Detection of Similar Terrorist Events

Vittoria Cozza[1] and Michelangelo Rubino[2]

[1] Department of Computer Science, University of Pisa, 56124 Pisa, IT
cozza@di.unipi.it
[2] Expert System, 41123 Modena, IT
mrubino@expertsystem.it

Abstract. Event counting is significant when it allows us to discover and represent implicit knowledge. We realize that a particular event happens somewhere not just by mere chance, it is unlikely to be what we call as *accidental event*. E.g. the number of violent attacks and terrorist acts can give the measure of the safety for a given country and can help us to predict where and/or when similar events are likely to happen next time.
This work proposes an approach for detecting terrorist events sharing common details, available from open datasets, with the aim of merging their descriptions and counting them exactly. Events are aggregated according to a space-time-textual similarity function.

1 Introduction

Generally speaking, a high number of particular events within a geographical area can give you a clue about what that place is characterized by. We can use this logic likewise with events connected with terrorism, car bombs, suicide attacks and any other event identified as terrorist act allow us to realize whether a country is dangerous or not: the higher the number of violent acts, the greater the risk related to that place. This is the reason why we need a methodology to count this kind of events; the number of terrorist acts is therefore the indicator of safety for a place.

In the following, we propose a two-stage approach, that is first extracting useful information from events and representing it as space-time-textual records, then clustering these records according to a similarity function by combining space and time proximity and keyword relevance.

The first stage starts with the news contained in public datasets, where events consist of a short description including place, date, act and casualty, with no comments or personal opinions. Sometimes we can also find the group (typically an acronym) claiming the act. There is a point about the length of the description to be considered when we use these public databases: more details mean generally a longer description and a longer description means a more important event. In its turn, an important event is more likely to be found in each dataset we use and often to be found more times in the same dataset. This because of updates, especially when dealing with terrorist acts considered as relevant: daily updates can increase the news size. Here indeed, we need the second stage, whose purpose is grouping the occurrences of the same event. In short, one event, one record. Once we get single records, the similarity function can be exploited with

different settings either by using the three dimensions or by combining them two by two. This way, we have clusters of events with the same keywords to be placed on the basis of space and/or time: e.g. the increasing presence of female terrorists in a particular area starting from a particular event or date, the use of the AK-47 rifle in a region and so forth.

2 Related works

This work refers to existing approaches for information integration, e.g. entity resolution or deduplication, that also aims at finding real-world entities occurring in different forms in multiple data records. For a review [5]. In [6] authors deal with deduplication too. In particular, facing the problem of identifying redundant social network messages, they are able to identify whether one message subsumes information from another one (textual entailment) or they both convey the same information (paraphrase).

Despite these works that mainly focus on textual similarity, in terrorist event scenarios space and time components too can be considered as relevant information for duplicate events detection.

Several problems connected to scoring spatio-temporal data have already been studied in the fields of spatial keyword queries, time-dependent text queries and in sensor networks. In [1] the authors face the problem of efficiently processing spatial keyword queries with *AND semantic* after evaluating them. Cong et al. [2] proposed to use a hybrid index: an inverted file is associated to each R-tree node so that both location information and text can be used to prune the search space at query time. In [3] the authors introduced a new index named as Spatial Inverted Index (S2I) to efficiently process top-k spatial keyword queries. With regard to temporal ranking, the *time machine* proposed in [4], allows to retrieve documents according to keywords and those existing at a specific time. In [7] textual, temporal and spatial dimensions are combined all together.

3 Data sources

After the 11 September attacks, starting with GDT, many terrorism databases, such as WITS, CTC, ISVG, have been published, to make news collections publicly accessible for scholars and specialists. Our approach has been applied to a few free sources listed as follows:

1. CTC (Combating terrorism center) SENTINEL[3], in particular the short news contained below the section *Recent Highlights in Terrorist Activity* in the monthly journal Sentinel;
2. ISVG (The Institute for the Study of Violentv Groups)[4], a research center providing data about transnational terrorism through its Violent extremism Knowledge Base (VKB);

[3] CTC: http://www.ctc.usma.edu/sentinel
[4] ISVG: http://vkb.isvg.org/Special:IsvgSearch

3. NCTC (National Counter Terrorism Center), whose WITS (Worldwide Incidents Tracking System)[5] has been stopped since 2010.

Yet, there are also providers whose services are available with fee, the most popular is the Terrorism and Insurgency Centre by Jane's Information Group (JTIC). From these sources a common main structure of the news can be recognized, in particular it can be always identified a short description in natural language (text), a location, usually understood as the country (space), a date (time) and the publisher, that is the dataset source name, as in Table 1.

