Global Social Event Extraction and Analysis by Processing Online News

Bing Zhu, Yu Wang, Chenglong He

Science and Technology on Information Systems Engineering Lab, Nanjing, China
Nanjing Research Institute of Electronics Engineering, Nanjing, China
Email: bingzhazha@126.com

Abstract. The understanding of global social pattern can benefit the society operation a lot, including the domestic social governance, international situation awareness, risk assessment and forecasting, conflict resolution, crisis response and future policy planning. As the development of Internet, hundreds of millions of news could be found online every day, reporting the social events around the world, including political events, diplomatic events, cultural events, natural disasters, etc. However, by manual reading and analyzing, it is too difficult to deal with the vast amount of data and obtain valuable information quickly. Thus, in this paper we investigate the global social event extraction and analysis method based on the automatic processing of online news, including 1) event model building, 2) event information extraction and automatic classification based on English news text, 3) global social dynamic analysis and visualization based on event data. Finally, we constructed the method on the real global news data collected from more than 200 sites to evaluate their performance and interpret some underlying insights of the results.

Keywords: information extraction, event analysis, text processing

1 Introduction

Nowadays, as the rapid development of Internet technology and application, hundreds of millions of news could be found online every day, reporting the social events around the world, including political events, diplomatic events, cultural events, natural disasters, etc. Obviously, it is too difficult to deal with the vast amount of data and obtain valuable information quickly by manual reading and analyzing. Statistically, humans are able to identify and code about six to ten events per hour from news text, while machine automatic coding can bring an increase by a factor of about a million over human processing [1]. Therefore, benefiting from the increasing in both machine-readable text and computing power, the automatic event extraction and analysis has become a hot research topic, which will serve the society operation, situation awareness, risk assessment and future policy planning.

In this paper we investigate the global social event extraction and analysis method based on the automatic processing of online news. Among the enormous social events, we particularly pay attention to the international cooperation and conflict events of high concern, including economic cooperation, diplomatic cooperation, natural and man-made disasters, terrorist attacks, fights, etc. The method investigated including 1) event model building, 2) event information extraction and automatic classification based on English news text, 3) global social dynamic analysis and visualization based on event data. Finally, we constructed the method on the real global news data collected from more than 200 sites to evaluate their performance and interpreted some underlying insights of the results. We extract the social event information from the vast amount of online news data utilizing natural language processing method and then analyze the emerging social pattern. The contribution of this paper lies in two aspects:

- *Event modeling and information extraction:* We build a social event model, which can formulate an event as a brief and sufficient structural representation and benefit the fast query and correlation analysis. Based on the model, we propose a text processing method to extract structured event data from unstructured news text automatically. The structured event data comprise all the elements defined in the event model, including time, location, event actor(s), action and related persons, organizations, themes, etc.
- *Real evaluation:* We evaluate our method on a large-scale online news dataset collected from more than 200 sites. Analysing the extracted structured event data, we can find that we successfully identify the breaking social events and principal global social patterns, justifying the effectiveness of our event processing methods.

The remainder of the paper is organized as follows. In Sect. 2, we review some related work. We introduce the event model in Sect. 3 and the event element extraction method in Sect. 4. In Sect. 5, we evaluate our method on a real large-scale dataset and discuss the results. In Sect. 6, we draw a conclusion to this paper.

2 Related Work

In this section, we briefly review some related work. Political event data have long been used in the quantitative study of international politics, dating back to 1970s. Edward Azar's COPDAB(Conflict and Peace Data Bank) [2] and Charles McClelland's WEIS(World Event Interaction Survey) [3] are early efforts. In 1990s, two practical automated event data coding systems developed, that is the NSF-funded KEDS(Kansas Event Data System) [4] the proprietary VRA-Reader [5]. The KEDS was developed by Schrodt et al., utilizing machine-assisted approaches to generate the political data from news sources.

