Recorded Future
UNLOCK THE PREDICTIVE POWER OF THE WEB

Big Data for the Future: Unlocking the Predictive Power of the Web

Staffan Truvé, PhD
truve@recordedfuture.com

"Thus, what enables the wise sovereign and the good general to strike and conquer, and achieve things beyond the reach of ordinary men, is foreknowledge."(from The Art of War by Sun Tzu, Section 13)

Introduction

What if you could have advance knowledge of important events, or at least an improved understanding of what might happen in the near future? How could your organization change its operations if it knew in advance when its competitors would launch new products, or when there was an increased risk of a supplier not being able to meet its commitments?

Even after having existed for more than one and a half decades, the web continues to grow at tremendous speed, reflecting every aspect of human life, from large corporate mergers and natural disasters all the way down to someone having a bad cold or contemplating buying that new orange juice brand. We have truly entered the age of “big data”, where the primary challenge is not getting access to enough information but rather making sense and full use of that data. Niche examples of this is Google’s Flu Trends (http://www.google.org/flutrends/) and the fairly successful attempts to use twitter to predict box-office sales, but we generalize this to be doable in any domain.

Recorded Future is founded on the belief that there is predictive power in all the information that is being published on the web, and that it is just waiting to be unlocked. By applying temporal analytics to what is written about the future, and by algorithmically crowd-sourcing the resulting temporal information, we can draw conclusions and gain insight.

Therefore, we have developed a system, which analyzes information from many different sources on the web, along with tools which use that information to assist our customers in identifying and understanding historical developments and in formulating hypotheses about likely future events. We like to use the term “temporal analytics” to describe the time-oriented analysis tasks supported by our systems.

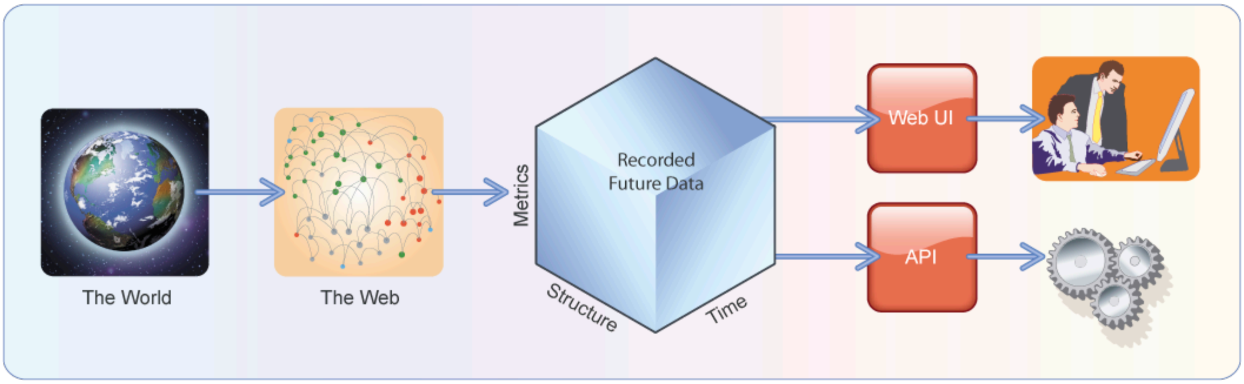
It is worth emphasizing that, unlike search engines, our mission is not to help our customers find web sites or documents, but to enable them to understand what is happening in the world - to record what the world knows about the future and make it available for analysis!
In the end, we of course always provide links to the underlying data to give credibility to our analysis.

Today, Recorded Future continuously harvests and performs real time analysis of news from more than 40,000 sources on the web, ranging from big media and government web sites to individual blogs and selected twitter streams. We have millions of entities and events, and more than 2 billion facts (timed entity and event instances) in our database.

This white paper describes the underlying philosophy and some of the key technologies of Recorded Future and its products. Our intended target audience is someone with basic technical understanding and, most importantly, a deep interest in using data to improve the operations of a company or organization.

Three Steps - Structure, Time, Metrics

The Internet contains a multitude of sources, ranging from the web sites of governments and "big media" all the way to individual bloggers and tweeters. To be able to use the massive volumes of unstructured text of the Internet for predictive analysis, we need to bring some order to it. We store this data in a "cube", and give access to it through our web interface and programmer's API. The focus of the web interface is to allow users to slice and dice aggregate views of the data, while the focus of the API is to allow our customers to apply statistics and machine learning algorithms to build models of the world. Both the UI and the API can thus be used to analyze historic events and predict likely future events - manually or algorithmically.



We have identified three major steps to bring order to all the textual information we harvest from the net:

- Structure
    - From unstructured text we extract structured information about *entities* (persons, organizations, companies, places etc.) and *events* (meetings, natural disasters, company acquisitions, political protest events etc.). We combine that with *ontological* information about the world (who is the leader of a certain country, what technologies can be used for night vision, in which country is a certain city located, etc.) to give a structured view of the world and events in it.

- Time
    - Temporal analysis allows us to detect what actual calendar time a text is referring to when an event is described. This can be both absolute times ("9:37AM, September 11, 2001", "at the end of 2012") and relative times ("three weeks from yesterday", "tomorrow"). Information about the publication time of a text combined with our linguistic analysis is used to map all events to a calendar time interval.

- Metrics
    - To be able to decide which information is most important, we calculate a number of metrics - numeric attributes of entities and events. A key metric is *momentum*, which is computed based on the volume of news around a certain entity or event, and also on the contexts in which is mentioned. Other metrics include sentiment (positive or negative tone, use of violent language, etc.).

