Twitris 2.0 : Semantically Empowered System for Understanding Perceptions From Social Data

Ashutosh Jadhav, Hemant Purohit, Pavan Kapanipathi, Pramod Ananthram, Ajith Ranabahu, Vinh Nguyen, Pablo N. Mendes, Alan Gary Smith, Michael Cooney, and Amit Sheth

Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) Center
Wright State University, Dayton, Ohio 45435
{ashutosh,hemant,pavan,pramod,ajith,vinh,pablo,alan,michael,amit}@knoesis.org

Abstract. We present Twitris 2.0 [1], a Semantic Web application that facilitates understanding of social perceptions by Semantics-based processing of massive amounts of event-centric data. Twitris 2.0 addresses challenges in large scale processing of social data, preserving spatio-temporal-thematic properties. Twitris 2.0 also covers context based semantic integration of multiple Web resources and expose semantically enriched social data to the public domain. Semantic Web technologies enable the system's integration and analysis abilities.

1 Introduction

With the wide adoption of Web 2.0 and social networking platforms, the vast amount of social data volunteered by users opens an exciting opportunity to extract social perceptions. In particular, emergence of microblogging platforms such as Twitter, friendfeed etc. have revolutionized how unfiltered, real-time information is disseminated and consumed by citizens. Citizens themselves are now empowered to act as sensors, reporting a variety of perceived observations. This phenomenon is different from the traditional centralized information dissemination and consumption environments where citizens primarily act as consumers of reported information from several authoritative sources. Twitter, has therefore emerged as the pre-eminent medium for sharing citizen-sensor observations, as was demonstrated in a variety of situations ranging from Mumbai terrorist attack to Iran elections [6].

While the decentralized information diffusion model offered by twitter has gained momentum and has created avenues for experiential data sharing, millions of observations, shared through tweets, create significant information overload. In many cases it becomes nearly impossible to make sense of the information around a topic of interest. This problem is further compounded by the fact that tweets increasingly integrate other social networking sites (flickr, twitpics) and general Web content(news, Wikipedia, blogs) through embedded links and metadata. Given this data deluge, analyzing the numerous social signals carried by tweets

[1] http://twitris.knoesis.org

and associated content to find out what is being said about an event (theme), where (spatial), when (temporal), how are key concerns (topics of discussion) changing over a period of time and whether there are regional differences in the opinions on a given topic, can be extremely challenging.

Furthermore a variety of spatially-sensitive facets such as culture, language and history as well as time sensitive events such as political debates and crisis/disaster monitoring and response, influence the local perceptions, complicating the analysis of Twitter data. For example in the Presidential elections in United States, citizen observations relayed from various location offered multiple, and often complementary viewpoints. What is more, these viewpoints evolved over time (before, during and after the elections) and with the occurrence of other events. Expectedly, a prohibitively large number of tweets expressed various perspectives on the elections, making informed decisions based on aggregation of tweets a challenging task.

In response to this growing data deluge, we have developed Twitris (currently Twitris 2.0) with the vision of performing semantics-empowered analysis of a broad variety of social media content. Specifically, Twitris aims to capture semantics (i.e., meaning and understanding) with spatial, temporal, thematic dimensions, user intentions and sentiments, networking behavior (user interactions patterns and features such as information diffusion and centrality) and other information present in social media. Semantic Web technologies enable its core integration, analysis and data/knowledge sharing abilities. Twitris 2.0, focuses only on content centric analysis , leveraging the relevant Semantic Web technologies, background knowledge, languages, tools where appropriate.

Twitris 2.0 is a Semantic Social Web approach to detect social signals by analyzing massive, event-centric data through
a. Analysis of casual text with spatio-temporal-thematic (STT) bias, to extract event descriptors.
b. Capturing semantics from three contexts(internal context, external context and mined internal context) associated with tweets(Section 2.2).
c. Use of deep semantics (using automatically created domain models) to understand the meaning of standard event descriptors.
d. Use of shallow semantics(semantically annotated entities) for knowledge discovery and representation.
e. Exposure of processed social data to the public domain, complying with semantic Web standards.
f. Semantic Integration of multiple external Web resources (news, articles, images and videos) utilizing the semantic similarity between contexts.

2 Twitris 2.0 Overview

Twitris 2.0 is developed as a multi-layered system where each component acts as part of a pipeline. The processing starts by extracting tweets. These tweets are then processed, exposed via the LOD cloud or explored via a rich user interface

(Figure 1). Each of the illustrated components are described in detail in the subsequent sections.