A later generation of event datasets combines new data with a sharper substantive focus, including location as an important major role. Event datasets

such as the ACLED(Armed Conflict Location and Event Dataset) [6], the geo-referenced event dataset released by the Uppsala Conflict Data Program [7] and the SCAD(Social Conflict in Africa Database) [8] all list events with precise spatial coordinates, making it possible to study the spatial patterns of events and the international and subnational relationships. However, all these datasets rely largely on human coding of news reports and thus require significant effort and time. [9]

Recently, there is a growing interest in utilizing news reports to extract and analyze event data. We refer to [10], [9], [11], [12] for some recent advances in this domain. Schrodt et al. [10] set up a quite influential project GDELT(Global Data on Events, Location and Tone), providing up-to-date real-world events data with its richness in covering all countries globally. Preprocessing of raw texts from media sources, has been already done by representing all the event types and actors using a standardized code set. The abstraction actually removes much domain-specific information and thereby makes a systematic analytical methodology become available. However, many details of the events are also lost and make it difficult to mine deeper semantic information and insight, which inspires us to do the event coding and analysis work. [9], [11] and [12] utilize the GDELT dataset to do some political conflict analysis.

3 Event Model

We build an event model to formulate the structured representation of social events extracted from unstructured news reports, which has laid an important foundation to the management and analysis of the various event data. Briefly, the model captures two actors and the action performed by Actor1 upon Actor2 when and where. Table. 1 shows the major elements of the event model and a piece of data sample.

Table 1. A piece of data sample

ID	PostDate	EventDate	StoryNum	Actor1Name	Actor1Country	Actor1Lat
2015111300261	20151114	20151113	130	*Paris*	*FRA*	48.855562

Actor1Long	Actor2Name	Actor2Country	Actor2Lat	Actor2Long	Action
2.352361	*ISIS*	*SYR*	35.051041	38.450062	*terrorist attack*

ActionCountry	ActionLat	ActionLong	Category	Title	URL
FRA	48.855562	2.352361	*conflict*	*newstitle*	*http://news.report.url*

- *ID:* the global unique identifier of each event.
- *PostTime:* the Date when the news reported the event in YYYYMMDD format.
- *EventTime:* the Date when the event took place in YYYYMMDD format.
- *StoryNum:* the number of duplicated stories reporting the same event from different news sources.

- *Actor1Name:* the actual name of the Actor1, e.g. the full name of a country/ state/ city, a person, an organization, etc.
- *Actor1Country:* the 3-character ISO 3166 code for the country affiliation of *Actor*1.
- *Actor1Lat:* the centroid latitude of the landmark for mapping $Actor1Country$.
- *Actor1Long:* the centroid longitude of the landmark for mapping $Actor1Country$.
- *Actor2Name:* the actual name of the Actor2, e.g. the full name of a country/ state/ city, a person, an organization, etc.
- *Actor2Country:* the 3-character ISO 3166 code for the country affiliation of *Actor*2.
- *Actor2Lat:* the centroid latitude of the landmark for mapping $Actor2Country$.
- *Actor2Long:* the centroid longitude of the landmark for mapping $Actor2Country$.
- *Action:* the actual action(what *Actor*1 did to *Actor*2).
- *ActionCountry:* the 3-character ISO 3166 code for the country affiliation of *Action* took place.
- *ActionLat:* the centroid latitude of the landmark for mapping $ActionCountry$.
- *ActionLong:* the centroid longitude of the landmark for mapping $ActionCountry$.
- *Category:* the category the action classified to. Four categories are investigated in this paper, i.e. *Support, Cooperation, Object, Conflict.*
- *Title:* the title of the news from which the event was extracted.
- *URL:* the URL where the news report collected.

4 Event Extraction

In this section, we look into the method to extract social events from news reports based on the model built in Sect. 3 and classify the event into four interested categories, i.e. international or civil support, cooperation, object and Conflict event. As Fig. 1 shows, first We use a story filter to remove the theme irrelative news texts, such as the sports new, business news, etc. Then, we use the natural language tools to produce a parse tree and identify the named entities upon the theme-related texts. Based on the action and actor dictionaries(generated from event and actor ontologies), event elements are extracted automatically. Considering both the actor and action attributes, put the event that can be classified into one of the four interested categories into the dataset. Finally, event pattern analysis and visualization can be carried out.

4.1 Story filter

News sources always contain a large number of stories which do not contain interested social events(international or civil support, cooperation, object and Conflict events). Thus, we use standard supervised text classification algorithm - the support vector machine - to deal with the theme filter issue. We utilize several sets of positive and negative cases to train the classifiers, rather than some detailed specification. It is pretty difficult to distinguish some sports and entertainment story from the interested political ones, because of the news metaphors.