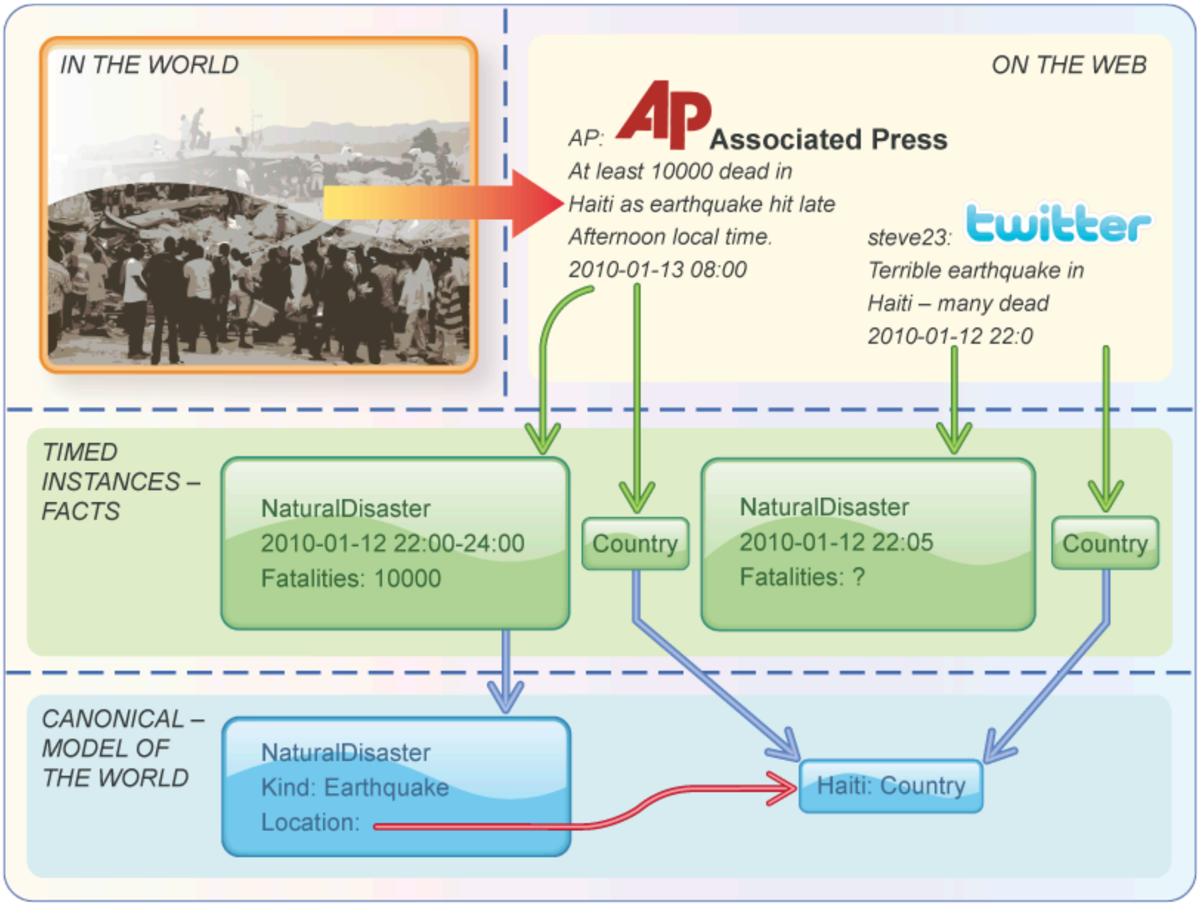
When we have structure, time and metrics, we can use this information to answer questions no other system can address, like “Which heads of state visited Libya in 2010?”, “What pharma companies are releasing new products in the first quarter of 2012”, and “What do French bloggers say about the earthquake in Haiti?”.

The following sections describe these three components in more detail. In this presentation we will primarily use English examples, even though Recorded Future today also analyzes sources in Chinese and Arabic, and is in the process of adding support for additional languages.

Structure

Recorded Future goes beyond explicit link analysis and adds implicit link analysis, by looking at the “invisible links” between documents that talk about the same, or related, entities and events. It’s similar to Plato’s distinction between the real world and the world of ideas (http://en.wikipedia.org/wiki/Platonic_idealism), or the mirror worlds of David Gelerntner.

We model the world by separating the documents and their content from what they discuss – the *canonical* entities and events, which correspond to people, companies, meetings, business transactions, political activities etc. in the real world. For each canonical entity and event there will be a number of timed instances, that relate mentions of the canonical entities/events to the source text where they were mentioned and the time they occurred (see example below).



Some entities are actual, physical things like people, places, and products. Other entities are more abstract, like technologies, companies, organizations, and positions. On top of this we maintain an ontology, which describes “the way the world is organized” - what entities exist, how can they be grouped and related within a hierarchy, and subdivided according to similarities and differences. For example, we keep track of which cities and regions are located in which countries, who the heads of state and ministers of governments are, what technologies can be grouped into a certain technology area, and so on. We derive this ontology both from our general information analysis, and from harvesting special sources, e.g. for geographical relationships, corporate officers, and world leaders.

Events describe things happening in the world and are organized by types. These could include earnings events -- “The earnings report for GM next week is expected to be a disappointment”. Or they may be disasters, donations or meetings -- “An earthquake hit Japan,” “Paris Hilton said that she plans to donate more money to charity,” “Steve Jobs and Steve Ballmer are meeting in London tonight.” Each event type is characterized by a number of attributes, and each event instance is associated with a time interval.