2.1 Tweet Extraction

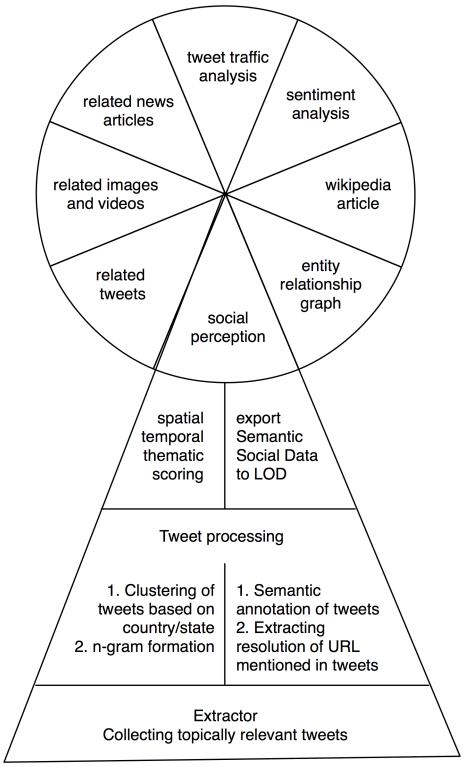

Fig. 1. Twitris 2.0 Functional Overview

Twitris 2.0 has implemented a near real-time extractor to fetch topically relevant tweets using Twitter Search API. The volume of messages on a popular topic in Twitter coupled with the short nature (140 characters), poses significant challenges for extraction and processing. Identifying topically relevant posts is a significant extraction challenge. Twitris 2.0 has a continuous topic identification and update strategy, that starts by searching for concepts in DBpedia that are pertinent to the events. For example, we identify DBpedia concepts that are relevant to healthcare and healthcare debate in the United States. This set of identified concepts form the semantic keyword cluster. While the use of semantic models enhances the precision of our extraction, users of Twitter often employ words and terms that are casual in nature and not found in semantic models such as DBpedia. We employ statistical techniques to identify additional seed keywords for the extraction. To do this, we extract tweets for a fixed time period (a day in the current implementation). The statistical keyword cluster is created using the keywords extracted from the tweets, along with hashtags. This set is enhanced using Google Insights for Search [2], a service that provides top searched and trending keywords across specific regions, categories, time frames and properties.

This Twitris 2.0 data is exposed in RDF format and published on LOD cloud as a part of SSD (Section 2.3).

[2] http://www.google.com/insights/search/

2.2 Tweet Processing

Finding Spatio-Temporal-Thematic(STT) Event Descriptors Casual text form of the tweet content necessitates the need to go beyond conventional text processing approaches [3, 1]. Moreover, presence of twitter conventions such as mentions (denoted by @), shortened URL resources, user names, hashtags etc requires us to preserve their semantics while identifying and processing them. Also, conversational practices such as retweeting and mentions tend to create a statistically significant bias in the corpus due to repetition of the text. To perform statistical computations such as TFIDF computation on a changing corpus, requires fundamental changes in the way these computations are defined and performed.

Given such challenging computation, Twitris 2.0 performs three step processing for finding N-gram summaries from tweets. First, it creates the Spatio-Temporal clusters of the tweets corpus surrounding an event, since every event is different and we want to preserve social perceptions that generated this data. TFIDF computation is performed to fetch the n-grams from this set. Second step involves the association of spatial, temporal and thematic bias to these n-grams by means of enhancing the weights, while preserving the contextual relevance of these event descriptors to the event. Finally, we create domain models automatically considering the event context and prominent event descriptors using Doozer [7]. These domain models are used to facilitate fine grained browsing of concepts and as an evidence to calculate the weight of extracted terms. We enhance the weights of descriptors that share a relationship with one or more terms from the semantic keyword cluster. Further details of the text processing algorithm are available in [4].

Tweet Traffic Analysis An interesting observation to make on an event is its popularity over a period of time. Twitris 2.0 keeps track of number of tweets collected per day for each event. For intuitive visualization of this information, a graph of tweet count is generated on a time line for each event.

