Real-time #SemanticWeb in <= 140 chars

Joshua Shinavier
josh@fortytwo.net

RPI Tetherless World Constellation, Troy, NY USA
http://tw.rpi.edu/

Abstract. Instant publication, coupled with the prospect of widespread geotagging in microblogs presents new opportunities for real-time and location-based services. Much as these services are changing the nature of search on the World Wide Web, so the Semantic Web is sure to be both challenged and enhanced by real-time content. This paper introduces a semantic data aggregator which brings together a collection of compact formats for structured microblog content with Semantic Web vocabularies and best practices in order to augment the Semantic Web with real-time, user-driven data. Emerging formats, modeling issues, publication and data ownership of microblogging content, and basic techniques for real-time, real-place semantic search are discussed.

1 Introduction

Compared to the World Wide Web, the Semantic Web is lacking in user-driven content. Recent analyses [11][17] of the Linked Data cloud indicate that it does not exhibit the power-law distributions or strong connectivity typical of naturally-evolving networks.

Meanwhile, Web 2.0 services channel large amounts of potentially valuable user-driven data every day. Semantic wikis and the Microformats[1] community aim to bridge this gap by enabling users to add small amounts of semantic data to their content, while most of the work on *semantic microblogging* thus far has focused on representing users, microblogs and microblog posts in the Semantic Web: essentially, on doing for microblogs what SIOC [5] has done for blogs. This paper focuses on the complementary approach of harvesting semantic data *embedded in* the content of microblog posts, or doing for microblogs what microformats do for Web pages.

2 Nanoformats for the Semantic Web

A number of compact formats, variously called *nanoformats*[2], *picoformats*,[3] or *microsyntax*,[4] have been proposed to allow users to express structured content

[1] http://microformats.org/
[2] http://microformats.org/wiki/microblogging-nanoformats
[3] http://microformats.org/wiki/picoformats
[4] http://www.microsyntax.org/

or issue service-specific commands in microblog posts. Examples in widespread use include @usernames for addressing or mentioning a particular user, and #hashtags for generic concepts. So-called *triple tags* even allow the expression of something like an RDF triple. These formats are subject to a tradeoff between simplicity and expressivity which heavily impacts community uptake.

Twitter Data [10] is an open proposal for embedding structured data in Twitter messages. According to its FAQ, "the purpose of Twitter Data is to enable community-driven efforts to arrive at conventions for common pieces of data that are embeddable in Twitter by formal means". To kick-start this process, Twitter Data introduces a concrete syntax based on key/value pairs. For instance,

I love the #twitterdata proposal! $vote +1

RDF-like triples are possible using explicit subjects:

@romeo $foaf>loves @juliet

MicroTurtle [12] is a particularly compact serialization format for RDF, suitable for embedding general-purpose RDF data in microblog posts. It makes use of hard-coded CURIE [2] prefixes as well as keywords for terms in common vocabularies such as FOAF [7], Dublin Core [4], and OpenVocab.[5] For example:

Wow! Great band! #mttl #music <#me> ♥ [→<http://theholdsteady.com/> #altrock] .

This expresses, in a named graph tagged "music", that the author of the post likes a band, tagged "altrock", with the given homepage.

smesher [6] is a semantic microblogging platform which collects structured data from microblog posts in a local RDF store. Its syntax includes key/value pairs similar to Twitter Data's which are readily translated into RDF statements:

RT @sue: I can #offer a #ride to the #mbc09 #from=Berlin #to=Hamburg

Smesher users can query their data using SPARQL and filter it to create customized data streams.

TwitLogic currently supports a near-natural-language format which is intended to be particularly memorable and unobtrusive. Structured content is expressed by annotating a user name, hashtag or URL with a parenthetical "afterthought" resembling a relative clause. For example:

[5] `http://open.vocab.org/terms/`
[6] `http://smesher.org/`

Great convo with @lidingpku (creator of #swoogle) about ranking algos.

We make the assumption that hashtags such as #swoogle are semantically stronger than "ordinary" tags, in that microblog users who really want to refer their readers to a specific concept tend to avoid ambiguous tags. Ideally, the syntax should be natural enough so as not to distract the reader, yet contrived enough to minimize false positives:

#sioclog (see http://bit.ly/2uAWo2) makes Linked Data from IRC logs.