Some events can range over long, imprecise time periods (such as a product being released sometime in 2012), whereas other are very short and precise in time (such as a scheduled earnings call for a public company). Events can be both historic (things that have already happened) and planned/expected for the future. Future events are either scheduled (such as election days or earnings calls conferences) or speculative (such as guesses about future product release days or assumptions about when we reach “peak oil”).
Some events may also affect our ontology. A career change event may tell us that a CEO has stepped down or Sarah Palin has become a presidential candidate. An event such as this would tells us to modify a person’s Position attribute in the ontology.

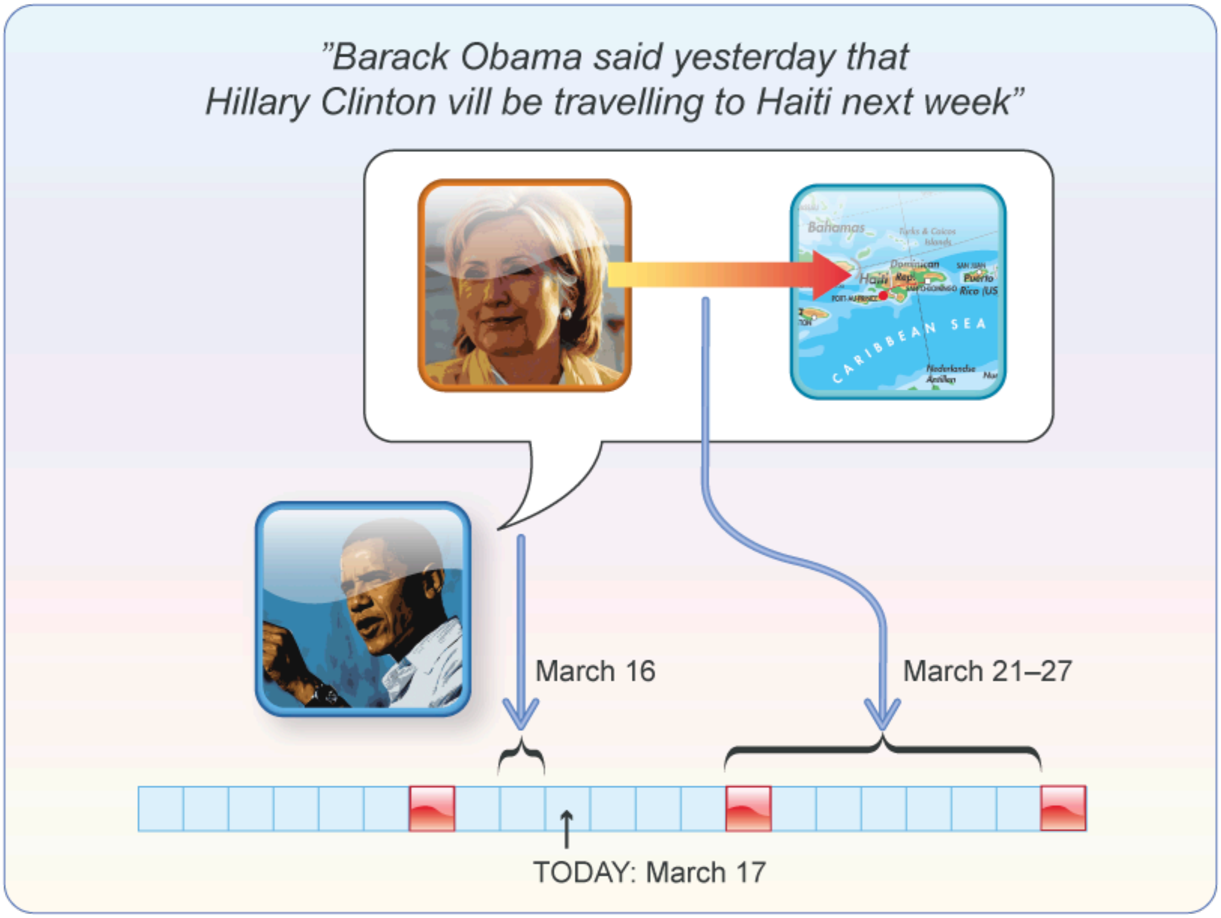
Time

Recorded Future also introduces a *temporal analytic* to documents. Each document contains references to not only publication date, but also unstructured indications of when an event has taken place or is expected to take place. These references compose the *temporal analytic.*

All temporal information is represented as a *time interval* (from START to STOP, e.g. from 2012-04-01 to 2012-05-30) and with a certain *precision*, indicating how exact the information is (e.g. year, month, day, or minute). All time is represented as Coordinated Universal Time (UTC). Temporal information in different time zones is transformed into UTC before being represented in the system.

Time expressions can be either absolute (“9:37AM, October 12, 2008”, “at the end of 2012”) or relative (“three weeks from now”, “tomorrow”). To make sense of relative time references we need to know when a text is written, and in most cases we use the publication time of a document to set that point in time.

Here is a sample text (“*Barack Obama said yesterday that Hillary Clinton will be travelling to Haiti next week.*”) and its representation as two events (a quotation event in the past and a travel event in the future) on a timeline:


”Barack Obama said yesterday that
Hillary Clinton vill be travelling to Haiti next week”
Bahamas
Haiti
CARIBBEAN SEA
March 16
March 21–27
TODAY: March 17

Recorded Future uses a rule based system to analyze text and infer what actual calendar time is intended. We can often identify a precise date or time in the temporal expression, but not always. Expressions such as “early 2011” and “late tomorrow night” are somewhat fuzzy, but we translate them into fixed calendar time intervals and set the precision to reflect their imprecise nature.