Semantic Context Analysis Twitter enables sense making by capturing semantics from three contexts associated with a tweet: internal context, external context and mined internal context as described in the following section.
Internal Context: Context obtained by analyzing directly mentioned content in the tweet. These include
1. Images, videos and other Web content directly linked from the tweet. Tweets often include shortened URLs to refer to Web resources. Twitris 2.0 extracts all these URLs and resolves them if necessary. We perform following analysis on the URLs: **a**. Fetch the title of referenced article and hyperlink it to the URL; **b**. Extract entities mentioned in the title of external reference page using OpenCalais [3]; **c**. Collect images and videos links by performing semantic and syntactic analysis of URLs. Metadata for each URL is retrieved from the relevant

[3] http://www.opencalais.com/

API if available (such as the twitipic.com image API). The retrieved metadata (title, description) is used to check semantic similarity between the image and event context. Images are annotated using Dublin Core(DC) vocabulary [4] so they can be exposed in the form of RDF triples through a SPARQL endpoint, allowing us to query images based on events and enable data reuse. The same technique is used for collecting event related videos using the Youtube API.
2. Related tweets found to contain *event descriptors* present in the current tweet. These phrases, called *event descriptors*, are found after the spatio-temporal-thematic analysis.
3. Semantically annotated entities mentioned in the tweet using Named Entity Recognition (NER). Unstructured nature of tweet content is transformed to structured representation by semantically annotating the tweets with DBpedia entities [2]. Meaning Of A Tag (MOAT)[5] is used to model the structured data. Twitris 2.0 incorporates background knowledge from DBpedia by linking entities mentioned in the tweets to DBpedia, which provides meaningful relationships to other entities in the knowledge base.
External Context: Context obtained from external sources by semantically following the theme of the current tweet. External context facilitate understanding of tweet content in broader sense. These includes Google news, Wikipedia articles and other Web content that are not mentioned in the tweet but relevant to the theme the current tweet belongs to.
Mined Internal Context: Context obtained by mining the internal context.
1. Sentiment analysis comprises of finding the sentiment polarity of the tweet with respect to an event of interest using machine learning techniques. For each event related tweet, first all possible on-topic sentiment units of the tweet are extracted as its features. A sentiment lexicon learned from the domain-specific corpus is employed by the classifier to identify actual sentiment units and remove noise. Finally, the tweet is classified as objective, positive, neutral or negative by the classifier using a lexicon-based classification algorithm. As an initial demonstration and evaluation of this method, we have run this feature on the events from movie domain, and the result shows our approach outperforms several baseline methods significantly.
2. The entity-relationship graph is created using the semantically annotated DBpedia entities in the tweets, which can be used to facilitate fine grained browsing of concepts. This entity-relationship graph supports knowledge representation and discovery. The graph is constructed considering relationships upto 2 hops and displayed using the RelFinder[6] user interface.

2.3 Semantic Social Data(SSD)

Demonstration version of Twitris 2.0 involves over 16 millions tweets with metadata (ID, content, author information, event name, published data, latitude and

[4] http://dublincore.org/
[5] http://moat-project.org/
[6] http://relfinder.dbpedia.org/

longitude) pertaining to different events. These tweets are semantically enriched with annotations corresponding to the three types of context. All the Twitris 2.0 data is exposed in RDF format and published on LOD as a part of SSD. Based on tweet location information, SSD dataset is connected with GeoNames [7] dataset and we are planning to integrate FOAF[8] dataset with SSD based on the author information.

The STT analysis performed on the tweets can be reused by the community for a variety of analysis tasks. One of the major contribution of Twitris 2.0 is to publish social perceptions with the context as a dynamic dataset on LOD. At present LOD can answer historical data related questions such as *where was Barack Obama born?*. In future, our dataset will enable answering time sensitive questions like *where is Barack Obama now?*. Consider a case where people are tweeting about Obamas speech at Chicago. Twitris 2.0, on incorporating work in Twarql[9] [2] , can capture Obama and Chicago as DBpedia entities and based on the timestamp associated with tweets, the dataset can answer such questions.

3 User Interface and Visualization

The primary objective of the Twitris 2.0 user interface is to integrate the results of the data analysis (extracted descriptors and surrounding discussions) with emerging visualization paradigms to facilitate sensemaking. The current user interface facilitates effective browsing of the when, where, and what slices of social perceptions behind an event and includes UI components to illustrate the theme, time and space. The UI also includes widgets for media (Related tweets, News and Referenced articles, Wikipedia articles), entity-relationship graph, images, videos, traffic volume and sentiments. These widgets are dynamically populated depending on the users selections.

4 Technical Choices

Twitris 2.0 back-end is developed using PHP and Java. PHP is used where lightweight functionality is required and Java is used primarily for the text processing and RDF handling components due to the availability of robust libraries. Virtuoso[10] is used as the SPARQL end point after a performance evaluation with other alternative open source solutions. MySQL is the primary data storage. The front-end is built with JavaScript, primarily using the JQuery library for the convenience it provides over common operations. We have used Web services provided by Twitter, Yahoo! BOSS, Google News, twitpic, Youtube, DBpedia and Wikipedia.