The following produces a minimal review of a movie in terms of the RDF Review Vocabulary:[7]

Who would have guessed such a funny movie as #Zombieland (3/4) could be made around zombies?

TwitLogic will eventually take advantage of some of the other formats described above, as well as pre-existing conventions such as "tag++". The best approach to the chicken-and-egg problem of semantic nanoformats may be to promote and build tools to support a variety of formats, see what "sticks", and then take steps to keep up with any community-driven conventions which may arise.

3 A user-driven Semantic Web knowledge base

There are several kinds of structured content which can be gathered from a microblogging service such as Twitter:

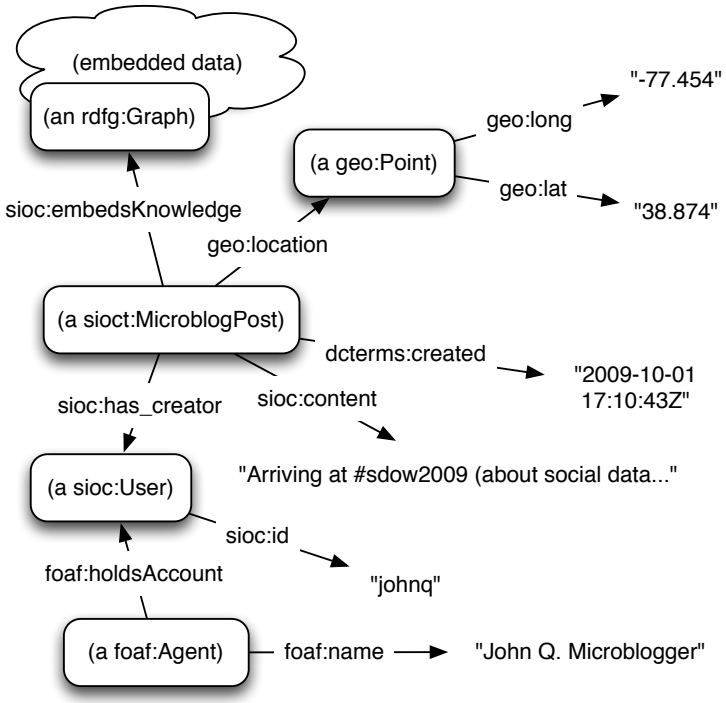
1. authoritative information about microblogging *accounts* and the *people* who hold them. SemanticTweet[8] is an example of a service which publishes the social network information provided by Twitter on the Semantic Web.
2. authoritative information about microblog *feeds* and individual *posts*. The SMOB semantic microblogging system [15], for one, represents microblog content at this level.
3. user-created information *embedded* in the text of a microblog post. Gathering such user-driven "statements about the world" and using them to populate the Semantic Web is the main goal of TwitLogic. People, accounts, and microblog posts are included in the knowledge base only as contextual metadata to enhance information discovery and provide author attribution for the embedded data.

[7] `http://vocab.org/review/terms`
[8] `http://semantictweet.com/`

3.1 Representing microblog content in RDF

The schema used for TwitLogic's knowledge base draws upon a recent discussion[9] on the Semantic Web mailing list about RDF vocabularies for modeling microblogging resources. In particular, it makes use of a collection of terms from the FOAF, SIOC, Dublin Core, Named Graphs [8] and Basic Geo [6] vocabularies. The sioc:embedsKnowledge property, which has been proposed in

Fig. 1. Embedded data and its metadata



connection with UfoWiki [16], serves to associate a microblog post with any structured data that has been extracted from it, in the form of a named graph containing the extracted RDF statements. This link not only provides source metadata for those statements, but also connects them with a *timestamp* and, potentially, a *placestamp* which are useful in searching and filtering. The use of geo:location as depicted is questionable, although convenient.

[9] `http://lists.w3.org/Archives/Public/semantic-web/2009Sep/0174.html`

3.2 Publishing the knowledge base as Linked Data

From the moment it is added to the TwitLogic knowledge base, the embedded data and contextual metadata of a microblog post are made available in accordance with best practices for publishing Linked Data on the Web [3]. Also available are a voiD [1] description of the data set as a whole, as well as owl:sameAs links into related data sets (currently, SemanticTweet's). As named graphs are an essential component of the knowledge base, TriX and TriG serializations of the data are provided alongside the more common RDF formats.

