Integrating Web 2.0 Data into Linked Open Data Cloud via Clustering

Eirini Giannakidou and Athena Vakali

Aristotle University of Thessaloniki
Department of Informatics, Greece
{eirgiann,avakali}@csd.auth.gr

Abstract. In this position paper, we claim that useful information mined from social tagging systems can be published and shared as Linked Data, in order to be further exploited. The proposed approach aims: i) to be applied on any social tagging platform, and ii) utilize information that concerns the collective activity of users, not individual transactions. Such information involves topics of interests, emerging trends, events, communities around events/trends. To mine this knowledge from tagging data, a number of clustering algorithms are suggested.

Keywords: linked data, web 2.0, tagging, clustering

1 Introduction

With the advent of Web 2.0, tagging practices constitute a collective fashion of metadata creation, which is suggestive of individual users interests, since users tend to tag content that they find interesting. As more and more people have supported this surge, tag data provides a rich information source to study social patterns and emergent drifts/directions in the web user community. The main problem though regarding this information is that it is kept isolated in a number of proprietary systems and formats and, thus, it cannot be reused and further exploited.

Research carried out in the field of Semantic Web supports applications to provide extensibility, flexibility, interoperability and reusability. To address such issues in the tag space, many researchers claim that the application of mature semantic web technologies (e.g. formal representation of content, formal relations) on tagged data could add great value to the latter, as it may render a kind of structure to them, remove ambiguities and promote re-usage of content. A number of standards and initiatives in this direction involve SKOS [13], FOAF [12], SIOC [14] and MOAT [11]. OpenID [15] is also a step towards this direction, by providing an authorization mode, which is currently accepted by many sites, allowing, thus, consolidation of users' digital identities. However, despite the fact that interoperability between tagging systems is a subject of research, these approaches have not found, yet, widespread application and, so far, there is no common agreement on a formal representation of tagging activity between social tagging systems.

Lately, the Linked Data initiative developed which focuses on pragmatic approaches that help applications to share and connect data. Specifically, it suggests to publish data online using open standards, but, also, link to other existing datasets. In this position paper, we claim the publishing as Linked Data of information mined from tagging data via advanced clustering methods. The proposed approach can be applied on tagging data of any web 2.0 platform, provided that it offers API access to such data. Its application is expected to enable combination of data from different data sources in a standardized manner and, thus, allow the development of services that provide added-value to user community across application and domain boundaries. The rest of the paper is organized as follows. First, the current status in Linked Data and its applications is described, concluding in a quotation of issues that have not been addressed so far. Then, our proposal is given for publishing information from web 2.0 data as Linked Data. Finally, the paper ends with some applications and some general conclusions of the proposed approach.

2 The Linked Data Landscape in 2010 and what is missing

In 2006, Tim Berners-Lee presented *Design Notes* on how to write Linked Data on the Web, in which he outlined 4 basic principles that should be followed and are summarized as follows: Each resource should have an HTTP URI, so that it can be dereferenced either by users or by agents, as well as a semantic description which should be represented in standard formats such as RDF/XML. Further, the use of links that associate representations of the same resource in different datasets is encouraged, to promote discovery of related information regarding the resource at hand. Following this article, a number of research initiatives appeared that aimed at the production of linked data out of distributed datasets in the web, the more widely known being the Linking Open Data (LOD) project[1]. In October 2007, the datasets in LOD project consisted of over two billion RDF triples, which were interlinked by over two million RDF links. By September 2010 this had grown to 25 billion RDF triples, interlinked by around 395 million RDF links. Moreover, a number of applications were developed in which the aforementioned principles are followed, as described in [2].

