Type of the presentation: In-use

Title of the presentation: Global media analytics using Event Registry

Summary of the presentation:

Event Registry (http://eventregistry.org/) is a system that contains a repository of world events that are automatically identified by analyzing news articles published globally. For each event it can extract from articles core information, such as event location, date, who is involved and what is it about. The system provides extensive event search options, numerous visualizations and aggregations of search results. Event detection is done in real time and currently supports information extraction from 15 languages.

Extended abstract:

There are thousands of news articles written and published every day by news publishers and agencies all across the world. They are written in various languages and discuss all possible topics. A large percentage of these articles are discussing world events - current, past and future. There is no generally accepted definition of an event, but one intuitive definition is that an event is any significant happening in the world. Two instances of an event are, for example, Felix Baumgartner's jump from a helium balloon on October 14, 2012 and bombings during the Boston marathon on April 15, 2013.

The unstructured form in which information is provided in news articles is human-friendly, but it's not easy to process for computers. Since there is no structured representation of the information provided by the news articles, it is very hard to identify events based on their location, topic, date or some relevant set of entities. To help in this task we present in this paper a system called Event Registry. It is able to collect news articles from thousands of news sources and identify in them the events that are being discussed. Information about the events is automatically extracted from the articles and stored in a database. The events can then be found by specifying search conditions such as an entities, topics, locations or dates. Events matching the criteria can be listed as well as summarized and visualized in different ways in order to provide additional insights. In the rest of the paper we briefly describe the individual parts of the pipeline and the possible uses of the system.

Data collection and preprocessing

For collecting data we use the NewsFeed[1] service which collects news articles from around 100.000 news sources. The number of collected articles ranges between 100.000 and 200.000 articles per day. The collected articles are in various languages, where most represented languages are English, German, Spanish and Chinese. The collected articles are then processed with a set of natural language processing tools. These are responsible for named entity detection and linking, date identification and categorization into a topic taxonomy. Since collected articles are in various languages we also perform cross-lingual analytics. By applying canonical correlation analysis we are able to compute similarity between articles in different languages. This functionality (which is available for over 100 languages) allows us to identify the most similar articles other languages which is important in the later task of cross-lingual event extraction.

Event construction

After the articles are enriched we try to find groups of articles describing the same event and extract from the articles event information. In order to identify groups of articles describing the same event

[1] http://newsfeed.ijs.si/

we implemented an online clustering algorithm. Learning features used for clustering are based on the article title, body, date and detected named entities. Each new document is compared with the existing centroids of the clusters. If the closest centroid is similar enough, the article is put into the existing cluster, otherwise a new cluster is created consisting of the article. As soon as a cluster reaches a certain size (depending on the language, usually around 5 articles) we start considering the group of articles as an event. The intuition for treating such a group of articles as an event is that since a sufficient number of news publishers wrote similar content then it's very likely that the content of the articles describes an event that occurred.

Clustering of articles is done separately for each language. Since clusters in different languages can describe the same event we provide methodology that is able to identify clusters describing the same event and merge them into a single event. The approach is based on computing how similar articles in the two tested clusters are, as well as taking into account the similarity of entities mentioned in the two clusters. Using an SVM model we are able to achieve 90% accuracy in merging of clusters.

Event information extraction

When a sufficiently large cluster of articles is identified an event with a unique id is constructed and the articles from the cluster are assigned to it. To extract event information we analyze the articles assigned to the event. Event title and a short text snippet are determined by finding the article closest to the center of the cluster (medoid article) and using its title and first paragraph. For the event date we analyze the detected date references in the articles. If the most frequently detected date is frequent enough then we use it as the event date. If no date passes the threshold then we use the average date of the article in the cluster as the event date. To determine the event location we find frequently detected named entities that are known to be locations (based on GeoNames). An SVM model is again used to determine which location is the event location based on the entity frequency and their text position. As a way of summarizing who and what the event is about we analyze all detected named entities and keywords in the articles and compute their weight based on their frequency. Events are also about different topics (sports event vs. bombing report). To categorize the events we used the DMoz taxonomy which provides categorization into 5.000 different topics. Events with all the extracted information about them are then stored in the Event Registry database and are searchable using the search API.

How to use Event Registry

Event Registry can be used through the web interface or through its REST API. The web interface provides ways for identifying events of interest based on various search criteria, such as searching by entity, keyword, topic, event location, date and size. Search results can also be visualized in numerous ways, such as by displaying map of event locations, a timeline of events, clusters of events and trending of concepts and topics. Examples of some visualizations are shown in the Figure 1.

If the information is to be used for additional analysis, the data can also be accessed using the API. The API access offers the same functionalities as the web interface, except that the resulting data is provided in JSON format.

Functionalities provided by Event Registry can be used in various ways. Some examples include:

- Real-time stream of events. Events are identified immediately after articles about them are written which provides very small latency in event detection. When an event is identified, the core information about the event (what, where, when, who) can be provided in structured form.

