topical features to cluster the tweets by using term frequency–inverse document frequency (TF-IDF), and used support vector machines to train the model, in order to improve the generic analysis of "trending topic", such as prioritization, ranking, and filtering of extracted content in tweets.

TEDAS [9] analyzed the spatial and temporal feature of tweets to detect new events and identify the importance of events especially for crime and disasters, such as floods and law-breaking activities. The basic idea is to extract tweets using Twitter API and return the tweets which is correlated to the given topic, such as crime. In other words, the system is looking for tweets which is correlated to the given topic and presents events in the map. The shortcoming for this method is that it is not designed for real-time, and it is often outdated.

Ozdikis et al. [10] enhanced the technique of lexicon-semantic of tweet by applying document similarity and clustering algorithms. The basic idea is to cluster TF-IDF vectors, and the distance metric is cosine similarity. Then, calculate all the cosine similarity of all vectors. Similar vectors are assumed to be semantically correlated. Finally, the method identifies the event with largest similar clusters. The shortcoming is obvious, which is that noisy terms really affect the accuracy.

EvenTweet [11] detected the most important local events (public events or emergency situations) in real-time from a Twitter stream by spatial-temporal characteristics, using tweets during 2012 UEFA European Football Championship. The basic idea for this approach is to put the keywords in the same geo-tagged tweets into the same cluster and score these keywords based on the level of burstiness, which means that deviations from the mean. Each cluster will get a score, which equals to the sum of all the scores of keywords in this cluster, and then, top K cluster would be the candidate cluster.

Zhang et al. [15] also described a method called GEOBURST to effectively detect real-time local event form geotagged tweet. This method is different from the method called EvenTweet mentioned above, since they considered not only spatial impact, but also semantic correlations between keywords. The spatial contribution is measured by a kernel function, and the semantic part is identified by the algorithm called “random walk” on a keyword graph. Besides, EvenTweet couldn’t detect local events in real time [15], but GEOBURST could. The first step of GEOBURST is to find all geo-topic clusters, which are spatial close and semantically correlated, as candidate events,

and then ranks all the candidates based on historical activity timeline, and finally returns top K events. The method is improved by two versions, one is called GEOBURST+ [16], which improved the ranking method. The ranking method performs keyword embedding to get the semantics of content. Then, considering both activity timeline and embedding of keywords, which greatly improves the detection effectiveness. Another version called Triovecevent [17] is presented, which considers multimodal embeddings of spatial, temporal, and semantic characteristics in the same latent space. If two aspects are highly correlated, their representations would be close. The model also applies a novel Bayesian mixture clustering model that can continuously updates the clustering results. However, there is still something to improve. For example, in order to capture the short text semantics, the way of embedding in Triovecevent is random walk, which could suffer from text sparsity [17]. Another issue is that it is difficult to filter the uninteresting activities from geo-tagged tweets. Lots of geo-tagged tweets just show people's routine life like eating, travelling, so some activities that we are interested such as traffic crash would not be selected as top K event candidates. Building a perfect function to filter uninteresting activity is hard.

2. Anomaly based Event Detection

Briefly speaking, this method is to identify abnormal observations [1], including "unexpected word usage, irregular spatial activity or different sentiment levels" [1].

Valkanas et al. [18] [19] used sentiment analysis to detect event and trends in social media. The idea why Valkanas uses sentiment analysis for event detection is

that the sentiment level would fluctuate when people comment on events [1], so when the sentiment level is different from the average, we can think that an event happens, and it affects the sentiment level of people who are close to that event. The basic idea for this method is to constantly identify the Probability Density Function (pdf) of sentiment levels and detect any outliers [19] from this pdf.

Detecting peaks of hashtags in Twitter [20] is another example for anomaly-based event detection, because a sudden surge of hashtag is likely to be an important event, but there is a big shortcoming, which is that they only used hashtag as parameters, and the content of tweets is discarded. The basic idea for this method is to extract hashtags using Map-Reduce and create their time interval every 5 minutes; and then, use Discrete Wavelet Transformation to detect the bursts of hashtags in the time interval which indicate events. This method cannot handle real-time task but focus on batch data analysis.

Vavliakis et al. [20] described their approach for the MediaEval Benchmark 2012 task. The task containing 167 thousand images from Flickr is to detect specific 3 events from these images. Firstly, they did some pre-processing for image text, including removing common words and translating into English using Google Translate API. Then, they used Latent Dirichlet Allocation (LDA) to extract topics from the images and used peak detection for each topic. If a topic received much more images, then we can consider that topic can be considered as an event. The disadvantage for this approach is the extremely expensive computation cost for LDA, which is unavailable for large data volume.

DeLLe [14] is presented as a method to automatically detect latest local events from geotagged tweet. There are two modules in the method: seeker and ranker. Seeker would first find the irregular locations which have a large number of similar unexpected tweets, and then ranker would rank these selected locations to get top-level locations which are considered as candidates. It is evaluated on some cities like Seattle and New York, and the results shows competitive effectiveness than baseline approach. The method is different from other methods, because seeker not only accounts for historical patterns, but also predicts the expected number of tweets; and then, compares the actual number of tweets with the predicted value to identify the existence of unusualness [14]. Besides, this method also considers other places when an event happens, instead of only focusing on the fixed location.

3. multiscale event detection

Only using fixed spatial methods in event detection couldn't help to detect events at a larger scale. For example, a fixed spatial method in event detection may only handle the problem in a fixed location such as a country level instead of global level. A larger scale event detection can detect events at a district or a city. Therefore, event detection at different scales is better to adapt to the requirement of real-life event [22]. For example, Visheratin et al. [23] built an enhanced quadtree -convolutional quadtree called "ConvTree" to handle the problem of different scales in event detection. The study used 60 million geotagged Instagram posts in the New York city to find a large range of event detection from local (a wedding party) to city (baseball game) and even country level (Christmas) events.

