Building Terrorist Knowledge Graph from Global Terrorism Database and Wikipedia

Tian Xia
School of Information Resource Management
Renmin University of China
Beijing, China
xiat@ruc.edu.cn

Yijun Gu
School of Information Technology and Cyber Security
People's Public Security University of China
Beijing, China
guyijun@ppsuc.edu.cn

***Abstract*—The Global Terrorism Database (GTD) is the most important dataset in counter-terrorism domain. Existed studies based on GTD focused on terrorism influences, data statistics and visualization, and terrorism event mining such as classification and clustering. In this paper, we build a terrorism knowledge graph(TKG) from GTD and Wikipedia. Compared with GTD, TKG enhaced the organizations of terrorism entities and relationships, and enriched the description by attatching Wikipedia knowledges. Therefore, TKG can better the understanding of terrorism attacks for both human beings and machine processing like graph mining and knowledge reasoning.**

***Index Terms*—Terrorism knowledge graph, terrorism dataset, GTD, Wikipedia**

I. Introduction

Terrorism is a serious expansive threat to global security. In order to increase the understanding of terrorist violence, the Study of Terrorism and Responses to Terrorism (START) of the United States has released the open-source Global Terrorism Database abbreviated as GTD online, so that the terrorism can be more readily studied and defeated [1]. Today, GTD includes systematic data on terrorist incidents around the world from year 1970 through 2017, and will be annual updated in the future. Due to its openness and comprehensive data characteristics, GTD plays an important role in counter-terrorism research.

The original data of GTD is saved into an Excel file in the form of a two-dimensional table. Though GTD has already well organized, there are still some shortcomings in use: First, GTD puts all entities and attributes of any terrorism incident into a data row, missing explicit relationships among entities; Second, GTD only contains data directly related to terrorism, and lacks background information like religious and demographic characteristics of attack place.

In this paper, we try to solve above problems by building terrorism knowledge graph(TKG). The GTD column fields are divided into entities and attributes, and different relationships are introduced to connect the entity nodes in knowledge graph. At the same time, the general knowledge extracted from Wikipedia are added into TKG to provide more background information, therefore, researchers can more easily obtain the possible information they need.

The rest of this paper is organized as follows: Section 2 introduces related work using GTD. Section 3 describes the proposed methodology and the flowchart to construct the terrorism knowledge graph. Section 4 describes the implementation detail, and conclusion and future work are described in section 5.

II. Related Work

Current research on GTD dataset can be divided into three main categories.

- Influences caused by terrorist incidents.
  Terrorist attacks can significantly influences investor sentiment and behavior at stock market. Konstantinos find that the negative effect of terrorist activity was substantially amplified as the level of psychosocial effects increased [4]. Charles et al. explored the reaction of hospitality stocks to the terrorist activities and find that investor sentiment affected by the terrorist events played a substantial role in hospitality stock returns [5]. Terrorism also has a indirect impact on the job market as shown in paper [6].
- GTD data statistics and visualization.
  Important patterns and trends of terrorist activities can be found through statistical analysis and visualization of GTD dataset. For example, Webb et al. described the spatio-temporal trends in terrorist incidents in the United States, and examined their characteristics like location, target type, weapon and attack type [7]. Ivana et al. compared the trends identified in GTD with the trends reflected through the social media, and suggested some media bias and public perception on terrorism [8]. For visual analysis of GTD, social network analysis techniques are usually used to output human friendly graph results as shown in [9], [10].
- Machine learning based terrorist event mining.
  Recently, some researchers applied machine learning method to predicate the terrorist event based on GTD dataset, so that the experts can find more hidden patterns and make better decisions against terrorist. Guohui et al. investigated the terrorist attacks and the relationships between the fatality and the influencing factors to better the understanding of terrorism [11]. Mo et al. applied Support Vector Machine, Naive Bayes and Logistic Regression algorithms to predicate the attack types [12], and Kim et

978-1-7281-2504-6/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: University of Maryland College Park. Downloaded on July 06,2023 at 02:10:50 UTC from IEEE Xplore. Restrictions apply.

al. used deep learning based named entity recognition to label terrorism incident automatically [13].

To the best of our knowledge, though there are a lot of counter-terrorism studies based on GTD dataset, the proposed research is the first to strengthen the construction of terrorism data itself to provide more powerful data ability.

III. Methods

Knowledge graph is a large network of entities, attributes and their semantic relationships. It's a powerful tool that changes the way we do data organization, retrieval and analysis. Therefore, we build a Terrorism Knowledge Graph (TKG) from GTD and Wikipedia, aim at semantically representing the truth of terrorism information in the form of machine readable graph structure. The whole processing flowchart is shown in figure 1.

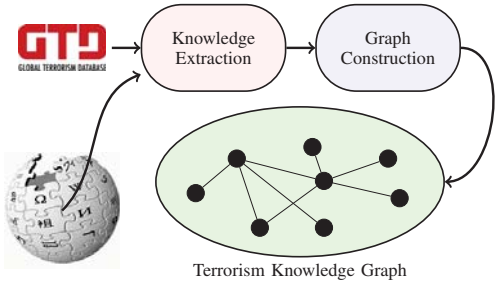

Fig. 1. Processing flowchart for terrorism knowledge graph construction

A. TKG Definition

TGK can be defined as a triple: $G =< G_s, G_d, R)$ where $G_s =< V_s, E_s >$ represents schema graph, which determines the types/concepts of nodes and edges that can appear in graph G, V_s is the set of concepts(nodes) and E_s is the set of all possible relationships(edges).

