Dynamic Associative Relationships on the Linked Open Data Web

Pablo N. Mendes, Pavan Kapanipathi, Delroy Cameron, Amit P. Sheth
Kno.e.sis Center, Wright State University
3640 Col. Glenn Hwy, Dayton, OH 45410 USA
pablo,pavan,delroy,amit@knoesis.org

ABSTRACT

In this work we approach relationships on the Linked Open Data Web as key facilitators of information exploration. **Linked Open Data (LOD)** principles contribute to a shift in paradigm for information representation and access, enhancing the ability of users and computers to connect, browse and query data on the Web through standard languages and protocols.

We present a brief discussion on the current relationship types on the Web, and observe the need for the extraction of a particular type. We focus on **trending socially contextual relationships** since they highlight the dynamics of the social interaction between users creating, linking, and consuming information on the Web.

Our extraction method is based on the identification of co-occurring entity mentions in microblog posts. Real time entity and relationship extraction can be useful tools for the on demand extension of the Web of Data. We demonstrate the usefulness of this approach in the context of two applications: database exploration and semantic browsing for exploratory search.

Keywords

real-time web, linked data, semantic web, microblogging, social media

1. INTRODUCTION

On the emerging Web of Data, the first class objects are not pages, but elements representing real world entities (e.g. a book in a library). The Linked Open Data (LOD) [6] initiative defines a set of good practices for data sharing and linking on this Web of Data. Currently, LOD sources comprise more than 13 billion triples.

However, it has been a common misconception to approach LOD as a data sharing revolution. In fact, offering structured data for third party consumption has been en vogue since the Web2.0 era of RESTful services. The leap made by the LOD initiative was the promotion of standards for (semi)structured data source linkage, much like hyperlinks and HTTP did for textual documents. LOD relationships properties linking two URIs - have the ability to lead a user from one piece of information to the next and therefore lay down the paths to be followed on the Web of objects.

In current LOD sources, many of the existing relationships are definitional, i.e. intended to describe what an object is. For example, the description for *Sprite Zero*[1] states it is a carbonated drink and, amongst other properties, links to some ingredients. Definitional relationships keep the navigation constrained to a tightly bound static set of facts. For the Web of Data to be useful for the regular user, definitional relationships are necessary, but not sufficient. For example, what products compete with *Sprite Zero*, what are the sale points for this product? In such cases, there is demand for less constrained, associative relationships that will lead the user to more loosely related, but relevant resources.

Particularly, there is a need for relationships representing more dynamic information that may change with context, time, location and social aspects. For instance, users are often interested in opinions or links to related products (e.g. Sprite is better than Sierra Mist[2]). That association between the two products is non-definitional, may only be interesting in certain regions (where both products are sold), or in specific times (when a new product is released).

We are particularly interested in these kinds of relationships, which we will call **socially contextual relationships**. Such relationships can be useful to contextually enhance browsing, retrieval and exploratory tasks in general by providing users with socially relevant cues. In the aforementioned example, the connection between Sprite Zero and related products should be influenced by the user's social network. If some friend prefers Sierra Mist, that is likely to be an interesting fact to be presented when the user is browsing Sprite Zero. One question that comes to mind is where can those relationships be obtained?

Microblogging services such as Twitter[4] have emerged as the preferred medium for distribution and consumption of unfiltered information including disseminating breaking news, individual opinions, organizing communities, etc. It composes an information base on top of which web intelligence platforms can be constructed. Microblog users distribute information in high volume, in real time and in textual format. The task of making sense of the aggregation of this information is not trivial.

In this work we present our implementation of a mixed retrieval/extraction system for the real-time extraction of relationships for a given LOD resource. Our implementation is composed mainly by rule based and lexicon based extractors. We describe an application integrating data from Freebase

Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.