In ISVG and WITS, that provide data in a semistructured format, the city of the event is not detected neither, but only the country name. This because it is not always possible to extract this entity, due to a few reasons, mainly to misspelling or inaccuracy or, even, lack of this piece of information. Moreover, it has to be considered that often the date these sources provide is not correct, this because if an event happens late evening, the date is likely to be the date the news has been reported (that is the day after the current one), not the date the event has happened. The same piece of news may be identified with the first date in a source and with the second date (the real time of the event) in another one. For a deeper overview, Daniel J. Mabrey, executive Director of

Table 1: Similar events

id	description
n1	IRAQ - April 4, 2008 -On 4 April 2008, in As Sa'diyah, Diyala, Iraq, a suicide bomber detonated an improvised explosive device (IED) he was wearing near a funeral procession in Sed Himreen cemetery, killing 20 civilians and wounding 30 others. No group claimed responsibility." - WITS
n2	IRAQ - April 4, 2008 -April 4, 2008 (IRAQ): Iraq police officer funeral suicide attack kills nine people A SUICIDE bomber attacked a funeral of an Iraqi policeman in Hamrin in Diyala province on 4 April, Reuters reported. The blast killed nine people and wounded 30 others" - JTIC
n3	PHILIPPINES - May 29, 2008 - Blast in southern Philippines leaves three people dead On 29 May, a bomb blast struck a building in front of the Edwin Andrews Airbase in Zamboanga City in the southern Philippines, killing three people and injuring 17 others. According to the local Filipino newspaper the- JTIC
n4	PHILIPPINES - May 29, 2008 - Philippines bomb kills three people SUSPECTED militants detonated a bomb targeting a building in front of the Edwin Andrews Airbase in Zamboanga City in the southern Philippines on 29 May, killing three people and injuring 17 others, the Philippine Inquirer reported. - JTIC

the Institute for the Study of Violent Groups, has examined and compared these incidents databases in [8], where he highlights the differences among them in organizing terrorist information. In addition, terrorist events providers, besides these datasets, typically have lists and taxonomies about groups, weapons, people connected to terrorism

[5] WITS: https://wits.nctc.gov/FederalDiscoverWITS/index.do?N=0

and that can be used to create customized dictionaries of frequent and less-frequent domain words, as we explain later on.

4 Methodology

Given a terrorist event dataset as shown in Table 1, the goal is identifying the records representing the same event and joining them into a single record. Analysing just the space dimension can return inaccuracy, as well as analysing just the time dimension. On the other hand, the simple keyword analysis is not enough to our goal, consequently we need to use the three dimensions: space, time and text have to be intersected to get one record for one event.

Technically, we preprocess event records with NLP tools to extract relevant information in structured format when not already available: time, space, keywords. In our knowledge base the event is an ennuple (id; p; l; t; k) where id is a unique event identifier, p is the publisher, l is the city location, t the time when the news was published and k the list of keywords extracted from text. For instance, given semi-structured news as in Table 1 we extract structured data as in Table 2.

The event description includes place, event and casualty, seldom the group blaming the act. From the description we first aim at extracting a more precise location, instead of getting just the country name. This can be done as a brute-force approach by simple searching whether any upper case word in the text correspond to a region or city name in the given country, in particular we extract geographical places i.e. by Geonames. Indeed geonames web service provides a function to search for places by name[6]. In our table the first two records show *Himreen* and *Hamrin*, the same location, though Himreen is a less known name in the local language for the town of Hamrin. In this case, the most widespread name is used (Hamrin). Please notice that when it is not possible to extract unambiguously the location l from text, we use the centroid nation location.

Furthermore, we extract keywords, neither identified as stop words, nor as locations, characterizing the event description and corresponding to the type of attack. The idea is to use a dictionary with the type of event or the terrorist group list for finding related words. This is the way how we can extract *bomb, police, officer, suicide, blast, militants* from the examples above.

Even if it is not possible to know it beforehand, another interesting set of keywords is represented by the facility types, which can be extracted starting from a consistent dataset of old news of the same domain, e.g. the less frequent words we can find in our sources over one year. If we consider the example, we refer to words such as *airbase, funeral, procession, cemetery, bus*, as well as *market, hotel, school*

Identifying the keywords follows the text analysis and extraction and would require a long discussion, though this is not included in this paper's goals. For further information we may suggest to refer to the literature about this subject: NLP tools for Entity Recognition (NERs), e.g. Stanford NLP group NER[7] or tanl NER [9, 10], indeed the NER goal is finding all proper nouns in a text and classifying them into categories of interest as e.g. *location*, *organization* and so forth.