[7] http://www.geonames.org/
[8] http://www.foaf-project.org/
[9] http://twarql.sf.net
[10] http://virtuoso.openlinksw.com/

5 Twitris 2.0 Statistics

Important statistics in the current data set is presented in Table 1.

Total number of extracted tweets	17.5 million
Processed Tweets	8 million
Cached unique location geocodes	595,301
Cached author locations	2.4 million
Extracted event descriptors	3.7 million
Extracted DBpedia entities	1.3 million
Extracted external URLs	649,165

Table 1. Statistics of the collected data

6 Conclusion and On-going Work

Twitris 2.0 is a powerful system for understanding social perceptions, starting with microblogs, spanning through news, blogs and other Web content. Semantic Web technologies are essential in processing the deluge of social data, preserving the spatio temporal thematic bias and provide the ability to realize and integrate multi-dimensional social perceptions while using as well as contributing to open Web of Data.

The system is currently being used for a number of People-Content-Network study experiments and being extended to integrate with SMS and other Web data used by a number of widely deployed open source projects. These include applications used by non governmental organizations (NGO) in developing countries for crisis management (in particular, Ushahidi.org, eMoksha.org and Kiirti.org). Twitris 2.0 is being extended with Twarql technology for limited real-time support and is being adapted for a cloud platform for much higher scalability. Some of these on-going efforts will also be demonstrated at ISWC 2010.

References

1. Gruhl, D., Nagarajan, M., Pieper, J., Robson, C., Sheth, A.: Context and domain knowledge enhanced entity spotting in informal text. The Semantic Web-ISWC 2009 pp. 260–276 (2009)
2. Mendes, P., Passant, A., Kapanipathi, P., Sheth, A.: Linked Open Social Signals. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2010. WI-IAT'10. (2010)
3. Nagarajan, M., Baid, K., Sheth, A., Wang, S.: Monetizing User Activity on Social Networks-Challenges and Experiences. In: Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on. vol. 1, pp. 92–99. IEEE (2009)
4. Nagarajan, M., Gomadam, K., Sheth, A., Ranabahu, A., Mutharaju, R., Jadhav, A.: Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. Web Information Systems Engineering-WISE 2009 pp. 539–553 (2009)
5. Nagarajan, M., Purohit, H., Sheth, A.: A Qualitative Examination of Topical Tweet and Retweet Practices (2010), http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1484

6. Sheth, A.: Citizen sensing, social signals, and enriching human experience. Internet Computing, IEEE 13(4), 87–92 (2009)
7. Thomas, C., Mehra, P., Brooks, R., Sheth, A.: Growing fields of interest-using an expand and reduce strategy for domain model extraction. In: Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on. vol. 1, pp. 496–502. IEEE (2009)

Appendix

Coverage of minimal requirements

1. Twitris 2.0 clearly provides a practical value to end users by letting them analyze the social data for a particular event and understand the regional perception of citizens. The citizen-perception is often different from the reports from traditional news media which lack the regional bias. It is also impossible to make sense of the raw social data due to the large volume, usually in the range of tens of millions per day.
2. The primary information sources used are Twitter, Google (Insights for Search, Geocode) and DBpedia. These data sources provide completely different data formats and semantics.
3. The data sets used are in the range of tens of millions.
4. The core theme of Twitris 2.0 is to make sense of the meaning of the social media, currently with Twitter as the focus. Tweets are conversational, regionally and temporally biased and opinionated. This data needs to be semantically processed (as discussed in previous sections) to derive a useful understanding.
5. Twitris 2.0 uses semantic technologies to represent the processed tweets, that are subsequently exposed in the LOD cloud.
6. Without semantics it would be near impossible to find a coherent linking of themes and tweets. The statistical processes alone are not sufficient (as demonstrated in Twitris 1.0) to find meaningful relationships in the tweets.

Coverage of additional desirable features

1. Twitris 2.0 provides a dynamic (Java script driven) Web user interface.
2. Twitris 2.0 functionality is clearly a value addition to traditional information retrieval.
3. Twitris 2.0 is the first line of work where semantic technologies have been incorporated with the statistical analysis techniques for social data analysis.
4. There are potential commercial applications in Twitris 2.0, primarily as either alternative or complementary source of information to traditional news media. When traditional media is blocked (as was the case during Iran elections), social media would be the primary source of information and highlights the importance of a system like Twitris 2.0. Combined with Twarql capabilities, Twitris 2.0 can significant enhance real-time data search that a number of companies are developing.
5. Twitris 2.0 UI incorporates images and other Web media where applicable.