$G_d =< V_d, E_d >$ represents the data graph, where V_d is the set of entities and literal strings, and E_d is the set of relationships between two entities or between entity and attribute value.

R represents the relationship between G_s and G_d. After instantiated the schema graph G_s with fact data, there comes the data graph G_d, which can be used in case study and terrorism mining.

G_d can also be represented by RDF standard under the restrict of G_s definition, i.e., using SPO(*subject, predicate, object*) triples to store the facts, where *subject* and *object* are entities and *predicate* is the relation between them.

GTD records up to 120 separate attributes of each incident under eight broad categories, such as incident date, region, tactic used in attack. To construct the schema graph G_s, all these attributes are summarized and divided into six categories:

1) Incident: includes ID, date, approximate date, summary text, latitude and longitude etc.
2) Weapon: describes the weapon used in attack, includes weapon name, type, sub type and description.
3) Perpetrator: describes perpetrator group or individual information.
4) Location: describes the attack place which includes region, country, state, city and their detailed information like population, religion and culture.
5) Target: describes the nature of the target such as government, military, police, business.
6) Damage: describes damage information like total number of fatalities and injured.

The edge relationships we defined include: *use_weapon, target_at, located, has_damage, attack_by*, most attributes are connected to corresponded entity via literal string value in practice.

B. Knowledge Extraction

Though GTD provides most important features related with terrorism events, experts still need more information else like weapon characteristics, detailed information about attacked city and perpetrator. Therefore, TKG uses three strategies in knowledge extraction to merge the GTD and Wikipedia together.

- Knowledge extraction from GTD structured data.
  Most GTD data can be extracted and converted into knowledge graph via heuristic rules. For example, we make "country" node in graph by GTD field *country* directly. However, several special fields contain hierarchical information like "Dynamite/TNT" and "Molotov Cocktail/Petrol Bomb", we split them into multiple nodes and maintain these hypernym relations.
- Knowledge extraction from GTD free text.
  GTD fields like *motive, added notes and cite sources* are made up of free text. In order to extract named entities and topic phrases, we use Stanford Core NLP to do lexical analysis and extract all noun phrases. Then, we rank all phrases by their frequency in all records and add meaningful items to TKG after filtering.
- Knowledge extraction from Wikipedia infobox.
  Although the content of a Wikipedia page is essentially unstructured, many pages still have structured information called infobox, which can be transformed into meaningful data. We use JWPL to extract the property and value pairs from infobox fragment, then, clean the data pairs by removing format tags like *"bold"* and *"br"*. The extracted pairs will be attached to the linked entity node as the background knowledge.

C. Entity Linking and Graph Construction

Due to the ambiguity of natural language, entity resolution and linking must be taken into account when attach Wikipedia knowledge to the entities extracted from GTD. We apply document embedding technique to cope with this problem.

Suppose $r = w_{r1}, w_{r2}, \cdots, w_{rm}$ represents a record of GTD, where w_{ri} represents the i^{th} word appeared in the free text of record r. $d = w_{d1}, w_{d2}, \cdots, w_{dn}$ represents a Wikipedia article, and w_{dj} represents the j^{th} word in article d. Let $\vec{w} = (v_{w1}, v_{w_2}, \cdots, v_{wv})$ represents the embedding of

Authorized licensed use limited to: University of Maryland College Park. Downloaded on July 06,2023 at 02:10:50 UTC from IEEE Xplore. Restrictions apply.

word w with length $|v|$ in pre-trained GloVe model [14], then the crentroid of record r and article d can be calculated as:

$$\vec{r} = \frac{1}{m}\Sigma_{i=1}^{m}\vec{w_{ri}}, \quad \vec{d} = \frac{1}{n}\Sigma_{j=1}^{n}\vec{w_{dj}}. \tag{1}$$

Given an entity $e \in r$, we first collect all Wikipedia articles $d_1, d_2, \cdots, d_n$ by article title and alias, and then choose the target article d by:

$$d = \arg\max_{d} sim(\vec{d}, \vec{r}|e \in r), \tag{2}$$

where $sim(\vec{d}, \vec{r})$ represents the similarity of document d and GTD record r, and can be calculated by cosine similarity:

$$sim(\vec{d}, \vec{r}) = \frac{\overrightarrow{d} \cdot \overrightarrow{r}}{|\overrightarrow{d}||\overrightarrow{r}|} = \frac{\sum_{i=1}^{|v|} v_{di} \cdot v_{ri}}{\sqrt{\sum_{i=1}^{|v|} v_{di}^2}\sqrt{\sum_{i=1}^{|v|} v_{ri}^2}}. \tag{3}$$

Using equation 2, we link weapon, perpetrator, place, extracted phrase of GTD to the most similar Wikipedia article, and attach the infobox structure data at the same time.