[6] Geonames search: http://www.geonames.org/export/geonames-search.html
[7] Standford NER: http://nlp.stanford.edu/software/CRF-NER.shtml

Table 2: Event database snapshot

id	publisher	location	time	keywords
n1	WITS	Diyala, IRAQ	April 4, 2008	suicide bomber IED funeral procession cemetery
n2	JTIC	Hamrin, Diyala, IRAQ	April 4, 2008	police officer funeral suicide attack bomber blast
n3	JTIC	Zamboanga City, PHILIPPINES	May 29, 2008	blast bomb building airbase
n4	JTIC	Zamboanga City, PHILIPPINES	May 29, 2008	bomb militants building airbase

At the second stage, given the dataset of event, we cluster data when closer for time and space and with high percentage of keywords expressing analogous concepts. The clustering works on the three publishers searching for different descriptions of the same event (E.g. the same terrorist act can be shorter in the first publisher database compared to the others or it can be characterized with the acronym IED in the first case and the compound word homemade bomb in the second one), complementary descriptions (E.g. a publisher can supplement a piece of news with details not contained anywhere) or updates as well.

The similarity function is the ST-IR ranking function from [7], as shown in the following.

Definition 1 (ST-IR rank). *Given a reference event n1 and a comparative event n2, the aggregation function τ returns a similarity score between n1 and n2, based on spatio-temporal and textual proximity:*

$$\tau(n1, n2) = \alpha \times \delta_s(n1, n2) + \beta \times \delta_t(n1, n2) + (1 - \alpha - \beta) \times \delta_w(n1, n2) \quad (1)$$

$$\textit{with } 0 \leq \alpha \leq 1 \textit{ and } 0 \leq \beta \leq 1 - \alpha.$$

$\delta_s(n1, n2)$ and $\delta_t(n1, n2)$ and $\delta_w(n1, n2)$, ranging between 0 and 1, represent three distance scores respectively for space, time and text.

To give an example, setting $\alpha = 0$ means not to give relevance to space rank and considering events similar only compared to time and text. As previously mentioned, these parameters can be modified to create different views on the basis of ones own needs.

Different functions can be used to model the spatial, temporal and textual score. For space and time proximity, we use Euclidean distance. The textual rank function is any of Jaccard or Cosine. It could be also interesting to consider other keywords having the same meaning by computing synonyms e.g. from Wordnet[8].

[8] Wordnet: http://wordnet.princeton.edu/

5 Conclusion and future works

This work highlights the importance of clustering similar terrorist events according to the three dimensions space, time and keywords, though each one should have a different weight in calcuating the overall score, based on qualitative and quantitative analysis.

By considering these three components, we can deduplicate events coming from different databases and aggregate them along more dimensions. Duplicate events detection has the advantage of counting exactly the events, avoiding duplicates and/or incomplete news.

As a future task, this approach could be implemented, optimized and tested over KBs enumerated in section 3.

References

[1] Chen, Y.y., Markowetz, A.: Efficient Query Processing in Geographic Web Search Engines. In: proc. of ACM Sigmod. (2006) 277–288

[2] G. Cong, C. S. Jensen, and D. Wu. Efficient Retrieval of the Top-k Most Relevant Spatial Web Objects. In: Int. Conf. on Very Large Data Bases (VLDB). (2009) 337–348

[3] Rocha-junior, B., Gkorgkas, O., Jonassen, S., Nø rv, K.: Efficient Processing of Top-k Spatial Keyword Queries. Proceedings of the International Symposium on Spatial and Temporal Databases, Springer, LNCS 6849 (2011)

[4] Berberich, K., Bedathur, S., Neumann, T., Weikum, G.: A time machine for text search. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07 (2007) 519

[5] G. Costa, A. Cuzzocrea, G. Manco and R. Ortale. Data De-duplication: A Review. *Learning Structure and Schemas from Documents*, Volume 375:385–412, 2011.

[6] F. M. Zanzotto, M. Pennacchiotti and K Tsioutsiouliklis. 2011. Linguistic redundancy in Twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 659-669.

[7] V. Cozza, A. Messina, D. Montesi, L. Arietta, and M. Magnani. Spatio-temporal keyword queries in social networks. In Springer, editor, *B. Catania, G. Guerrini, and J. Pokorn(Eds.): ADBIS 2013*, volume 8133 of *LNCS*, pages 70–83, 2013.

[8] Daniel. J. Mabrey Analyzing Terrorist Activities through Operational & Associational Coding of Events: Introducing the Institute for the Study of Violent Groups' Relational Database. Copyright 2010 - Institute for the Study of Violent Groups - All Rights Reserved

[9] G. Attardi, S. Dei Rossi, F. Dell'Orletta, E.M. Vecchi. The Tanl Named Entity Recognizer at Evalita 2009. In: Proc. of Workshop Evalita 2009, ISBN 978-88-903581-1-1, 2009.

[10] G. Attardi, G. Berardi, S. Dei Rossi, M. Simi. The Tanl Tagger for Named Entity Recognition on Transcribed Broadcast News at Evalita 2011. In B. Magnini et al. (Eds.), Proc. of Evalita 2011, LNCS 7689, pp. 116-125, 2012. ISBN 978-3-642-35827-2.