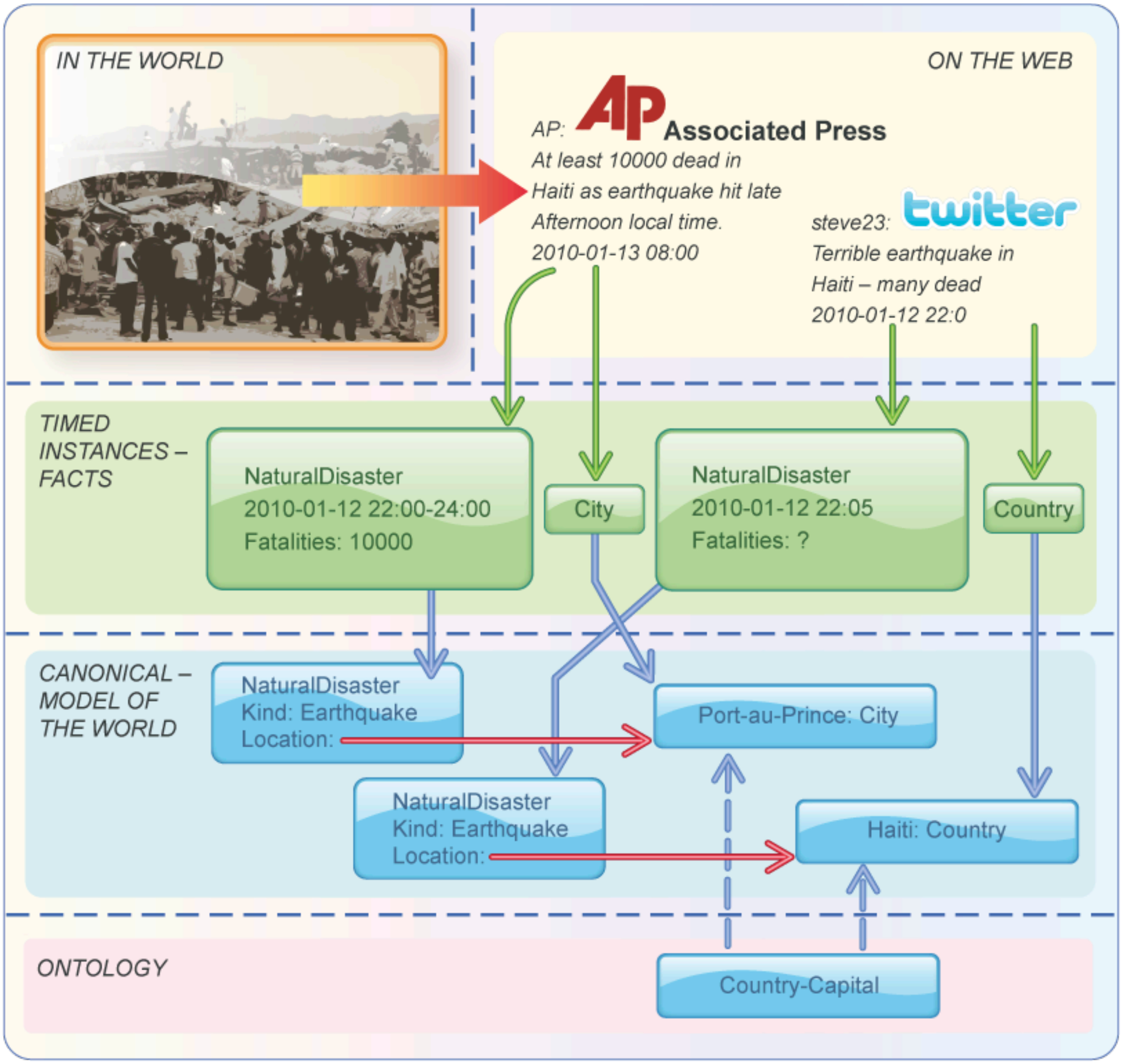
Culture also plays a role in time analysis - for example, when analyzing Chinese texts the Chinese lunisolar calendar needs to be taken into account, and for Arabic text the Islamic calendar must be understood and analyzed. Even for the western world there are differences, such as the first day of the week being Sunday in most countries and Monday in others - including the international ISO standard ISO-8601 for time. This can actually affect the interpretation of a simple expression such as “next week” uttered on a Sunday.

Similarly, dates are sometimes written as MONTH/DAY, and sometimes as DAY/MONTH - without the correct cultural context an expression like 5/12 is impossible to interpret unambiguously. Some dates actually become names of events, like “9/11”, and must be treated as such rather than actual dates. On twitter, some important events get hash-tagged by their date, like the Egyptian revolution’s “#jan25”.

The extensive rule set used by Recorded Future to analyze time in different languages is continuously updated, to capture more and more of the richness of human language.

Example

Enough theory - here's a more complex version of our Haiti example of how structure and time is extracted and represented:



A disastrous earthquake hits Haiti on January 12, 2010, at 16:53 local time (21:53 UTC). This is reported almost instantaneously by an alert twitter user (steve23), and also gets picked up next morning by AP (and many other sources, not shown here).