IV. Implementation

A. Datasets for building TKG

There are two datasets we used to build the TKG as follows:

1) GTD: We are working with the last GTD dataset which has 181691 records. The important attributes included in it are: country, city, latitude and longitude location, weapon, date(year, month, day) etc. The knowledge graph schema is manually designed by analyzing all GTD attributes.
2) Wikipedia: Wikipedia provides database dumps, which are the complete copies of all Wikipedia articles and linkages, in the form of wikitext source and metadata embedded in XML. The dump file to build TKG is enwiki-20180801-pages-articles-multistream.xml.bz2 (from 2018 August English dump), which can be downloaded from Wikipedia dump site. Besides GTD, extra descriptions of TKG are extracted from the semi-structured infobox and free text of Wikipedia articles.

B. Tools and process steps

The following tools are used in our implementation:

- Apache POI: extract GTD data from Excel file.
- JWPL(Java Wikipedia Library): convert the Wikipedia dump file to MySQL and parse the Wikipedia syntax.
- JanusGraph: a scalable graph database optimized for storing and querying graphs, we store all TKG nodes and edges in it.

We take two steps to construct TKG:

- Step 1: First scan GTD data and construct initial graph nodes and edges.
  We first read each record from the spreed sheet of GTD data, and insert or located the corresponded nodes into the graph, then we attach attributes to entity node, and connect the entities with specified relationships by manual coded rules.
- Step 2: Second scan GTD data and attach Wikipedia knowledge.
  Wikipedia articles are extracted from the dump file, and store the structured information of article into MySQL database, include article title, abstract, full content and key-value pairs from infobox. Then, we scan GTD again, and enhance all possible entities by adding the Wikipedia knowledge like perpetrator introduction, city information and weapon parameters and characteristics.

V. Conclusion

In this paper, we proposed the method for construct the terrorism knowledge graph TKG from GTD and Wikipedia, TKG will contribute for both machine processing and human analysis about terrorism. The future work includes:

- provide visual interface to help experts using TKG.
- apply graph embedding technique on TKG to improve the effect of terrorist attack predication.

Acknowledgment

This work was supported by the National Key R&D Program of China (NO. 2017YFC0820100).

References

[1] National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2018). Global Terrorism Database [Data file]. Retrieved from https://www.start.umd.edu/gtd.

[2] G. Lafree, L. Dugan, "Introducing the Global Terrorism Database," Terrorism and Political Violence, 2007, vol. 19, pp. 181-204.

[3] V. K. Borooah, "Terrorist Incidents in India, 19982004: A Quantitative Analysis of Fatality Rates," Terrorism and Political Violence, 2009, vol. 21, pp. 476–498.

[4] D. Konstantinos, "Terrorism Activity, Investor Sentiment, and Stock Returns," Review of Financial Economics, 2010, vol. 19, pp. 128–135.

[5] C. Charles, Y. Y. Zeng, "Impact of Terrorism on Hospitality Stocks and the Role of Investor Sentiment," Cornell Hospitality Quarterly, 2011, vol. 52, pp. 165–175.

[6] G. Robert, L. Dugan and G. LaFree, "The Impact of Terrorism on Italian Employment and Business Activity," Urban Studies, 2007, vol. 44, pp. 1093–1108.

[7] J. Webb, S. L. Cutter, "The Geography of U.S. Terrorist Incidents, 19702004," Terrorism and Political Violence, 2009, vol. 21, pp. 428–449.

[8] T. Ivana, S. Setu, and L. Xiao, "Are recent terrorism trends reflected in social media," 2017 14^{th} International Conference on Mobile Ad Hoc and Sensor Systems, Orlando, 2017, pp. 535–539.

[9] X. Wang, E. Miller, K. Smarick, W. Ribarsky, and R. Chang, "Investigative visual analysis of global terrorism," IEEE VGTC conference on Visualization, 2008, vol. 27, pp. 919–926.

[10] L. V. Hegde, N. Sreelakshmi, and K. Mahesh, "Visual Analytics of Terrorism Data," 2016 IEEE International Conference on Cloud Computing in Emerging Markets, 2016, pp. 90–94.

[11] L. Guohui, L. Song, C. Xudong, Y. Hui and Z. Heping, "Study on correlation factors that influence terrorist attack fatalities using Global Terrorism Database," Procedia Engineering, 2014, vol. 84, pp. 698–707.

[12] H. Mo, X. Meng, J. Li, and S. Zhao, "Terrorist Event Prediction Based on Revealing Data," International Conference on Big Data, 2017, pp. 239 – 244.

[13] I. Kim, W. M. Pottenger, and V. Behe, "Can a Student Outperform a Teacher? Deep Learning-based Named Entity Recognition using Automatic Labeling of the Global Terrorism," IEEE International Symposium on Technologies for Homeland Security, 2018, pp. 1–6.

[14] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," Proceedings of the EMNLP, 2014, pp. 1532–1543.

Authorized licensed use limited to: University of Maryland College Park. Downloaded on July 06,2023 at 02:10:50 UTC from IEEE Xplore. Restrictions apply.