[1] `http://dbpedia.org/resource/Sprite_Zero`
[2] `http://twitter.com/nkwhitten/status/8083355224`

and Twitter, and discuss the impact of our approach in two applications. First we discuss the use of D2R Server [2], a tool that maps relational databases to RDF and provides a Web interface for exploration of the underlying data. Finally we look at how our LOD expansion approach can impact the use of Scooner [8], a semantic browser focused on the task of exploratory search.

2. RELATIONSHIPS ON THE WEB OF DATA

Moving from the Web of pages to a Web of semantically rich objects will require taking a fresh look at relationships. In this section we provide a brief analysis of one snapshot of the Web of Data. We do not aim at a comprehensive, objective evaluation. Our intention is simply to reflect on the types of relationships we can currently encounter on the Web of Data, in order to make a connection to the types of relationships we consider most interesting for the discussion in this work.

We use the Billion Triples Challenge 2009 (BTC2009) dataset [1]. The BTC2009 dataset was crawled during February and March 2009 based on datasets provided by Falcon-S, Sindice, Swoogle, SWSE, and Watson using the Multi-Crawler framework. The creators of the BTC2009 dataset provide general statistics on 1.14 billion statements (triples) crawled, the 50 most frequent classes, and RDF properties (relationships). In order to broadly assess what kinds of information the relationships convey, we manually classified each relationship type in three axes: (i) **intent**: definitional or associative relationships. (ii) **continuity**: static or dynamic. (iii) **scope**: contextual or context-independent.

We used our subjective judgment to decide on the classification. When possible, we referred to authoritative descriptions of the relationships, but in many cases the definitions are ambiguous or non-existent definitions. In many cases the relationships have been clearly misused, as it is the case with the class *foaf:image*[3] that frequently appears used as a relationship in the BTC2009 dataset. Since it is impractical to investigate each and every one of the billion triples, our classification captures a subjective judgement on the perceived usage of each relationship on the schema level.

On the intent axis, we classified as **definitional** relationships those whose primary intent is to describe the identity (e.g. the URI, SSN), the essence of a concept (e.g. *rdf:type*), to define its meaning (e.g. *skos:definition*), or indicate symbols of reference (e.g. *foaf:name*, *foaf:depiction*). Associative relationships, on the other extreme, do not intend to describe the meaning of an object, but to make explicit some association between objects (e.g. *foaf:knows*, *rdfs:seeAlso*, *emph:nearbyFeatures*).

On the scope axis, we qualified as **contextual** the relationship that is dependent on or valid only within some context (e.g. *rdfs:label* is dependent on language,), and as **context independent** the relationship that is not expected to change with context (e.g. *foaf:name*).

Finally on the continuity axis, we assessed the character of the relationship with regard to time. Static relationships maintain their value, while dynamic relationships are subject to expiration or expected change (e.g. the president of a country). Although *George Bush* was very related to the resource *President* a few years ago, now it would be probably more accurate to link *President* to *Barack Obama*.

[3] `http://xmlns.com/foaf/spec/#term_Image`

On Figure 2 we show our subjective assessment of the top 50 relationships from the BTC2009 dataset according to the three aforementioned axes. The percentages represent the sum of the occurrences on the dataset, where an occurrence is defined by a triple containing the given relationship. Notice that **contextual dynamic associative** relationships (e.g. president, near) seem to be highly neglected. This observation encourages further research on this type of relationship. Particularly, we concentrate on relationships with the following character:

1. Its **dynamics** presents an uptrend at the current moment: for the entity *Barack Obama*, there is a wealth of definitional information. However, for this author's current spatio-temporal context, *Healthcare* and *Iraq War* would be more interesting related entities.

2. Its **context** is socially defined: Can we use someone's social network to boost the ranking of some relationships to a given resource? For instance, for the users looking at this resource, can we capture what the world suggests, toward the realization of a 'social' *rdfs:seeAlso*?

These **trending socially contextual relationships** are specially interesting since they highlight the dynamics of the interaction of human beings creating, linking, and consuming information on the Web. In the following section we present our approach for extracting one instance of those relationships with the potential of extending the Web of Data.