Recorded Future harvests both these news items, and uses its linguistic and temporal analysis to determine that they refer to two canonical events, both of type NaturalDisaster and kind Earthquake, but with different locations (Port-au-Prince vs. Haiti). Our ontological information does however tell us that Port-au-Prince is the capital of the country Haiti, so we can infer that these two events are related. Canonical NaturalDisaster events contain a location, which allows us to group together all such events occurring at a certain location. The number of fatalities and the time of the event is however not part of the canonical representation, but kept as part of the instance data. Note also that for the tweet, the publishing time is used as event time (since we have no other information to use), but for the AP article we apply our temporal reasoning to deduce that the event occurred sometime late in the evening (UTC) on the day before the news item was published.

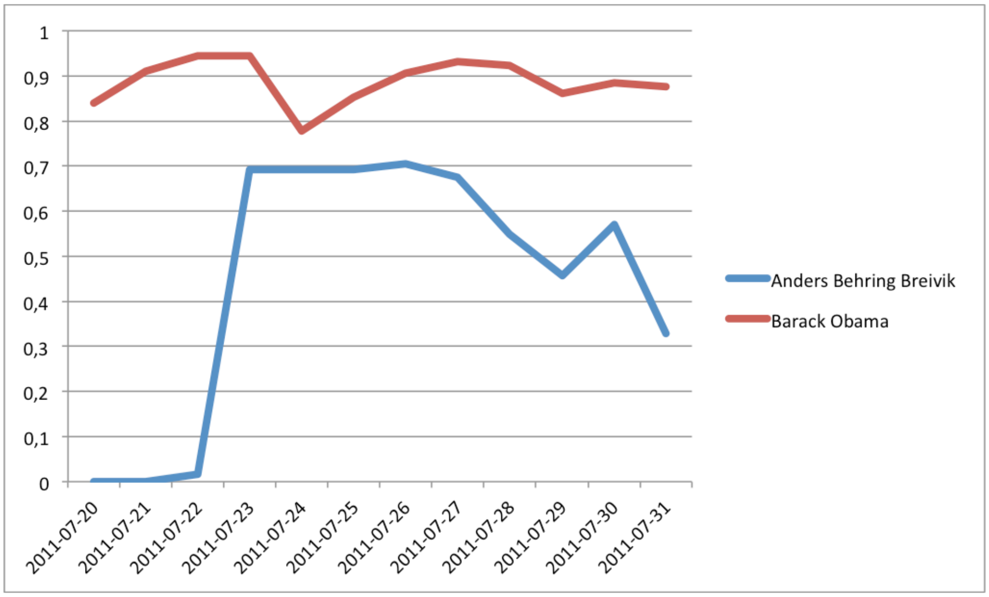
Metrics

Metrics are numeric attributes associated with either a canonical entity or event, or an entity or event instance. Metrics are used to order events and entities in terms of relevance (to ensure that the most important fact "float to the top"), and statistical algorithms based on metric values have been shown to have predictive power e.g. for stock prices.

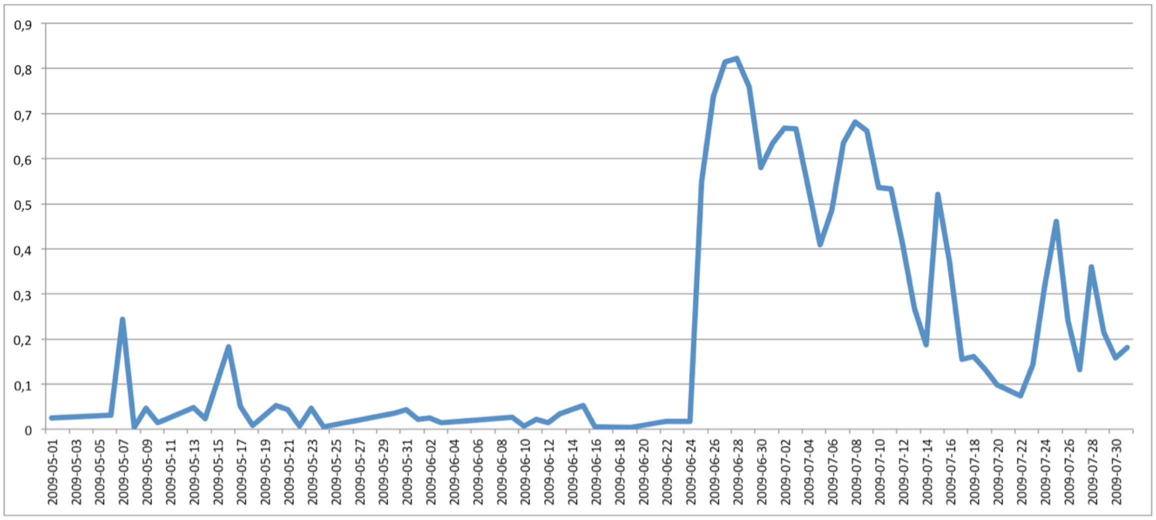
Momentum

To find relevant information in the sea of data produced by our system, we need a ranking metric. To this end, we have developed our momentum metric – a relevance measure between 0 and 1 for events and entities, which takes into account the flow of information about an entity/event, the credibility of the sources from which that information is obtained, the co-occurrence with other events and entities, and so on. Momentum is for example used to present results in most relevant order, and can also be used to find similarities between different events and entities. Momentum is a relative metric, and is computed within each entity or event category - hence it makes most sense to compare the momentum of two companies, but not the momentum of a person and a country. In general, entities or events with very high momentum are of course interesting and worthy of further investigation.