Conclusion

There are two basic methods to handle the problem of event detection in social media: clustering-based and anomaly-based event detection. In addition, multiscale event detection could adapt to the requirement of real-life event detection. We can find that there are still some common problems from above papers. One issue is that it is difficult to find an algorithm which is good enough to filter the uninteresting activities from geo-tagged tweets. Another issue is the way of embedding. From Triovecevent [17], the way of embedding is random walk, which could suffer from text sparsity.

For future work, we will explore other potential ways for embedding so that semantically similar tweets will also end up close in the vector space, such as doc2vec [24] or sent2vec [25], which are the latest embedding ways.

References

[1] Panagiotou, N., Katakis, I., & Gunopulos, D. (2016). Detecting events in online social networks: Definitions, trends and challenges. In *Solving Large Scale Learning Tasks. Challenges and Algorithms* (pp. 42-84). Springer, Cham.

[2] Bifet, A., & Frank, E. (2010, October). Sentiment knowledge discovery in twitter streaming data. In International conference on discovery science (pp. 1-15). Springer, Berlin, Heidelberg.

[3] Mathioudakis, M., & Koudas, N. (2010, June). Twittermonitor: trend detection over the twitter stream. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (pp. 1155-1158). ACM.

[4] Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E., & Yu, Y. (2012, August). Collaborative personalized tweet recommendation. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 661-670). ACM.

[5] Nurwidyantoro, A., & Winarko, E. (2013, June). Event detection in social media: A survey. In International Conference on ICT for Smart Society (pp. 1-5). IEEE.

[6] Dou, W., Wang, X., Ribarsky, W., & Zhou, M. (2012, October). Event detection in social media data. In IEEE VisWeek Workshop on Interactive Visual Text Analytics-Task Driven Analytics of Social Media Content (pp. 971-980).

[7] Petrović, S., Osborne, M., & Lavrenko, V. (2010, June). Streaming first story detection with application to twitter. In Human language technologies: The 2010 annual conference of the north american chapter of the association for

computational linguistics (pp. 181-189). Association for Computational Linguistics.

[8] Becker, H., Naaman, M., & Gravano, L. (2011, July). Beyond trending topics: Real-world event identification on twitter. In Fifth international AAAI conference on weblogs and social media.

[9] Li, R., Lei, K. H., Khadiwala, R., & Chang, K. C. C. (2012, April). Tedas: A twitter-based event detection and analysis system. In 2012 IEEE 28th International Conference on Data Engineering (pp. 1273-1276). IEEE.

[10] Ozdikis, O., Senkul, P., & Oguztuzun, H. (2012, August). Semantic expansion of tweet contents for enhanced event detection in twitter. In 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 20-24). IEEE.

[11] Abdelhaq, H., Sengstock, C., & Gertz, M. (2013). Eventweet: Online localized event detection from twitter. Proceedings of the VLDB Endowment, 6(12), 1326-1329.

[12] Walther, M., & Kaisser, M. (2013, March). Geo-spatial event detection in the twitter stream. In European conference on information retrieval (pp. 356-367). Springer, Berlin, Heidelberg.

[13] Xie, W., Zhu, F., Jiang, J., Lim, E. P., & Wang, K. (2016). Topicsketch: Real-time bursty topic detection from twitter. IEEE Transactions on Knowledge and Data Engineering, 28(8), 2216-2229.

[14] Wei, H., Zhou, H., Sankaranarayanan, J., Sengupta, S., & Samet, H. (2018, November). Detecting latest local events from geotagged tweet streams. In Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 520-523). ACM.

[15] Zhang, C., Zhou, G., Yuan, Q., Zhuang, H., Zheng, Y., Kaplan, L., ... & Han, J. (2016, July). Geoburst: Real-time local event detection in geo-tagged tweet streams. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 513-522). ACM.

[16] Zhang, C., Lei, D., Yuan, Q., Zhuang, H., Kaplan, L., Wang, S., & Han, J. (2018). Geoburst+: effective and real-time local event detection in geo-tagged tweet streams. ACM Transactions on Intelligent Systems and Technology (TIST), 9(3), 34.

[17] Zhang, C., Liu, L., Lei, D., Yuan, Q., Zhuang, H., Hanratty, T., & Han, J. (2017, August). Triovecevent: Embedding-based online local event detection in geo-tagged tweet streams. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 595-604). ACM.

[18] Valkanas, G., & Gunopulos, D. (2013). Event Detection from Social Media Data. IEEE Data Eng. Bull., 36(3), 51-58.

[19] Valkanas, G., & Gunopulos, D. (2013, October). How the live web feels about events. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management (pp. 639-648). ACM.

[20] Cordeiro, M. (2012, January). Twitter event detection: combining wavelet analysis and topic inference summarization. In Doctoral symposium on informatics engineering (pp. 11-16).

[21] Vavliakis, K. N., Tzima, F. A., & Mitkas, P. A. (2012). Event Detection via LDA for the MediaEval2012 SED Task. In MediaEval.

[22] Dong, X., Mavroeidis, D., Calabrese, F., & Frossard, P. (2015). Multiscale event detection in social media. Data Mining and Knowledge Discovery, 29(5), 1374-1405.

[23] Visheratin, A. A., Mukhina, K. D., Visheratina, A. K., Nasonov, D., & Boukhanovsky, A. V. (2018, November). Multiscale event detection using convolutional quadtrees and adaptive geogrids. In Proceedings of the 2nd ACM SIGSPATIAL Workshop on Analytics for Local Events and News (p. 1). ACM.

[24] Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196).

[25] Pagliardini, M., Gupta, P., & Jaggi, M. (2017). Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507*.