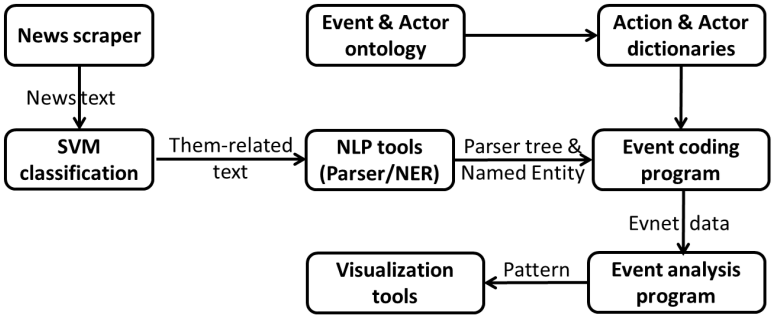

Fig. 1. the process flowchart of event extracting and analysis

4.2 Natural language process

One of the most popular NLP tools Stanford CoreNLP [13] is utilized to product the text parsing and named entity identification. The Penn Treebank-formatted parse tree(Fig. 2) is generated which will be used to extract the actors, action and locations based on the event model.

4.3 Event deduplication

Historically, deduplication has been an important part of the event data generation process when automated coding is used. It was originally used because of multiple reports of the same event - for example a meeting or a terror attack can easily generate tens of reports dealing with the same occurrence—and become even more important when multiple local sources are used, as is the case with news aggregators, since these will frequently reprint wire service stories, sometimes with editing, sometimes not. In the current environment of web-based sourcing, deduplication turns out to be a critical aspects, and if done sloppily, can actually result in very significant proliferation of events, very significantly increasing the level of noise in the dataset. However, the number of duplicates may be useful as an indicator of how important a story is, at least to the media. Consequently the our system keeps the number of duplicates in the field $StoryNum$ indicating the event's influence.

4.4 Geo-locating

Accurate location of a single event is much more difficult than identifying location words within a given set of text, because a text may mention several different

```
Dagolath's first Deputy Prime Minister Telemar left for
Minas Tirith on Wednesday for meetings of the joint transport
committee with Arnor, the Dagolathi news agency reported.
(ROOT
  (S
    (S
      (NP
        (NP (NNP Dagolath) (POS 's))
        (ADJP (JJ first))
        (NNP Deputy) (NNP Prime) (NNP Minister) (NNP Telemar))
      (VP (VBD left)
        (PP (IN for)
          (NP
            (NP (NNP Minas) (NNP Tirith))
            (PP (IN on)
              (NP (NNP Wednesday)))))
        (PP (IN for)
          (NP
            (NP (NNS meetings))
            (PP (IN of)
              (NP
                (NP (DT the) (JJ joint) (NN transport) (NN committee))
                (PP (IN with)
                  (NP (NNP Arnor)))))))))
    (, ,)
    (NP (DT the) (NNP Dagolathi) (NN news) (NN agency))
    (VP (VBD reported))
    (. .)))
```

Fig. 2. a Penn Treebank example

locations, but it is necessary to identify one as the single location for the event. As with the broader issue of geo-coding, this is not a solved problem and will likely require a fair amount of active research to solve. In order to geolocate the coded events, we makes use of the CLAVIN project(http://clavin.bericotechnologies.com/).

5 Evaluation

In this section, we carry out our method on a real large-scale online news dataset and evaluate the performance.

5.1 Event extraction performance

We conducted our system on the worldwide news collected from more than 200 sites, from November 1st to 30th, 2016, including BBC news, Xinhua news, Agence France-Presse, Associated Press, etc. News updated once per hour. In the data pre-processing, stories unrelated with the four interested social theme have been omitted. About 4000 news stories were processed and 7000 events were extracted per day. By artificial subjective validation, our method worked well with a 70 percent recall rate and a 80 percent precise rate.

5.2 Event Analysis

Based on the extracted structured event data, we did some analysis and successfully identified some breaking social events. Some visualization results can be found in Fig. 3-Fig. 6, indicating the principal global social patterns, which can justify the effectiveness of our event processing methods as well.