3.3 Data ownership

According to Twitter's terms of service,[10] "you own your own content", although that content may be freely copied, modified, and redistributed by Twitter and its partners. The data model described above supports authors' rights by providing attribution metadata for all user-generated content: the text of a microblog post is always associated with its author, and the RDF statements from embedded content in a post are always contained in a named graph which is associated with that post. Although TwitLogic does not rdfize every microblog post that passes through the system, it does maintain an RDF description of every tweet from which it has extracted content, so that attribution metadata is guaranteed, independently of the availability of the Linking Open Data[11] data sets into which TwitLogic links. In future, derived data such as search results will preserve the most relevant metadata through formal justification of results.

Given its diverse authorship, the TwitLogic knowledge base as a whole is published under the Open Data Commons[14] Public Domain Dedication and License (PDDL).

4 Real-time, real-place semantic search

Apart from collecting and publishing user-generated semantic data, we would also like to be able to search and reason on the data. SPARQL query answering, as well as a wide variety of Semantic Web reasoning techniques, are possible. In this environment, however, every user-generated statement is associated with a time-stamped and potentially[12] place-stamped microblog entry. We would like to take advantage of this metadata to score search results based on nearness in time and location to the context of the query.

Although the type of search may vary, the basic technique of TwitLogic is to keep a record of the *influence* of a named graph on intermediate results during query execution and to combine this with a measure of the *significance* of the graph, in terms of space and time, to produce a score for each query result.

[10] `http://twitter.com/tos`

[11] `http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/`

[12] `http://blog.twitter.com/2009/08/location-location-location.html`

A planned feature of TwitLogic is to use the PML Justification ontology [9] to annotate search results with the graphs most relevant to the derivation of each result. Overall significance of a graph in TwitLogic is computed as the product of its significance in time and space.

4.1 Time-based significance

TwitLogic's use of time metadata is based on an inverse relationship between the significance of a graph and the positive difference between the current (or reference) time and the timestamp of the graph. Specifically, the time-based significance, S_{time}, of a statement should have a value of 1 when there is no difference in time, and should approach a designated baseline value, b_{time}, as the difference goes to infinity. An exponential decay function has this behavior:

$$S_{time}(t) = b_{time} + (1 - b_{time}) \cdot 2^{-\frac{t}{t_h}}, \tag{1}$$

where t_h is the amount of time it takes for the significance of a statement to drop to half of its original value, disregarding the influence of b_{time}.

The resulting preference for the *most recently acquired* information is fitting for microblogging environments, in which it's generally not possible to "take back" statements which have been made in the past:[13] instead, one simply makes new statements.

In the current application, b_{time} is given a value greater than zero, as it's not desirable for old statements to "disappear" entirely. For example, a statement of a person's gender is usually as valid years from now as it is today. However, as new information becomes available, it will tend to supplant older information. This "freshness" is particularly advantageous for properties such as foaf:based_near, whose value is frequently subject to change.

4.2 Location-based significance

Our requirements for location-based significance are the same as those for time-based significance, except that geo-distance varies over a finite interval, whereas time-distance varies over an infinite one. S_{loc} should have a value of 1 at the current (or reference) location, and a baseline value of b_{loc} at the maximum distance d_{max}:

$$S_{loc}(d) = 1 + (b_{loc} - 1) \cdot \frac{d}{d_{max}} \tag{2}$$

where d is the great-circle distance from the reference location. Note that only distance, as opposed to actual position on the global map, is considered here.

[13] Some microblogging services, including Twitter, do allow users to delete their posts. Nonetheless, once a post is "out there", it is potentially out there for good, both in computer systems and human memory.

4.3 Closing the world of time and space

It is necessary to impose a baseline significance of b_{time} or b_{loc} on graphs with no time or place metadata. Equivalently, we can think of such graphs as occupying a time and place long, long ago or at the other end of the world. All of the data from the *rest* of the Semantic Web which we might like to search and reason on, resides here.

5 Implementation

An open-source implementation[14] of the ideas set forth in this paper is available online. The TwitLogic home page and demo can be found at:

`http://twitlogic.fortytwo.net/`

Noteworthy features of the implementation include:

1. the use of the Twitter streaming API[15] to acquire microblog posts in near-real-time
2. the use AllegroGraph[16] 3.0 for fast geolocation- and time-based queries as well as free-text search
3. efficient spreading activation algorithms in the Allegro Common Lisp environment
4. the use of XMPP Publish-Subscribe [13], to stream TriX-formatted data from the Java-based Twitter stream listener to the Lisp-based query environment

Other tools and services used include the Sesame v2.2 RDF toolkit, the ANTLR v3 parser generator, the Smack v3.1 and cl-xmpp v0.7 XMPP toolkits, OAuth Signpost v1.1 and the bit.ly URL-shortening API.

