A Tweet Consumer's Look At Twitter Through Linked Data Goggles Via Google Analytics

Thomas Steiner and Arnaud Brousseau

Google Germany GmbH, ABC-Str. 19, 20354 Hamburg, Germany
{tomac,arnaudb}@google.com

Abstract. Twitter Trends[1] allows for a global or local view on "what's happening in my world right now" from a tweet producers' point of view. In this paper, we discuss the possibility to complete the functionality provided by Twitter Trends by having a closer look at the other side: the tweet consumers' – i.e., readers' – point of view. While Twitter Trends works by analyzing the frequency of terms and their velocity of appearance in tweets being written, our approach is based on the popularity of extracted named entities (in the sense of Linked Data) in tweets being read. Our experimentation architecture takes advantage of the possibility to use a client-side browser extension to harvest and dissect tweets from users' timelines, i.e., tweets supposed to be read. Named entities are extracted via several third-party Natural Language Processing (NLP) Web services in parallel, and are then reported to Google Analytics, which is used to store, analyze, and compute trends by pivoting Analytics data, e.g., users' geographic location, with the recorded named entities.

1 Introduction

1.1 Twitter Trends

Twitter Trends was introduced by Twitter during the summer of 2008. This service was implemented to reflect "what's happening in my world right now" on the micro-blogging platform. To compute trends, Twitter uses an algorithm which analyzes the words and hashtags in tweets. Although the concrete implementation details are kept secret, it is known[2] that the algorithm considers three main characteristics of a word/hashtag to determine if it is part of a "trend":

[1] http://blog.twitter.com/2008/09/twitter-trends-tip.html

[2] http://blog.twitter.com/2010/12/to-trend-or-not-to-trend.html

– the quantity, i.e, the absolute number of appearances of a given word/hashtag among all users' tweets.
– the velocity, i.e, the frequency of appearance of that word/hashtag. The higher the frequency, the more popular the word/hashtag.
– the newness, e.g, the (at the time of writing brand new) hashtag #ipad2 will be given priority over an old and generic tag like #awesome to be featured as "trend". The freshness of trends is also assured by the fact that the frequency is computed over a recent set of tweets, although we have no further details of the exact computation period – Twitter keeps this information secret.

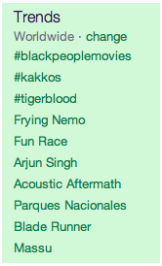

Fig. 1. Screenshot of Twitter Trends as of Friday, March 4, 2011, 4pm CET.

1.2 Google Chrome Extensions

Google Chrome extensions[3] are small software programs that can be installed to enrich the browsing experience with the Google Chrome browser. They are written using a combination of standard Web technologies, such as HTML, JavaScript, and CSS. Chrome extensions bundle all their resources into a single file that gets usually (but not necessarily) distributed through the Chrome Web Store. There are several types of extensions, for this paper we focus on extensions based on so-called content scripts. Content scripts are JavaScript programs that run in the context of Web pages, similar to the Firefox Greasemonkey extension[4]. By using the standard Document Object Model (DOM), they can read or modify details of the Web pages a user visits. Examples of such modifications are, e.g., changing hyperlinks to remove potential @target="_blank" attributes, or increasing the font size.

[3] Google Chrome Extensions: `http://code.google.com/chrome/extensions/index.html`. Text partly adapted from the description to be found there.
[4] Firefox Greasemonkey extension: `http://www.greasespot.net/`

1.3 Google Analytics

2 Twitter Swarm NLP Extension

With our Twitter Swarm NLP extension[5], we inject JavaScript code via a content script into the Twitter.com homepage. The extension first checks if the user is logged in, and if so, retrieves the tweets of the logged-in user's timeline one-by-one, and performs NLP analysis via a remote NLP Web service on each of the tweets. The extracted entities are then displayed on the righthand-pane of the Twitter.com homepage, and sent to Google Analytics for further processing.

2.1 Twitter Swarm NLP Web Service

We have created a wrapper NLP Web service that merges results from existing third-party NLP Web services, namely from OpenCalais[6], Zemanta[7], AlchemyAPI[8], and DBpedia Spotlight[9].

