Augmenting Open Government Data with Social Media Data

Evangelos Kalampokis[1,2], Efthimios Tambouris[1,2], Michael Hausenblas[3], Konstantinos Tarabanis[1,2]
[1] Information Technologies Institute, Centre for Research & Technology - Hellas (CERTH/ITI), Thessaloniki, Greece
[2] University of Macedonia, Greece
{ekal, tambouris, kat}@uom.gr
[3] DERI-NUI Galway, Ireland
michael.hausenblas@deri.org

Abstract. The liberation of data produced and collected by governments around the globe promises to enable the development of added value services for the society. In this position paper we discuss the possibilities and challenges of combining government and social media data for the development of innovative services.

Keywords: Open government data, social web, linked data.

1 Introduction

The Open Government Data (OGD) movement evangelizes the need for making public sector information freely available in open formats and ways that enable public access and reuse. It promises to enhance transparency, enable economic growth, improve citizens every day life and support public administration function. In the last years, a number of OGD portals have been launched around the globe providing different types of public information (e.g. statistics, expenses, etc.) following various technological and organizational approaches [1].

OGD is part of the ongoing evolution of the Web where platforms and sites provide access to their data in a structured way that facilitates reuse. Examples include Facebook's Graph API[1], Twitter's RESTful API[2] and also the Linking Open Data project[3]. In this context, OGD could be combined and integrated with other open data on the Web in order to enable the development of innovative services. In particular, OGD can be integrated with governmental data sets provided by different agencies and countries, with formal data published by highly trustworthy sources such as New York Times and also with data from social Web. Towards this end, a stage model has been proposed as a high-level roadmap for OGD endeavors [2].

[1] http://developers.facebook.com/docs/reference/api/
[2] http://dev.twitter.com/doc
[3] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

Social Web is one of the wealthiest data sources on the Web. In the last years, social media have grown in popularity with millions of users producing tremendous amounts of data in various forms such as text messages, tags and multimedia content. For example, Twitter[4] has currently100 million active users that publish more than 200 million posts per day (as of September 2011).

In this position paper we discuss the possibilities and challenges of combining data from governmental sources and social media in order innovative services to be developed and provided to the society. Towards this end we analyze the conceptual and technical requirements of the approach and we also present some use cases in order to make our point clear. The use cases exploit data from data.gov.uk portal.

2 Open Government and Social Media Data

The idea behind the integration of government and social media data is that although they both may refer to the same real world entity, the former provides *objective* facts while the latter *subjective* thoughts and opinions. The integration of both could enable the composition of two complementary points of view of the same problem and thus enable a better and more in depth understanding of it.

In order to clarify our approach we employ the scenario described in [3]. According to it the United Kingdom's government announces a bill on public budget cutting in police forces. Before its enactment the government needs to know the public opinion on this issue and more importantly the opinion of residents of areas with high crime rates. According to our approach, social data will be collected from Twitter before and after the announcement of the bill. In order to identify only those tweets that are posted by residents of areas with crime level above average we converge data from Data.gov.uk that provides crime levels and statistics in neighborhood areas in the 43 English and Wales's police forces through a RESTful API[5] and data from Twitter. By linking the "location" attribute of tweets to the "crime area" attribute of the Data.gov.uk dataset we can collect only the required tweets. Thereafter, the quantity and sentiment of the collected tweets is measured and also government datasets that could provide relevant objective facts are identified. In our case the same dataset from Data.gov.uk provides also data about the number of crime incidents in each area. By combining these three different values for every area, a model could be created which would provide an indication of public reaction in each area.

From a technical point of view *linked data* [4] could address the challenge of integrating data that refer to the same entity and are provided by different sources on the Web. In [3] a linked data based architecture has been proposed for the integration of open government and social media for supporting decision-making. The publication of OGD as linked data is fairly a straightforward task as the existing data is usually in structured formats. At the moment several portals such as data.gov.uk provide linked government data [5]. The creation, though, of linked data from social

[4] http://twitter.com

[5] http://data.gov.uk/apps/police-api

media content is a challenging task mainly because of the need for transforming unstructured text to structured data. This mainly refers to the extraction and disambiguation of real-world entities. In social media realm this task required advanced text analysis approaches and tools since the informal nature of the posts reduce the accuracy of existing NLP tools developed for analyzing casual text.

3 Exploring Data.gov.uk for 'joint points'

Data.gov.uk seems to be the most sophisticated OGD portal [1]. It currently contains more than 8.000 datasets. In order to identify possible 'join points' for diverse OGD datasets as well as for OGD datasets and social media data we have thoroughly explored and analyzed data.gov.uk. The result of this exercise is a list of the measured main real-world entities included in data.gov.uk along with the respective variables that describe them. Following an OLAP terminology[6] the measured values of the phenomena described in the data.gov.uk data sets along with the respective dimensions were identified. Common phenomena in data.gov.uk include health care, education, transport, employment, environment etc.

The most important dimension is the location which however is described in various geographical or administrative divisions e.g. local authorities, police force areas, criminal justice system areas, regions, government office regions (GOR), NHS board areas, municipals, assembly constituencies, crown prosecution service (CPS) areas etc. Other dimensions that also refer to real world entities are hospitals, drugs, sport clubs, colleges etc.

The convergence of different data sets has two preconditions:

- The conceptual relevance of the measured values and phenomena
- The matching of the dimensions. This matching is also known as the identification of 'joint points' for the datasets.

4 Use cases

The analysis of the data sets of data.gov.uk facilitates the identification of interesting use cases in which the proposed approach could be applied. Examples of such cases could be identification of areas with proneness to criminality, infer voting intentions, predict traffic conditions etc.

For the shake of brevity in this paper we only describe one of these scenarios. In elections the analysis objective data about demographics in different areas along with older election results could reveal patterns for the behavior of the voters. In addition the analysis of social media posts about the election could provide indications about voters' intention in different areas. Different indications for the outcome of the elections could be deduced by each of these analyses. However, the convergence of both the objective and subjective results could enable a more accurate prediction for the outcome of the elections.

[6] http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html

5 Conclusion

This position paper discusses the value and the possibilities of combining open government with social media data. It also highlights some technical requirements that are challenging regarding their implementation.

References

1. Kalampokis, E., Tambouris, E., Tarabanis, K.: A Classification Scheme for Open Government Data: Towards Linking Decentralized Data. International Journal of Web Engineering and Technology 6(3), 266–285 (2011)
2. Kalampokis, E., Tambouris E. and Tarabanis K., Open Government Data: A Stage Model. In: M. Janssen et al. (Eds): EGOV2011. LNCS 6846, 235-246, 2011.
3. Kalampokis, E., Hausenblas, M. and Tarabanis, K., Combining Social and Government Open Data for Participatory Decision-Making. In: E. Tambouris, A. Macintosh, and H. de Bruijn (Eds.): ePart 2011, LNCS 6847, 36–47, 2011.
4. Hausenblas, M.: Exploiting Linked Data to Build Web Applications. IEEE Internet Computing 13(4), 68–73 (2009)
5. Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W., mc schraefel.: Open Government Data and the Linked Data Web: Lessons from data.gov.uk. IEEE Intelligent Systems, http://doi.ieeecomputersociety.org/10.1109/MIS.2012.23