Analyzing and Integrating Social Media

AnHai Doan
University of Wisconsin-Madison and @WalmartLabs

I am interested in analyzing and integrating social media data at the semantic level, then providing such services on the cloud. This interest stems from my work at Wisconsin, Kosmix, and @WalmartLabs. Kosmix was a social media startup in the Bay Area. It was bought in 2011 by Walmart and converted into @WalmartLabs, a research and development lab that analyzes social and mobile data for e-commerce.

Semantic services on the cloud

By social media I mean data such as tweets, blogs, and Facebook updates. A lot of work has analyzed such data, but at the *keyword* level, to answer questions such as "how many tweets mention the word "Obama" today?". In contrast, I want to analyze and integrate such data at the *semantic* level, to answer questions such as "how many tweets mention President Obama today?". To answer this question, I would need to recognize that tweets that mention "Obama", "the pres", "BO", "the messiah", etc. all refer to the same person. In other words, I want to infer entities and relationships from the raw social media data, then leverage them to provide useful services.

Numerous examples of semantic analysis and integration of social media have been studied. Here's a small sample:

- **Information extraction and entity disambiguation**: Given a tweet such as "mel crashed his car. maserati is gone", recognize that "mel" is a person name and that "maserati" is a car name. Further, recognize that "mel" here refers to the person entity Mel Gibson, the Hollywood actor, and not some other Mel.
- **Event discovery**: Find interesting events in the Twittersphere. Global events include Japanese earthquake and the bin Laden assassination. Local events include a planned protest in a square in Syria and a book fair in Mountain View, California.
- **Event monitoring**: Once an event has been found, find all tweets related to that event and display them in a continuously rolling fashion, as they appear.
- **Statistics gathering**: How many tweets mentioned Mitt Romney in the past three hours? What is the overall sentiment of Florida voters with respect to Newt Gingrich in the past two days?

I am interested in developing such semantic services, and then deploying them on the cloud, for companies and end users. For example, an end user may be interested in monitoring all tweets related to an upcoming book fair in Mountain View. He or she can go to a Web site provided by us, define the book fair event, then subscribe to it. When we find any tweet related to this event, we will automatically send the tweet to the user.

Research directions

At Kosmix and @WalmartLabs we have studied many of the above problems. Our experience suggests the following research challenges.

Traditional integration challenges: We must build a giant knowledge base of entities and relationships, along the line of Wikipedia and Freebase, then use this knowledge base to analyze and integrate social media. For example, given the tweet "mel crashed his car", we can recognize that "mel" is a possible person name, because it appears as a person name in our knowledge base. Further, we can match "mel" into the Mel Gibson node in the knowledge base, thereby performing entity disambiguation.

While critical, developing this knowledge base raises numerous challenges. First, we must integrate data from numerous sources, such as IMDB, Wikipedia, Freebase, Musicbrainz, TripAdvisor, etc. into a coherent whole. What is a good end-to-end ETL methodology to integrate such data? How can we use big data techniques (such as Hadoop) in such integration, so that we can scale up to terabytes of data? Events in social media often unfold at real-time speed, in seconds or minutes. How can we refresh the ETL pipeline quickly, so that a change in a raw data source can be reflected in the knowledge base in seconds? Today there has been relatively little research on these challenges.

Social expansion challenges: Once the giant knowledge base has been built using "traditional" Web data (such as IMDB, Wikipedia, Musicbrainz), we must expand it with entities, relationships, and events emerging from the social media sphere. For example, we must continuously add Twitter users and Facebook users, as we discover them, into the knowledge base. We must discover interesting events in the Twittersphere, and add those to the knowledge base too.

If we view the knowledge base as our "understanding of the world" that can be used to help us analyze and integrate social data, then clearly this "understanding" must involve not just "old" entities, relationships, and events (that we find in the traditional Web data), but also "new" entities, relationships, and events that have just emerged in social media. Such social expansion of the knowledge base raises interesting challenges, such as how to match social media users (e.g., a Twitter user account) to person entities already existing in the knowledge base, and how to discover interesting events in the Twittersphere.

Social context challenges: Not only do we have to expand the knowledge base "socially", by adding "social" entities and relationships, but we also have to add social context to each node in the knowledge base. To see this, consider again the tweet "mel crashed his car". Given just this tweet, we can't really tell that "mel" here refers to Mel Gibson, because there is no keyword indicative of Mel Gibson in the tweet, such as "actor", "Oscar", "scandal", and "Hollywood". On the other hand, if we know that Mel Gibson has just crashed his car, and that in the past three hours, most tweets related to Mel Gibson mentioned words such as "crash", "car", and "maserati", then we can match "mel" to Mel Gibson with high confidence.

The above example suggests that for each node in the knowledge base (which can be visualized as a giant graph), we have to maintain a *social context*, a set of words that are most indicative of that node in the past few hours in the social media. The social context is the key that allows us to perform semantic analysis such as entity disambiguation. The challenge is how to construct these social contexts accurately, to maintain them on a very large scale, for hundreds of millions of nodes, and to ensure very low latency.

Crowdsourcing and human computing challenges: In building the knowledge base, expanding it with social elements, and adding social contexts, we will have to use not just algorithms, but humans as well, in a crowdsourcing fashion. How to crowdsource effectively is a major challenge.

Further, once we have deployed semantic services on the cloud, how to effectively engage the end users is also a major challenge. For example, when a user goes to our site to define the book event in Mountain View, what should we ask the user to do, what kind of information should he or she provide, so that we can effectively find tweets that are related to that event for the user?

Scaling challenges for big and fast data: Clearly we will have to deal with not just big data, but also fast data, such as high-speed streams of tweets, Facebook updates, Foursquare checkins, etc. MapReduce has proven to be an effective paradigm to write and execute big data programs. Can we develop a similar paradigm for fast data?

When we deploy semantic services on the cloud, we will have potentially thousands to millions of users taking advantage of the services (e.g., defining an event and monitoring the event). This will severely exacerbate the above big data and fast data challenges.