3. EXTENDING THE LOD WITH REAL TIME LINKS

In 1945, Vannevar Bush described the Memex device as one in which navigation across a document space would be driven by the users' interpretation of the content. According to Dr. Bush's understanding, the brain works by association: "With one item in its grasp, [the brain] snaps instantly to the next that is suggested by the association of thoughts." [7] Inspired by Dr. Bush's vision, towards realizing the Relationship Web vision [10], we approach the Linked Open Data Web from the perspective of a user attempting to navigate between entities to explore information, where new information is sought on a defined topic.

We propose to enhance LOD exploration with a similar service to that of the Memex device. As the user is exploring an LOD resource, this service will instantly 'snap' to other resources that are suggested by the user's context. The context of interest here is the social context. Imagine a Social Memex: a collective memory device that stores the knowledge of your social network. Or more ambitiously, a Global Memex, storing knowledge of all Web users including their social and other contexts. In the context of this work, we focus on tapping into the knowledge of Web users through microblog posts from Twitter, as at the time of writing it is a popular source with ramifications on many communities. Other use cases may call for other sources of social conversations and processing methods.

In simple terms, our approach is to gauge the change in association importance between any two given entities based on to the chatter involving those entities. As a user is browsing through an entity on the LOD, we search Twitter for the labels of that entity, i.e. the values for the definitional property *rdfs:label* of that entity. As we receive the search

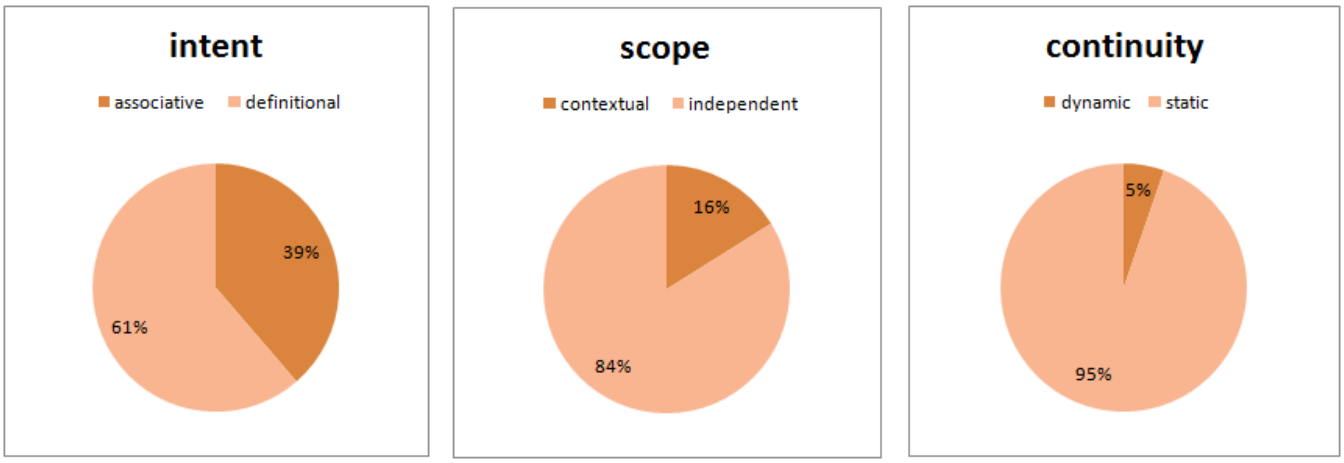

Figure 1: Distribution of the top 50 relationships in the Billion Triples Challenge 2009 dataset.

results, we extract all entity mentions in every tweet returned. We then aggregate the entity mentions and show a list of suggested 'trending associations' to other entities in the LOD based on some ranking strategy (e.g. frequency). When browsing the description for *Obama*, for instance, by searching Twitter we could extract entities like *Iran* and *Health care* that were commonly co-occurring in tweets in the United States. These links to other entities are real-time extensions to the LOD that aim at enhancing the user ability to explore the Web by taking advantage of contextually relevant information. Note that the extracted association may be a newly discovered association, or it may be an indication of higher contextual importance of an existing relationship.