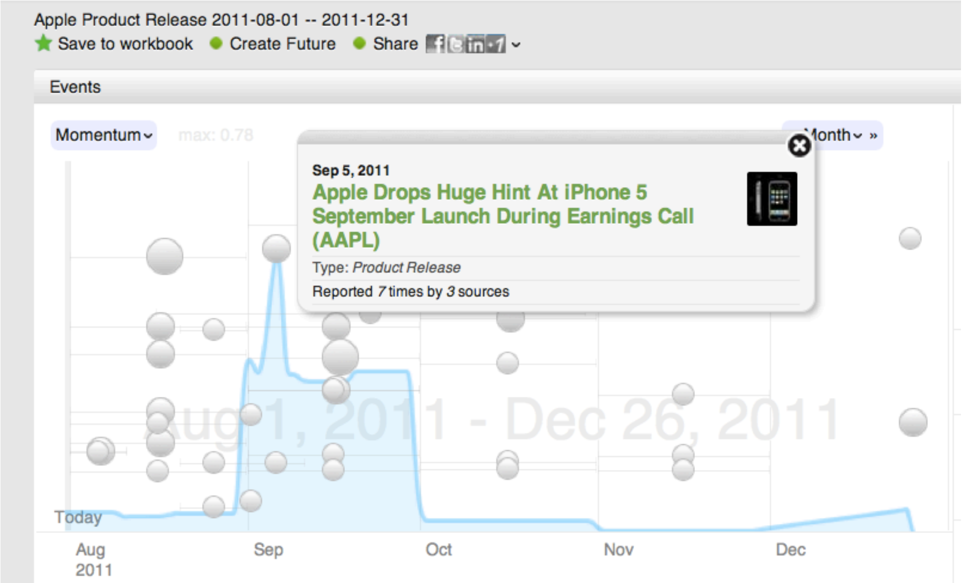
Momentum is our aggregate judgment of how interesting or important an entity or event is at a certain point in time – note that over time, the momentum measure of course changes, reflecting a dynamic world. As an example, here is a graph comparing the maximum momentum per day for Barack Obama and the Norwegian terrorist Anders Behring Breivik, for the last 12 days of July 2011:



This example illustrates the way momentum attempts to reflect both long and short term importance; Barack Obama is always one of the top persons in the news and thus always has high momentum, but the algorithm also reacts to total newcomers and highlights them with a high momentum value if they show enough presence in the media even during a very short period. Here's another example of how momentum can evolve - this is the momentum for Michael Jackson from the beginning of May 2009, through his untimely death on June 25th, and the month following that:



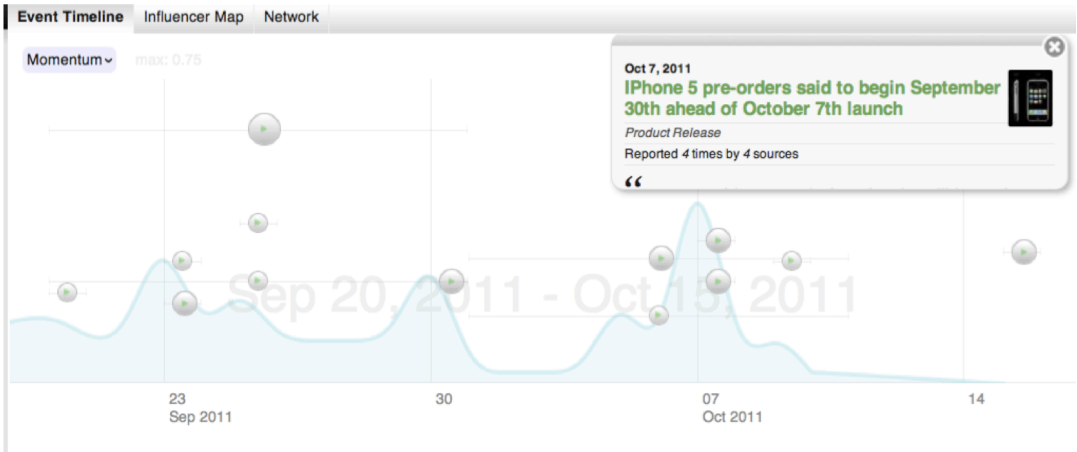
In these examples we can see how momentum evolves historically, but we can also view momentum for future events; as an example, this is the momentum curve for Apple product releases, as of August 1st 2011, and viewed in the Recorded Future web interface - note the big momentum for an iPhone 5 product release in September 2011:



Apple Product Release 2011-08-01 -- 2011-12-31

Sep 5, 2011
Apple Drops Huge Hint At iPhone 5 September Launch During Earnings Call (AAPL)
Type: *Product Release*
Reported *7* times by *3* sources

As of September 19th 2011, the iPhone 5 has not yet been released, and a modified curve can be seen in our system:



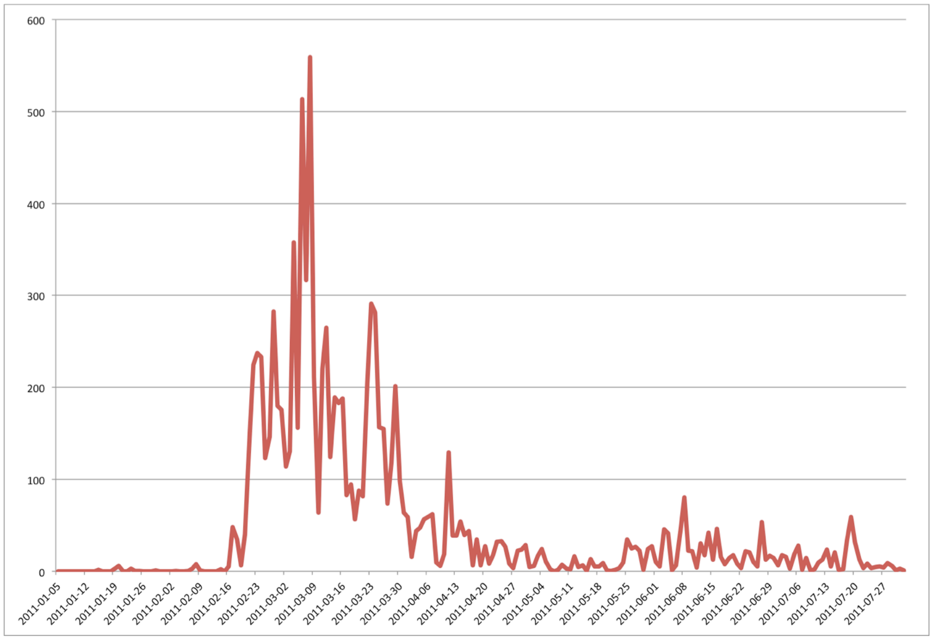
Oct 7, 2011
IPhone 5 pre-orders said to begin September 30th ahead of October 7th launch
Product Release
Reported *4* times by *4* sources
“

Sentiment

Another very useful metric is sentiment, which describes what attitude an author has towards his/her topic, and how strong that attitude is – the affective state of the author.

Currently, we compute positive and negative sentiment for all text fragments associated with entities and events (note that a sentence can contain both positive and negative sentiment).

As for all linguistic analysis, we might sometimes be mistaken in analyzing a single sentence, but on an aggregate level (e.g. the aggregated sentiment for an entity on a given day) we can ignore single individual incorrect results. As an example, here is the aggregated negative sentiment per day for “Muammar al-Gaddafi” during 2011, showing a burst of negative sentiment around the beginning of the Libyan crisis, and then dwindling off as media focus shifted elsewhere:



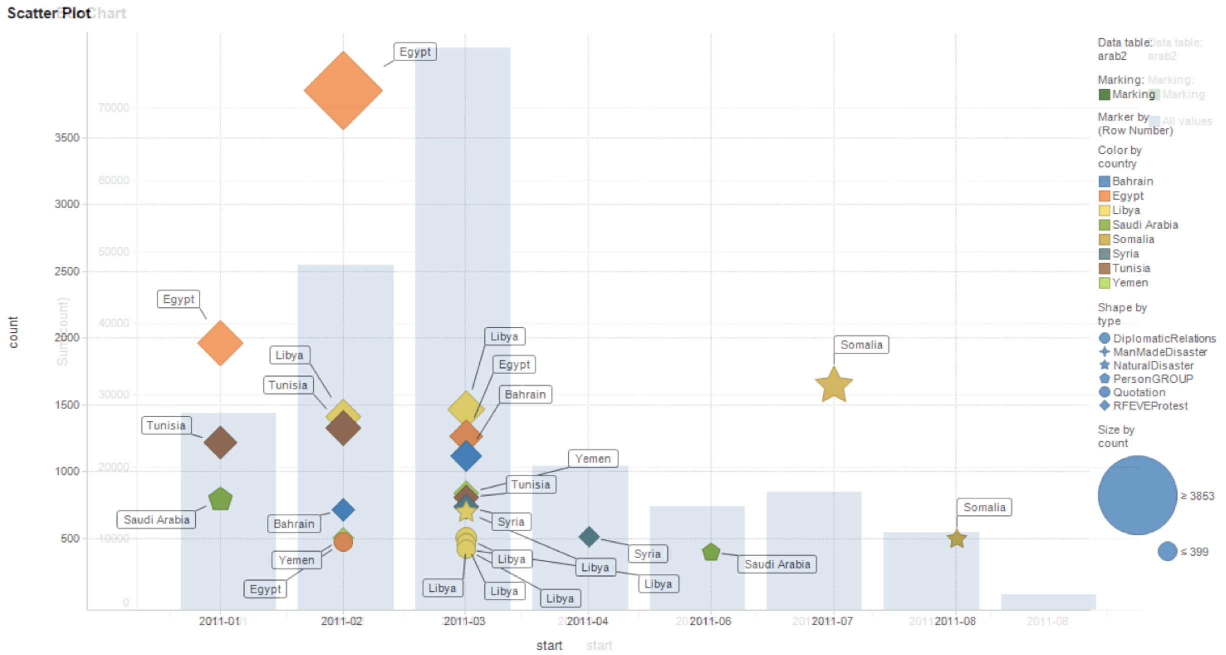
From Data to Analysis

The three dimensions of data - structure, time and metrics - described above comprise the base on which we can do analyses and create predictions about the future. The information can be used in several ways, e.g. by slicing out high momentum events for a certain company and time period, by aggregating weighted opinions about the likely timing of a future event, (using what we call algorithmic crowd sourcing), or by building statistical models to predict future happenings based on historical records of chains of events of similar kinds.