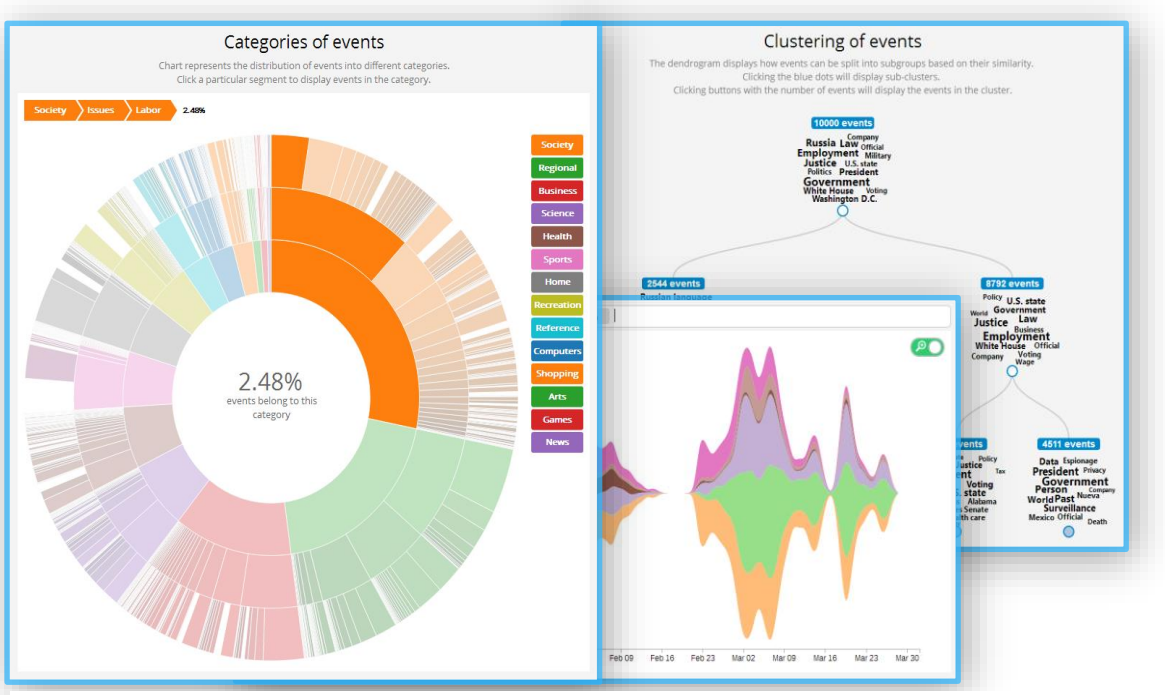

Figure 1 Example visualizations from Event Registry

- Custom stream of articles and events based on interests. Based on user's interests, a custom feed of articles and events can be obtained by specifying topics of interest.
- Detection of trends. By analyzing new articles we can identify what are the current entities and topics that are written about. By monitoring social activity we can also identify the articles and events that are shared the most on social media.
- News bias. Using the data in Event Registry we can compute various aspects in which news publishers are biased in their reporting. The types of biases include geographical bias (where events happened), topical bias (what types of events), objectivity bias (sentiment analysis), citation bias (what sources they cite) and others. We can also detect parts of the story they omit as well as detect plagiarism.
- Information diffusion. Having extensive coverage about individual events allows us to analyze how information spreads through different channels. We can identify which sources are hubs, which are opinion makers and which are simply repeating information from others.
- Nowcasting. Using the trending information about various topics and keywords mentioned in the news we are able to compute correlations as well as do predictive analytics of various socio-economic indicators.

Contributor names:

Gregor Leban is a senior researcher in the field of artificial intelligence, machine learning, information extraction, text mining and data visualization. He received his PhD. for his dissertation on the use of machine learning methods for automatic identification of interesting data visualizations. He is a key contributor for a popular open-source machine learning toolkit called Orange (http://orange.biolab.si/). He is the author of several conference papers and papers published in high-ranking SCI journals like Data Mining and Knowledge Discovery and Bioinformatics. He is also a major contributor on several FP7 European projects and the main author of the Event Registry system (http://eventregistry.org/).

Blaž Fortuna is a senior research assistant and a PhD student at JSI in the area of kernel methods, statistical learning and semantic Web with strong focus on text analysis. In the recent years he had several publications at international conferences and developed several software modules for scalable machine learning, cross-lingual information retrieval and classification, ontology learning and active learning which are part of Text Garden software environment. He is also major author to the OntoGen (http://ontogen.ijs.si) system for ontology learning and Document Atlas (http://docatlas.ijs.si) text visualization software.

Marko Grobelnik is an expert in the areas of analysis and knowledge discovery in large complex databases. In particular, the areas of expertise comprise: Data Mining, Text Mining, Semantic Technologies, Network Analysis, and Complex Data Visualization. Marko collaborates with major European and US academic institutions and consults industries such as British Telecom, Microsoft Research, Nature, New York Times, Bloomberg, and Accenture. Marko is author of several books in the area of machine learning, data mining, text mining and semantic technologies and authors of many scientific papers. He is also W3C AC representative for JSI, CEO of the company Quintelligence and co-founder of the company Cycorp Europe.