3.1 Extracting trending associations

Our data processing focuses on the extraction of entity mentions as descriptors of the content in tweets. The corpus for extraction is obtained dynamically as a user is browsing an LOD entity. We send a search query to the Twitter Search API, obtain and parse results in JSON, subsequently sending each tweet through our information extraction process. There is a strong requirement for performance, since the users are not likely to spend more than a few seconds until they take the next navigation step. We therefore choose a simple, but reasonably efficient entity mention recognition algorithm. We extract a list of known 'surface forms' (entity labels) from a wide coverage LOD subset and build an in memory representation optimized for string matching.

In our current implementation we use a set of 2M entities from Freebase [3] and DBpedia [5]. We load the entity set as a trie (prefix tree) in memory, allowing us to perform longest common substring match at time complexity $O(LT)$ where L is the number of characters and T is the number of tokens in the sentence provided as input.

One may ask if such a crude entity extraction approach would be able to capture event descriptors. Or someone may ask if the content in tweets would even contain any entities to provide meaningful evidence of associations. In order to shed some light on these issues, we collected **1,391,791** tweets on the health care topic using the crawling technique described by Nagarajan et al. [9]. To provide a gold standard of event descriptors, we collected **2,665** article abstracts from the New York Times **2009-07-25 to 2010-02-01**. We performed entity extraction on both datasets using the techniques described in this work. Figure 3.1 shows a series of entity frequencies per date. Each curve represents an entity **(president, senate)**. Each point on the X axis represents a date, starting from **2009-08-18** until **2009-10-30**. On the Y axis you see the frequency of occurrence of the given entity on the corresponding date. Although some shifts and mismatches can be observed, the overall trend seems to be consistent. Peeks of entities obtained from the tweets seem to correspond to peeks in the gold standard. That initial assessment provides encouragement for more in depth evaluations that will be performed in future work.

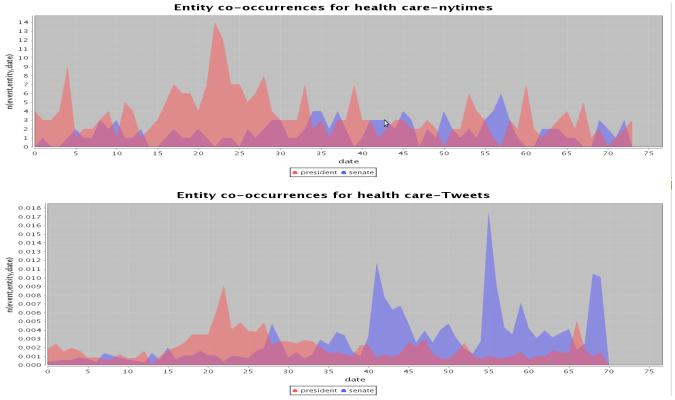

Figure 2: Distribution of the most frequent entities per date on the healthcare topic.

3.2 Ranking

After extracting the entity mentions from tweets, we aggregate the counts and rank the associations. The simplest approach for ranking is purely based on term frequency. The best ranking scores are attributed to the entities that co-occurred in higher frequency - i.e. appeared more frequently on the Twitter results on a search for the entity being browsed by the user.

Other alternative ranking methods may include the incorporation of domain model information to keep at lower ranks the 'usual suspects' definitional relationships that may seem uninteresting to the user. For example, *Obama* is an entity of type *President*, so this co-occurrence is highly expected and adds little information for the user.

