Relation Tracker - tracking the main entities and their relations through time

M. Besher Massri
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
besher.massri@ijs.si

Inna Novalija
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
inna.koval@ijs.si

Marko Grobelnik
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
marko.grobelnik@ijs.si

ABSTRACT

In this paper, we present Relation Tracker, a tool that tracks main entities [people and organizations] within each topic through time. The main types of relations between the entities are detected and observed in time. The tool provides multiple ways of visualizing this information with different scales and durations. The tool uses events data from Event Registry as a source of information, with the aim of getting holistic insights about the searched topic.

KEYWORDS

Information Retrieval, Visualization, Event Registry, Wikifier, Dmoz Taxonomy

1. INTRODUCTION

Every day, tremendous amounts of news and information are being streamed throughout the Internet, which is requiring the implementation of more tools to aggregate this information. With technology advancement, those tools have been increasing in complexity and options provided. However, there has been a demand for tools that give simple yet holistic summary of the searched topic in order to acquire general insights about it.

Hence, we provide the Relation Tracker tool that tries to achieve this goal; it is based on the data from Event Registry [1], which is a system for real-time collection, annotation and analysis of content published by global news outlets. The tool presented in this paper takes the events and groups them into topics, and within each topic, it provides an interactive graph that shows the main entities of each topic at each time and the main topic of relations between those entities. In addition, a summary information about entities and their relationship is visualized through different graphs to help understand more about the topic.

The remainder of this paper is structured as follows. In section 2, we show the related work done in this area. In section 3, we provide a description of the used data. Section 4 explains the methodology and main challenges that were involved in this work. Next, we explain the visualization features of the tool in section 5. Finally, we conclude the paper and discuss potential future work.

2. RELATED WORK

Similar works have been done in the area of visualizing information extracted from news. We see in [2] a tool for efficient visualization of large amount of articles as a graph of connected entities extracted from articles, enriched with additional contextual information provided as characteristic keywords, for a quick detection of information from the original articles.

Regarding classifying news, we observe in [3] a new technique that uses Deep Learning to increase the accuracy of prediction of online news popularity.
In the paper explaining Event Registry [1], we see how articles from different languages are grouped into events and the main information and characteristics about them are extracted. Additionally, a graphical interface is implemented which allows search for events and visualize the results in multiple ways that together give a holistic view about events.

This work begins with the events as a starting point, and it is one more step on the same path; it groups events further into topics and trends, then it focuses on tracking how some entities are appearing as main entities regarding the selected topic, and how the relationship between them is changing through time.

3. DESCRIPTION OF DATA

We used part of the events from Event Registry as our main source of data. We obtained a dataset of ~ 1.8 million events as a list of JSON files, with event's dates between Jan 2015 and July 2016. Each event consists of general information like title, event date, total article count, etc., and a list of concepts that characterize the event, which is split into entity concepts and non-entity concepts. Entity concepts are people, organizations, and locations related to the event. Whereas non-entity concepts represent abstract terms that define the topic of the event, like technology, education, and investment. Those concepts were extracted using JSI Wikifier [4] which is a service that enables semantic annotation of the textual data in different languages. In addition, each concept has a score that represents the relevancy of that concept to the event.

4. METHODOLOGY

4.1 Clustering and Formatting Data

To group the events into topics, we used K-Means clustering algorithm, where each event is represented as a sparse vector of the non-entity concepts it has, with the weights equal to their scores in that event. The constant number of topics is set experimentally to be 100 clusters, in a balance between mixed clusters and repeated clusters. Each cluster describes a set of events that fall under the same topic, whereas the centroid vector of each cluster represents the main characteristics of it. To name the clusters, we used category classifier service from Event Registry, which uses Dmoz Taxonomy [5], a multilingual open-content directory of World Wide Web links, that is used to classify texts and webpages into different categories; for each cluster, we formed a text consisting of the components of its centroid vector, taking into account their weights within the vector. The resulted cluster names were ranged from technology and business to refugees and society, and clusters were exported as a JSON file for processing them in the visualization part.