Fig. 3 shows the global social event heatmap in November, 2016, illustrating France and Syria as the hottest districts all over the world. It is perfectly corresponding to the fact that Paris suffered a severe terrorist attack organized by the Syria ISIS. Then, we select all the event data concerning "terrorist attack" in the source news title in November around the world, and depict the interaction network between countries in Fig. 4. Obviously, France is the central node in the network, which is mainly connected with Syria, Belgium, United States, etc. Fig. 5 shows the timeline of events in four categories took place in France in November. Nov. 13th, the date the Paris terrorist attack happened, turns out to be the turning point, after which the conflict and object events increased significantly. Each point in Fig. 5 indicating an event, whose radius is proportional to the influence of the event(in fact, the number of duplicated reports). The word cloud Figure(Fig. 6)gives the key words about terrorist attack in November, whose font size is proportional to its importance. Thus, we can find that "Islamic state", "French President Francois Holl", "Middle East", "United States", etc. are the most significant key words describing the terrorist attack issue in November.

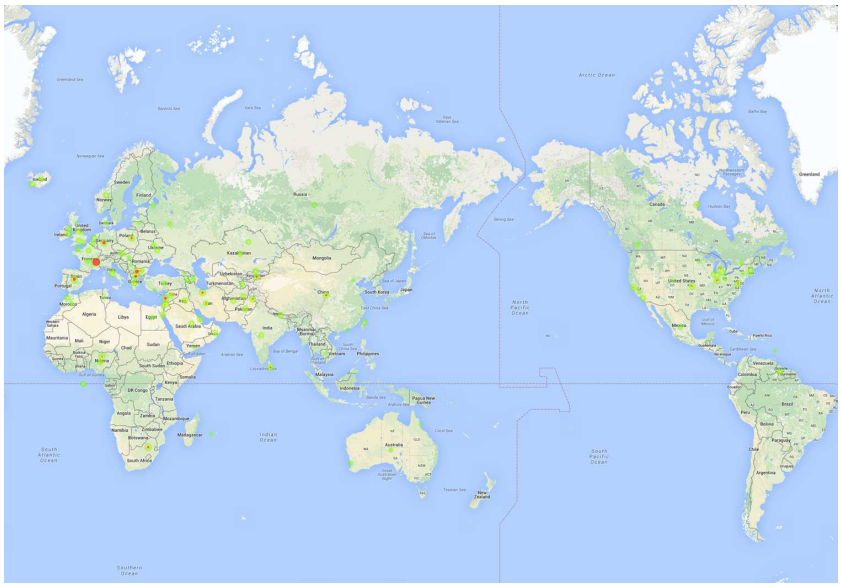

Fig. 3. Global social event heatmap in November, 2016

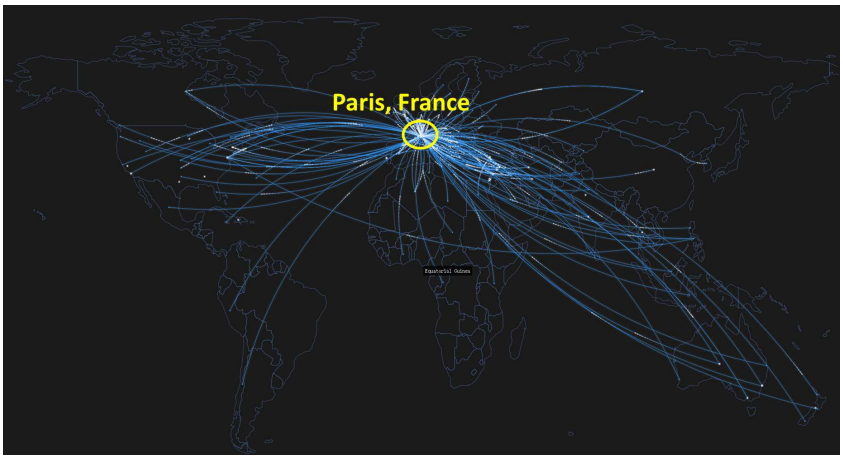

Fig. 4. International terrorist attack interaction network in November, 2016

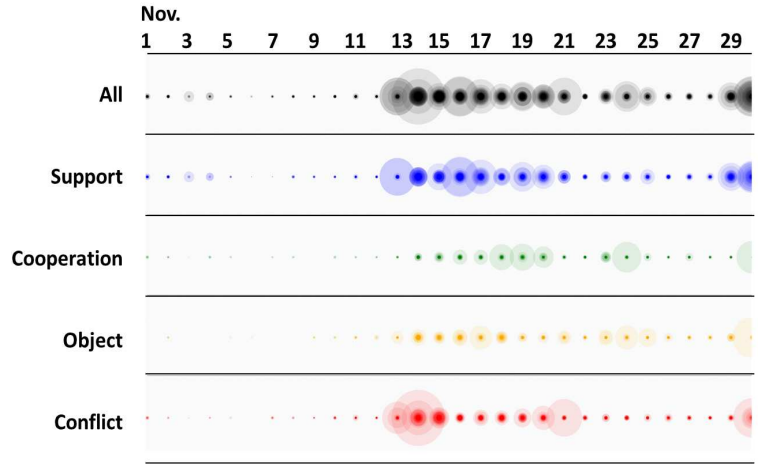

Fig. 5. Timeline of events in four categories in France, November, 2016

Fig. 6. Key words about terrorist attack in November, 2016

6 Conclusion

In this paper we investigate the global social event extraction and analysis method based on the automatic processing of online news. Among the enormous social events, we particularly pay attention to the international support, cooperation, object and conflict events of high concern, including economic cooperation, diplomatic cooperation, natural and man-made disasters, terrorist attacks, fights, etc. We extract the social event information from the vast amount of online news data utilizing natural language processing method and then analyze the emerging social pattern. Specifically, the method investigated including 1) event model building, 2) event information extraction and automatic classification based on English news text, 3) global social dynamic analysis and visualization based on event data. Finally, we constructed the method on the real global news data collected from more than 200 sites to evaluate their performance and interpreted some underlying insights of the results, which turned out to validate the effectiveness of the methods proposed.