In this work we implemented a simple ranking method that searches Twitter for the latest tweets (e.g. 100) mentioning the 'focus' resource (e.g. *Obama*), removes the resources that occur on the description of that focus resource (e.g. *President*), and ranks higher the resources with higher frequency of co-occurrence between with the 'focus' entity.

An interesting approach would be to prioritize entities that are trending on a certain time slice, but that were not popular before. However, realizing this approach in real time would require support from the Twitter API for searches in the past, a feature not present at the moment of writing. In order to circumvent that limitation, for focused domains the storage and use of historical data may be a viable alternative. This is a possible future direction for improvement.

3.3 Persisting useful associations

As users find new relationships and blaze new trails on the Web, they may find it useful to store some interesting associations to complement the model, enhance the description of an entity or simply provide a bookmark for further explorations. Human intervention can also be used as a way to validate an extracted association, in a similar way that clicks on search result page URLs validate the result's relevance to the keywords. The support for user feedback is dependent on the application. We present two example implementations on Section 4.

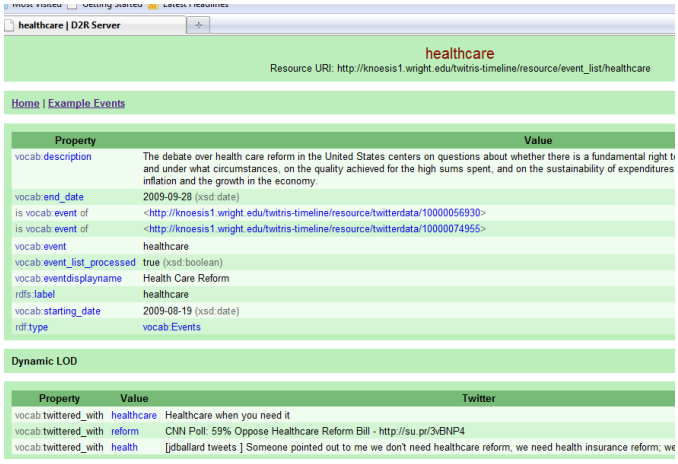

Figure 3: Screenshot of a page generated by the D2R Server, showing the Dynamic LOD extension at the bottom.

4. APPLICATIONS

4.1 Connecting Databases to the Real Time Web

While exploring a local database, users may be interested in discovering connections to the external world, particularly those that are more relevant at the current moment, for one reason or another. Take as an example a brand manager that would like to explore other products that are being compared to the product he or she manages. What are people saying about my product? Relationships from databases to the Web discovered in real time can provide valuable entry points for exploration of the Web of Data.

In order to support this and other use cases involving relational databases, we extended the D2R Server exploration application [2] for incorporating the functionality discussed in this paper. When the user is visualizing a resource in D2R, our extension automatically extracts the resource label, queries Twitter, performs entity extraction, aggregates and ranks the associations. The feedback incorporation mechanism is implemented as a 'Save' button that stores in the database a set of quadruples capturing the association and the context in which it was discovered.

```
Resource  twittered_with  Resource  G1
G1  created_by  user :_
G1  created_at  time :_
G1  geo location :_
```

Figure 4: Quads describing relationship to be saved.

4.2 Enhancing Exploratory Search Experience

Users engaging in exploratory search typically start with a tentative keyword query to gain access to documents as entry points into the information space. Since information that satisfies the user's need is often spread across the corpus, users spend considerable time exploring their environment to better understand it, selectively seeking and passively obtaining cues about their next steps. As highlighted by White et al. [11], for exploratory searchers, information-seeking is not just about the destination (e.g., the document or group of documents containing the required information). The knowledge acquired on their journey is also important (e.g., the results lists they view or the documents they encounter on their navigation path).

Figure 5: Triple Navigation using Scooner

In related research we have developed Scooner [8], an application supporting a new paradigm for exploratory search. Scooner uses semantic metadata to meaningfully link documents using named relationships between entities in those

documents. Ultimately, Scooner provides users with an avenue for browsing dynamically generated content by on-demand querying the knowledge base (e.g. LOD through SPARQL Endpoints) and document corpus (e.g. the Web through Yahoo! Search Boss API).