2.2 Dealing With Extracted Entities On the Client Side

2.3 Dealing With Extracted Entities On the Google Analytics Side

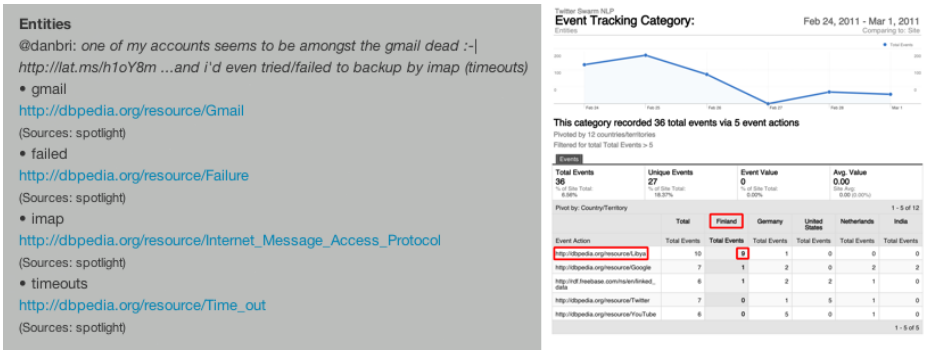

Fig. 2. Left: Screenshot of the extracted entites of a particular tweet as displayed by the Twitter Swarm NLP Extension. Right: Entities pivoted by country. The entity represented by the URL `http://dbpedia.org/resource/Lybia` appeared in 9 tweets on timelines of users located in Finland (red borders in the screenshot).

[5] `https://chrome.google.com/webstore/detail/dpbphenfafkflfmdlanimlemacankjol`
[6] `http://www.opencalais.com/`
[7] `http://www.opencalais.com/`
[8] `http://www.alchemyapi.com/`
[9] `http://dbpedia.org/spotlight`

3 Related Work

3.1 Linked Open Social Signals (TWARQL)

Previous work of Mendes et al. [1] has shown a possible implementation of real-time information both pushed and pulled from Twitter. TWARQL[10] is based on a distributed architectures which features:

- a client-side application which typically a Javacript-enabled web browser
- a "Social Sensor Server" to receive tweets and filter them according to the user's request. It is worth noting here that TWARQL filtering is based on web-semantic technologies: SPARQL, hash-tag resolution through glossaries and LOD cloud are used to extract the highest amount of information possible from the Twitter Streaming API.
- a number of distributed PuSH hubs which update clients as information flows (pushed-information model)
- another server – "Semantic Publisher" – registers user's interest and updates the hubs. The updated information is eventually displayed on the user's screen.

3.2 Twopular

Twopular[11] is an experiment by Martin Dudek with the objective of analyzing current Twitter trends. Therefore Twopular uses the OpenCalais Web service in a five minute interval with the most recent tweets behind the at the particular moment current Twitter trends in order to determine tags for these trends. By having tags for Twitter trends, another way of searching for trends – and more importantly the possibility to interrelate trends based on tag similarity – gets enabled. The author sees the feature more like a "linguistic experiment"[12], however states that the first results seem to be promising. In our tests of the service we could affirm the author's self-assessment, e.g., the Twitter trend (at the time of writing, March 7, 2011, 2PM CET) "Prince Andrew" was mapped to the OpenCalais tag "Prince Andrew, Duke of York", and related tags were, among others, "Sex offender registration", and "X-Offender", where the story behind the trend was that people were tweeting about Prince Andrew of England, whose close friend was found to be a pedophile.

4 Conclusion

Contributions: time filters (via Analytics), geographical pivoting (via Analytics)

As seen in the Related Work section, semantic analysis of a (real-time) Twitter stream is not new and has been successfully exploited to analyse tweets produced by the Twitter community. What we propose here is an insight into tweets consumers' interests to provide a more accurate view of Twitter trends.

[10] http://wiki.knoesis.org/index.php/Twarql

[11] http://twopular.com/

[12] http://twopular.com/blog/?p=308

Fig. 3. Screenshot of the Twopular trends page.

References

1. P. N. Mendes, A. Passant, P. Kapanipathi, and A. P. Sheth. Linked open social signals. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1:224–231, 2010.