Event Evolution and Archiving

Omar Alonso
Microsoft
omalonso@microsoft.com

1. INTRODUCTION

Over the past couple of years, social networks like Twitter and Facebook, have seen phenomenal growth, making them one of the largest sources of public opinion and real time information on the Internet. Trending topics and hashtags provide a very strong signal for important events that would later be covered in news articles. If the event is of general interest, it is highly likely that it would eventually have its own Wikipedia article. This information flow, from social posts (e.g., tweets or Facebook posts) to a final encyclopedia-like entry through a series of news articles is a new way of producing and consuming content.

Social posts provide real-time information but, at the same time, can be overwhelming for two reasons: there is no context for the uninformed user and there is a lot of noise for the informed user looking the latest update. News articles are less frequent but provide more context and are ranked by topicality and freshness. Wikipedia articles are curated documents that described the topic to a certain extent but maybe difficult to consume when looking at the progression of an event.

Current tools are not very well suited for discovering situations, summarizing unfolding events, or providing the evolution of a story and related topics. We are interested in combining social and news data to construct a new document that captures the backbone of a story over time with all associated information. More specifically, we would like to provide a system for searching and retrieving stories about numbers (e.g., "earthquakes over 7.5 in summer", "acquisitions of 1B dollars"), on-going events (e.g., "Brexit", "US elections"), or entities (e.g., "Barack Obama", "Microsoft") that are presented in a sequence of related events in temporal order. In contrast to knowledge-based solutions that retrieve facts, we are interested in stories and related subtopics.

We describe current efforts at Microsoft to develop infrastructure that supports querying and retrieving evolving stories about events. The aim of the system is to fulfill the above information seeking scenarios by algorithmically generating the core of the story as it evolves over time by using selected relevant content derived from social data and news articles. This new data asset is a new type of document that is not as encyclopedic-centric like a Wikipedia entry but, instead, more dynamic to the many data components of the story, always up-to-date, and constructed in a fashion that allows different aggregations and applications.

We define a story as a document that contains the following data elements:

- Named-entities: people, places, and organizations mentioned in the story.
- Relevant text: fragments or paragraphs that describe a concept or elaborate on certain specifics.
- Temporal information that anchors a salient matter in time.
- Quantitative information that can be used to synthesize specific facts.

While identifying numbers in text is easy, detecting numbers we can trust can be a difficult task, especially when the source is large and noisy. Extracting quantfrags, a small piece of text which contains quantitative information, can be very useful for identifying precise data points (e.g., "2011 earthquake magnitude 9.0", "12 people killed in Charlie Hebdo attack") that can serve for comparison purposes or to find related and relevant information and/or stories [2].

StoryDB is a system that consists of the following components:

- Generation of a topical timeline using judgment aggregation
- Extraction of relevant quantitative information
- Construction of a synthetic document per story
- A query language that allows the retrieval of stories by different sizes and granules

We implemented a back-end pipeline that generates the sketch of the story by removing a large percentage of the noise from social data. The initial sketch is derived using a combination of efficient filters, exact and near-duplicate content removal and clustering that provide substantial savings in storage and subsequent computational costs by reducing the size of the data while preserving majority of the interesting content by using a number of machine learning classifiers.

Once the sketch is constructed, an algorithm that uses social information as relevance surrogates to generate an informative timeline in conjunction with the entities and quantfrags, produces the story. Stories are updated given the frequency of the event, which is indicated by its trending activity in social data.

We hypothesize that a system that captures the evolution of a story as perceived by the crowd in social media along with editorially edited articles, provides a new way of consuming and archiving information on a topic with diverse perspectives. StoryDB can also be thought as underlying infrastructure to support information cartography [1].

2. REFERNCES

[1] D. Shahaf, C. Guestrin, E. Horvitz, J. Leskovec. "Information Cartography", CACM 58(11), 2015.

[2] T. Sellam and O. Alonso. "Raimond: Quantitative Data Extraction from Twitter to Describe Events", Proc. of ICWE 2015.