4.2 Choosing the Main Entities

Under any topic, the top entities at each duration of time has to be chosen. At first, the concepts were filtered from outliers like publishers and news agencies. Then, an initial importance value has been set for each concept based on two parameters: the TF-IDF score of concept with respect to each event, and the number of articles each event contains. If we denote the set of events that occur in the interval of time D by E_D, the number of articles that event e contains is A_e, the TF-IDF score of concept c at event e by $S_{c,e}$, then the importance value of each item with respect to the interval D is calculated by the formula:

$$Imp_{init}(c)_D = \sum_{\substack{e \in E_D \\ e\ has\ concept\ c}} S_{c,e} \quad * \sum_{\substack{e \in E_D \\ e\ has\ concept\ c}} A_e$$

The TF-IDF function is used to give importance to the concept based on its relevance to the events, and the number of articles is used to give more importance to the events that have more articles talking about it, and hence, more importance to the concepts that it has. We decided on using the product of summation rather than summation of product because of its computation efficiency while still producing good results. However, to prevent the case where all the chosen entities get nominated because of one or two big events (which results in a bias towards those few events), a modification to the importance value formula has been made by introducing another parameter, which is the links between concepts (whenever two concepts occur in the same event, there is a link between them). Each concept now affects negatively the other concepts it is linked to by an amount equal to the initial importance value divided by the number of neighbors. If we denote the set of neighbors of concept c during the interval of time D by $N_{c,D}$, then the negative importance value is defined by:

$$Imp_{neg}(c)_D = \sum_{c' \in N_{c,D}} \frac{Imp_{init}(c')_D}{|N_{c',D}|}$$

The final score is just the initial importance value minus the negative importance value, which is then used to sort and nominate the top entities.

$$Imp_{final}(c)_D = Imp_{init}(c)_D \; - Imp_{neg}(c)_D$$

4.3 Detecting the Characteristics of Relationship

The main goal was to model the relationship between any two entities through a vector of words where two entities are collocated. Since the relationship between two entities at any given time is based on the shared events between them, and each event is characterized by a set of concepts, we decided on using those concepts - specifically the abstract or the non-entity concepts - to characterize such relationships. For each pair, we aggregated all the non-entity concepts from the shared events between them, and each one of them was assigned a value based on the number of events it is mentioned in and its score in those events. Those concepts were sorted and ranked depending on their values, and the top ones were chosen as the main features of the relationship. In addition, these values of the concepts were used to rank the shared events and extract the top ones; by giving each event a value equal to the aggregated values (the ones calculated in previous step) of all non-entity concepts it has. To summarize the set of characteristics, we classified them using Dmoz category classifier in a similar way to what we have done in determining the names of the clusters. These categories were used to label the relationship between the entities, indicating the main topic of the shared events between them.

5. VISUALIZING THE RESULTS

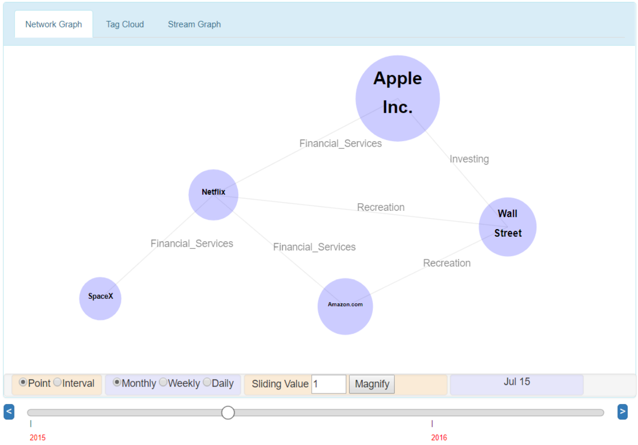
To access a topic, a search bar is provided to select among the list of extracted topics from clustering step. Once the user selects a topic, a default date is chosen and a network graph is shown explaining the topic.

5.1 Characteristics of the Main Graph

Since the tool's main goal is to show the top entities and their relations, the network graph is the best choice for this matter. Following that, we have built an interactive network graph that has the following features:

- The main entities within that topic at the selected interval of time are represented by the vertices of the graph.
- The size of the vertices reflects the importance value of each entity, scaled to a suitable ratio to fit in the canvas.
- The colors represent the type of the entity, whether it is a person [red] or an organization [blue].
- The links between the entities represent the existence of shared events in that interval of time between them under that topic, and hence indicating some form of relations. The thickness of the links is proportional to the number of shared events, whereas the labels are the ones calculated in previous section.

Figure 1 presents top companies with relevant relations in July 2015 found among business news.



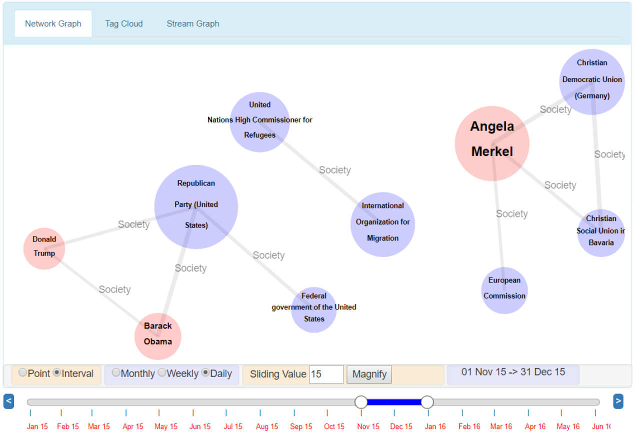
● Point ○ Interval ● Monthly ○ Weekly ○ Daily Sliding Value 1 Jul 15

Figure 1: Top companies in July 2015 and their relations under the business topic.