Acknowledgement. This research work was supported by National Natural Science Foundation of China under Grant No. 61402426 and partially supported by Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] Schrodt, P.A., Beieler, J., Idris, M.: Threesa charm?: Open event data coding with el: Diablo, petrarch, and the open event data alliance. In: ISA Annual Convention. (2014)
[2] Azar, E.E.: The conflict and peace data bank (copdab) project. Journal of Conflict Resolution **24**(1) (1980) 143–152
[3] McClelland, C.A.: World event/interaction survey codebook (icpsr 5211). inter-university consortium for political and social research. Ann Arbor (1976)
[4] Gerner, D.J., Schrodt, P.A., Francisco, R.A., Weddle, J.L.: Machine coding of event data using regional and international sources. International Studies Quarterly (1994) 91–119
[5] King, G., Lowe, W.: An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. International Organization **57**(03) (2003) 617–642
[6] Raleigh, C., Linke, A., Hegre, H., Karlsen, J.: Introducing acled: An armed conflict location and event dataset special data feature. Journal of peace research **47**(5) (2010) 651–660
[7] Sundberg, R., Melander, E.: Introducing the ucdp georeferenced event dataset. Journal of Peace Research **50**(4) (2013) 523–532
[8] Salehyan, I., Hendrix, C.S., Hamner, J., Case, C., Linebarger, C., Stull, E., Williams, J.: Social conflict in africa: A new database. International Interactions **38**(4) (2012) 503–511
[9] Hammond, J., Weidmann, N.B.: Using machine-coded event data for the micro-level study of political violence. Research & Politics **1**(2) (2014) 2053168014539924

[10] Leetaru, K., Schrodt, P.A.: Gdelt: Global data on events, location, and tone, 1979–2012. In: ISA Annual Convention. Volume 2., Citeseer (2013)
[11] Jiang, L., Mai, F.: Discovering bilateral and multilateral causal events in gdelt
[12] Keertipati, S., Savarimuthu, B.T.R., Purvis, M., Purvis, M.: Multi-level analysis of peace and conflict data in gdelt. In: Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, ACM (2014) 33
[13] Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. (2014) 55–60