Several examples of how our system can be used for analysis - both in our web interface and using our API - can be found on our blogs:

http://www.analysisintelligence.com/
http://www.predictivesignals.com/
http://blog.recordedfuture.com

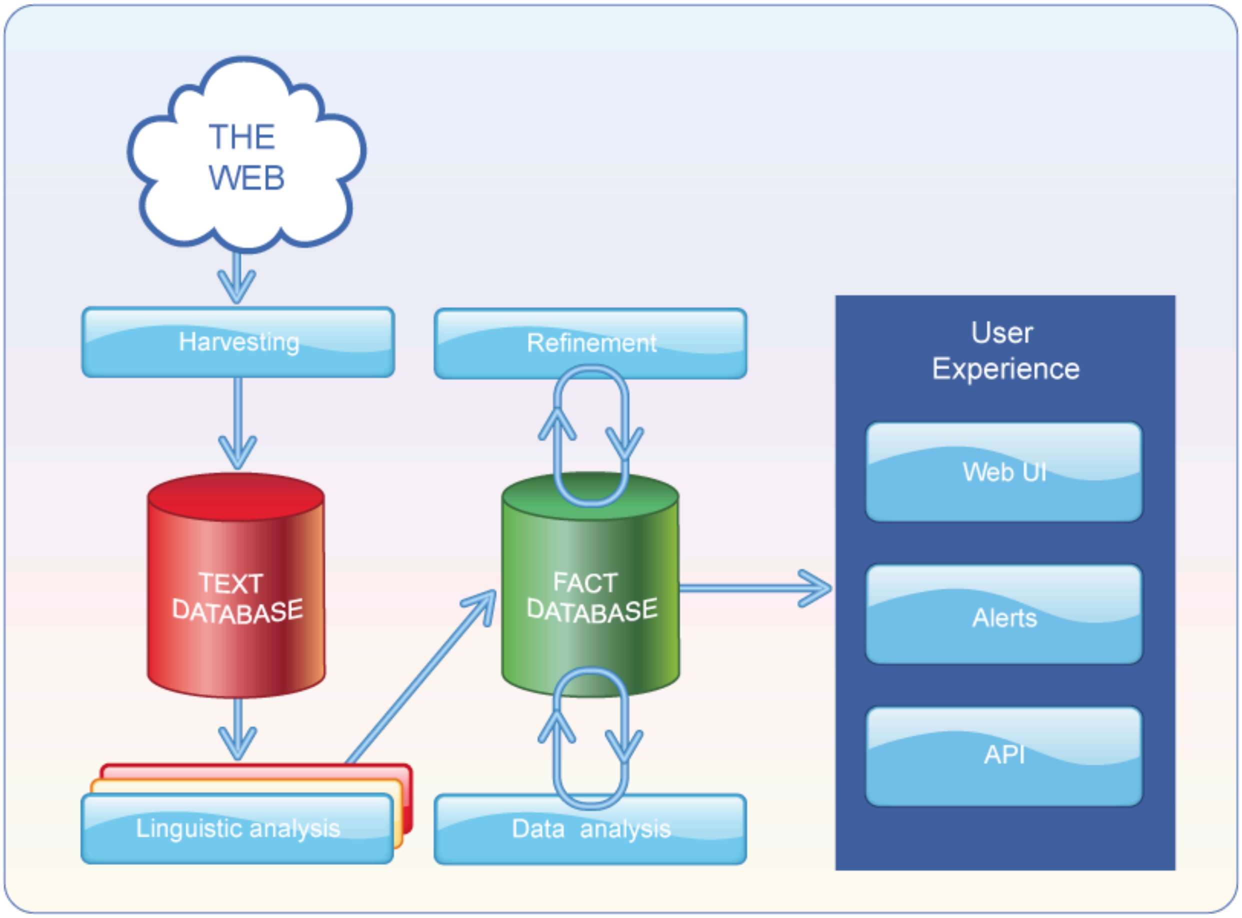
One example of how data can be extracted using our API and visualized in a 3rd party system (in this case, Spotfire) is the following timeline for “Arab Spring”:



Here, events for 22 Arab nations, from the beginning of 2011 and up till the beginning of September 2011, were extracted and clustered. The grey bar chart shows the total volume of events per month, and the icons are clusters of events, per country. The icon shape signifies the event type (‘diamond’ for Protest events, ‘star’ for Natural Disasters, ‘circle’ for Diplomatic Relations, ‘pentagon’ for Person related events, etc.). We can easily see how the events move from country to country, and also which events are most common.

System Overview

The Recorded Future system is centered around the fact database. Here, we store all canonical events and entities together with information about all their instances, and clusters. Intermediately, text harvested from the web is stored in another database, before being processed.



There are five major blocks of system components:

1. Harvesting – in which text documents are retrieved from various sources on the web and stored in the text database. We harvest everything from big media and government RSS feeds and down to blogs and selected twitter streams. Some more structured data sources, such as SEC filings and economic calendars, are also harvested.
2. Linguistic analysis – in which the retrieved texts are analyzed to detect event and entity instances, time and location, text sentiment etc. This is the step that takes us from the text domain to the data domain. This is also the only language dependent component of the system; as we add support for multiple languages new modules are introduced here. We use industry leading linguistics platforms for some of the underlying analysis, and combine them with our own analyses when necessary.
3. Refinement – in which data is analyzed to obtain more information; this includes calculating the momentum of entities, events, documents and sources, synonym detection, and ontology analysis.
4. Data analysis – in which different statistical and AI based models are applied to the data to detect anomalies in the data and to generate predictions about the future, based either on actual statements in the texts or other models for generalizing trends or hypothesizing from previous examples.

5. User experience – the different user interfaces to the system, including the web interface, our Futures alert mechanism, and the API for interfacing to other systems.

A Final Word

Recorded Future brings a paradigm shift to analytics, by focusing on the web as a data source, and on time as an essential aspect of the analyst's work. Sophisticated linguistic and statistical analyses combined with innovative user interfaces and a powerful API brings new opportunities to both human analysts and developers of analytics systems. We continuously develop all these aspects of our system to bring new tools into the analysts' hands - the future has only just begun!

If you can look into the seeds of time and say,
which grain will grow, and which will not,
speak then to me. (Macbeth, Act 1 Scene 3)