5.2 Main Functionality

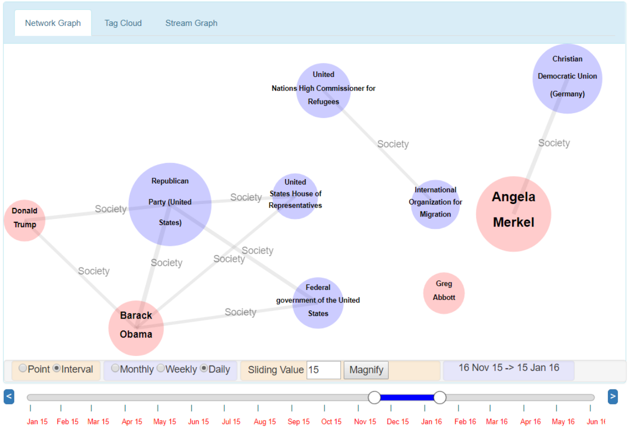
As the tool is concerned about tracking the changes with time. The graph is supported with a slide bar that allows the user to choose from the dates where there is at least one event occurred with respect to the selected topic. Different scales for moving dates are also provided; the user can choose to move day by day, week by week, or month by month and see the changes accordingly. In addition, the user can choose a specific interval of time, and track how the entities and their relations are changing when the interval moves slightly with respect to its length. An interval magnifier is also given if the user wants to get a closer look at the changes that happen in a small interval.

An example illustrating that can be seen in Figures 2 and 3. In Figure 2, we see the top 10 entities under the refugee topic in the last two months of 2015. When the interval is moved by 15 days, we notice that some of the entities disappear, like European Commission, indicating that they are no longer among the top 10 entities, whereas "United States House of Representitive" entity emerges and connects to "Barack Obama" and "Repulican Party". The change in size indicates the change in the importance value of each one, while Society is the general theme among all labels.



○ Point ● Interval ○ Monthly ○ Weekly ● Daily Sliding Value 15 01 Nov 15 -> 31 Dec 15

Figure 2: Top entities for the last two months of 2015 under the refugee topic.

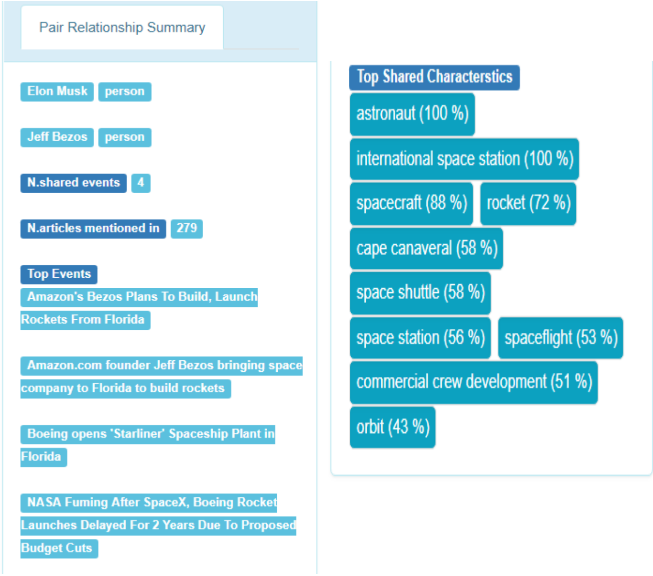


○ Point ● Interval ○ Monthly ○ Weekly ● Daily Sliding Value 15 16 Nov 15 -> 15 Jan 16

Figure 3: The changes in top entities under the same topic after moving the interval for 15 days.

5.3 Displaying Relation Information

Whenever the user selects a pair of entities, detailed information about their relationship in the selected interval of time is given, such as the number of shared events and articles, along with the top events both concepts were mentioned in. Also, the top shared characteristics that shape the relationship between them at this period is shown and sorted by percentage of importance. As seen in Figure 4; when selecting Jeff Bezos and Elon Musk under the space topic between January and September 2015, we see a list of the top events that involve both of them during this period. We see also that the relationship between them is mainly about sending astronauts by rockets to the international space station, as it can be understood from the top shared characteristics.