The current demo includes the Linked Data interface to the knowledge base, as well as a RESTful query API. Twitter users can access the API by tweeting at the @twit_logic user, which responds to correctly-formatted queries with the URL of a results page. Query syntax is a work in progress, as is the PML-based results format. As tweet-based queries are answered in real-time and will soon contain a placestamp for those users who have opted into Twitter's geolocation functionality, no extra syntax is required to make TwitLogic queries time- and location-sensitive.

Future work is likely to include a trust-based mechanism for selecting the subset of users that TwitLogic "listens to" through Twitter's rate-limited API, as well as a trust-based significance factor for query evaluation.

---

[14] `http://github.com/joshsh/twitlogic`
[15] `http://apiwiki.twitter.com/Streaming-API-Documentation`
[16] `http://www.franz.com/agraph/allegrograph/`

6 Acknowledgements

This project has been supported by Franz Inc. as well as RPI's Tetherless World Constellation. Special thanks go to Jans Aasman, James A. Hendler, Deborah L. McGuinness and Steve Haflich for their contributions to the concept and implementation of TwitLogic, and to Li Ding, Jie Bao, Marko A. Rodriguez, Alvaro Graves, Gregory Todd Williams, Jesse Weaver and Xixi Luo for their helpful comments and feedback.

References

1. Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao, *Describing linked datasets: On the design and usage of voiD, the "Vocabulary of Interlinked Datasets"*, 2nd International Workshop on Linked Data on the Web (Madrid, Spain), April 2009.
2. Mark Birbeck and Shane McCarron, *CURIE syntax 1.0*, http://www.w3.org/TR/curie/, January 2009.
3. Christian Bizer, Richard Cyganiak, and Tom Heath, *How to publish Linked Data on the Web*, http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/.
4. DMCI Usage Board, *DCMI metadata terms*, http://dublincore.org/documents/dcmi-terms/, January 2008.
5. Uldis Bojars and John Breslin, *SIOC core ontology specification*, http://rdfs.org/sioc/spec/, January 2009.
6. Dan Brickley, *Basic Geo (WGS84 lat/long) vocabulary*, http://www.w3.org/2003/01/geo/.
7. Dan Brickley and Libby Miller, *FOAF vocabulary specification*, http://xmlns.com/foaf/spec/, November 2007.
8. Jeremy Carroll, *Named graphs*, http://www.w3.org/2004/03/trix/.
9. Paulo Pinheiro da Silva, Deborah L. McGuinness, and Richard Fikes, *A proof markup language for Semantic Web services*, Information Systems **31** (2006), no. 4-5, 381–395.
10. Todd Fast and Jiri Kopsa, *Twitter Data – a simple, open proposal for embedding data in Twitter messages*, http://twitterdata.org/, May 2009.
11. Harry Halpin, *A query-driven characterization of Linked Data*, 2nd International Workshop on Linked Data on the Web (Madrid, Spain), April 2009.
12. Toby Inkster, *Buzzword.org.uk draft: MicroTurtle (μttl)*, http://buzzword.org.uk/2009/microturtle/spec, June 2009.
13. Peter Millard, Peter Saint-Andre, and Ralph Meijer, *XEP-0060: Publish-subscribe*, http://xmpp.org/extensions/xep-0060.html, September 2008.
14. Paul Miller, Rob Styles, and Tom Heath, *Open Data Commons, a license for open data*, 1st International Workshop on Linked Data on the Web (Beijing, China), April 2008.
15. Alexandre Passant, Tuukka Hastrup, Uldis Bojars, and John Breslin, *Microblogging: a Semantic Web and distributed approach*, Proceedings of the 4th Workshop on Scripting for the Semantic Web, June 2008.
16. Alexandre Passant and Philippe Laublet, *Towards an interlinked semantic wiki farm*, 3rd Semantic Wiki Workshop, 2008.
17. Marko Rodriguez, *A graph analysis of the Linked Data cloud*, KRS-2009-01, February 2009.