With a simple modification to Scooner's configuration, we can use the approach described in this work to suggest alternative browsing paths based on real time processing of social signals. Besides other relevant LOD sources, we can provide an additional SPARQL endpoint to Scooner, routing its navigation queries also to our API. By querying our system, Scooner will be able to suggest contextual relationships that complement the data from other SPARQL endpoints.

For example, with the user's Twitter account name at hand, we can dynamically search the public tweets in their network (i.e. their followers and those they are following), then perform information extraction and ranking on such tweets, suggesting trending entities within their social context. This is a realization of socially contextual relationships as browsing options for Scooner's users.

As the user interacts with Scooner, the system records each user move resulting on a semantic trail log. The system offers 'thumbs up' and 'thumbs down' commands at the interface, where a user can choose to persist or remove a given relationship from the trail log. These saved navigation paths represent searchable, evolving artifacts of knowledge that enhance and extend the Linked Open Data from that user's perspective. We perceive this as a realization of the aforementioned Social Memex vision.

5. CONCLUSION

Linked Open Data principles are an important advancement towards a global space of associated data, leading users from one piece of relevant information to the next. There is a need, however, for relationships representing more dynamic information that may change with topic, time, location and social context.

In this work we have discussed the real time processing of microblog posts for the extraction of associative relationships for on demand enrichment of the LOD cloud, increasing connectivity with contextually relevant, trending information. We presented two applications, demonstrating the usefulness of trending socially contextual relationships for browsing relational databases and performing exploratory search tasks on the Web.

Future work includes the incorporation of information extraction techniques for named entity recognition and relationship extraction, addition of other textual sources beyond microblogs, rigorous evaluation of the techniques for accuracy and real-time performance, as well as the expansion of the applications to other use cases.

6. ACKNOWLEDGMENTS

Many thanks to Christopher Thomas, Orsolya Szabo and Cory Henson for their valuable comments on earlier versions of this manuscript.

Thanks to the members of the Twitris project[4] for motivation and experience exchanged: Karthik Gomadam, Meena Nagarajan, Ashutosh Jadhav, Wenbo Wang, Raghava Mutharaju, Pramod Anantharam and Vinh Nguygen.

[4] `http://twitris.knoesis.org`

7. REFERENCES

[1] Billion triple challenge 2009 dataset and statistics. `http://vmlion25.deri.ie/`, Mar. 2009.

[2] D2r server publishing relational databases on the semantic web. `http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/`, Mar. 2010.

[3] Freebase - a wealth of free data. `http://www.freebase.com/`, Mar. 2010.

[4] Twitter. `http://twitter.com/`, Mar. 2010.

[5] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *In 6th International Semantic Web Conference, Busan, Korea*, pages 11–15. Springer, 2007.

[6] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.

[7] V. Bush. As we may think. *The Atlantic Monthly*, July 1945.

[8] D. Cameron, P. N. Mendes, A. P. Sheth, and V. Chan. Semantics-empowered text exploration for knowledge discovery. In *48th ACM Southeast Conference*, 2010.

[9] M. Nagarajan, K. Gomadam, A. P. Sheth, A. Ranabahu, R. Mutharaju, and A. Jadhav. Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. In *WISE*, pages 539–553, 2009.

[10] A. P. Sheth and C. Ramakrishnan. Relationship web: Blazing semantic trails between web resources. *IEEE Internet Computing*, 11:77–81, 2007.

[11] R. W. White, S. M. Drucker, G. Marchionini, M. Hearst, and Schraefel. Exploratory search and hci: designing and evaluating interfaces to support exploratory search interaction. In *CHI '07: CHI '07 extended abstracts on Human factors in computing systems*, pages 2877–2880, New York, NY, USA, 2007. ACM.