Pair Relationship Summary

Elon Musk person

Jeff Bezos person

N.shared events 4

N.articles mentioned in 279

Top Events

Amazon's Bezos Plans To Build, Launch Rockets From Florida

Amazon.com founder Jeff Bezos bringing space company to Florida to build rockets

Boeing opens 'Starliner' Spaceship Plant in Florida

NASA Fuming After SpaceX, Boeing Rocket Launches Delayed For 2 Years Due To Proposed Budget Cuts

Top Shared Characterstics

astronaut (100 %)

international space station (100 %)

spacecraft (88 %) rocket (72 %)

cape canaveral (58 %)

space shuttle (58 %)

space station (56 %) spaceflight (53 %)

commercial crew development (51 %)

orbit (43 %)

Figure 4: Relationship summary about Jeff Bezos and Elon Musk between January and September 2015 under the Space topic.

To illustrate how the importance of those top features with respect to the relationship is changing through time, a stream graph is used as shown in Figure 5.

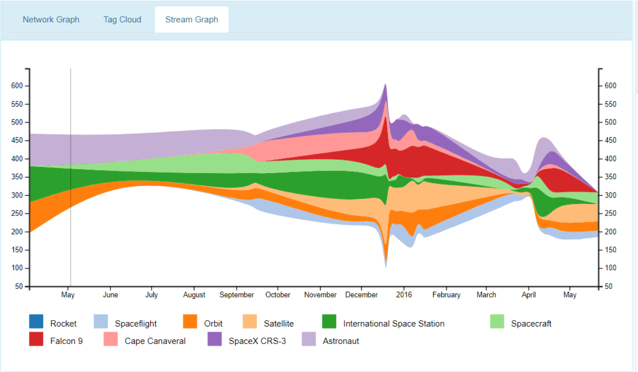


Figure 5: Stream graph showing how the effect of the main features on the relationship between Jeff Bezos and Elon Musk is changing through time.

Finally, the set of all characteristics that affect the relationship is visualized in a tag cloud to give a big picture about it. Figure 6 shows the tag cloud of the same relationship mentioned above.

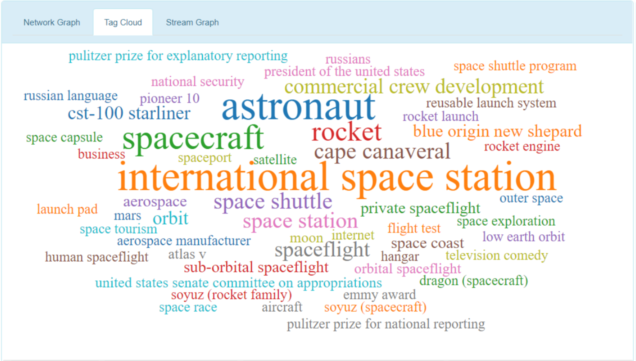


Figure 6: Tag cloud illustrating a general view about all the characteristics that affects the relationship between Jeff Bezos and Elon Musk under the space topic.

6. CONCLUSION AND FUTURE WORK

In this paper, we provide a tool that uses events data from Event Registry to show the main entities within each topic, and how the characteristics of relationship among them is changing through time. However, there are a couple of limitation to the tool that we want to improve in the future. Although we were able to detect the characteristics of the relationship between entities and how they are changing through time, the main type of relation that we used to label the links were very broad and hence rarely changing-improving the methodology for relation extraction and observation of relations in time will be the subject of future work. In addition, we limited the search space for topics for the 100 topics we obtained from clustering, we would like to generalize the search by enabling searches for any concept or keyword with different options to filter the search.

7. ACKNOWLEDGMENTS

This work was supported by the euBusinessGraph (ICT-732003-IA) project [6].

8. REFERENCES

[1] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion). ACM, New York, NY, USA, 107-110. DOI: https://doi.org/10.1145/2567948.2577024

[2] Marko Grobelnik and Dunja Mladenić. 2004. Visualization of news articles. Informatica 28.

[3] Sandeep Kaur and Navdeep Kaur Khiva. 2016. Online news classification using Deep Learning Technique. IRJET 03/10 (Oct 2016).

[4] Janez Brank, Gregor Leban and Marko Grobelnik. 2017. Annotating documents with relevant Wikipedia concepts. In Proceedings of siKDD2017. Ljubljana, Slovenia.

[5] Dmoz, open directory project, http://dmoz-odp.org/ (accessed in July, 2018)

[6] euBusinessGraph project, http://eubusinessgraph.eu/ (accessed in July, 2018).