A detailed review on linked data initiatives and applications can be found in [1]. As can be seen there, currently, the initiatives on linked data approaches are roughly categorized into two groups: i) approaches that mainly reuse content of datasets in the LOD cloud [3–5], and ii) approaches that enable associating HTTP URIs with User Generated Content (UGC), resulting, thus, in structured UGC [7, 6]. All such efforts that concern publication of linked data (i.e. turning raw web data into linked data) and their consumption in domain-specific or more generic frameworks have created a new landscape on the web map that contains tools, technologies and datasets following these principles and keeps on

[1] http://linkeddata.org/

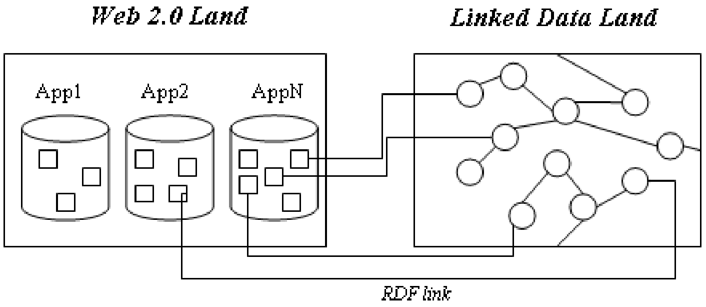

Fig. 1. Current status in the information binding between Web 2.0 the Linked Data Landscapes on the web map.

expanding as more and more applications and sites embrace linked data. Figure 1 illustrates schematically the current type of information binding between web 2.0 and linked data lands.

We will now focus on what is currently missing from the *Linked Data* landscape, concerning the interaction with datasets from the *Web 2.0* landscape. Given the popularity of web 2.0 platforms and the useful information that can be extracted from these sites, as reported in the related literature([16, 17]), we quote 2 shortcomings in the current binding interaction between these two landscapes:

- no mappings on the abundance of web 2.0 platforms to Linked Data (cf. App1 in Figure 1)
- no mappings on data aggregation level (currently, the mappings are on individual data level, so the "*Wisdom of the Crowds*" is lost)

To address the aforementioned issues, we propose a clustering approach in the next section.

3 Using Clustering for integrating Web 2.0 to Linked Data

The current shared vision for the future is one of semantically-rich information and service oriented architecture for global information systems. As individual web 2.0 platforms are not embracing semantic tools to unambiguously define their data, this vision is still far away from reality. Here we present an approach for integrating web 2.0 data to linked data, and, thus, get structural information out of raw web 2.0 data. An asset of the presented approach is that it does not bring any burden on individual web 2.0 platforms; instead it can be applied on each one of them, providing that the platform offers access to the user data.

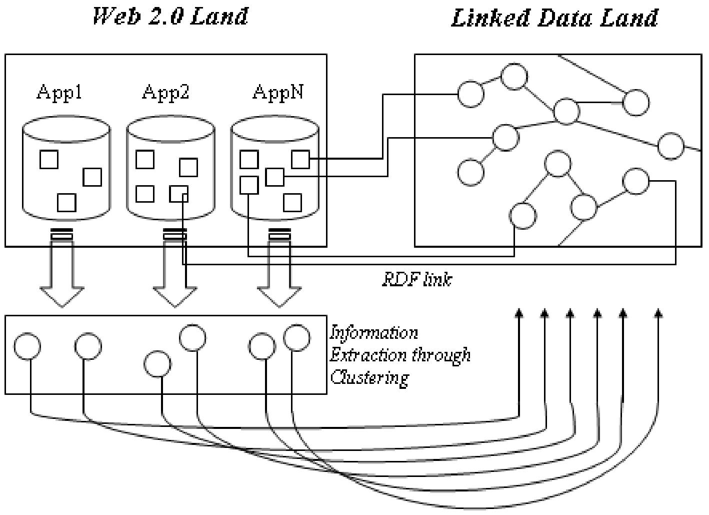

Fig. 2. Information binding between Web 2.0 the Linked Data Landscapes on the web map through clustering.

Furthermore, the proposed approach focuses on publishing as linked data not individual pieces of data, but information that is extracted from the collective user activity in a web 2.0 platform through clustering. The idea is given schematically in Figure 2 and summarized in the following steps:

- Apply clustering techniques on a web 2.0 platform data, to extract interested patterns in the data
- Map the terms mined in the extracted patterns to terms already in the LOD cloud
- Publish the mined information as linked data and create RDF links to the mapped terms in the LOD cloud

Clustering approaches have been heavily researched for mining useful information - such as topics, trends, user communities around specific topics and so on - out of tagging data in web 2.0 platforms. The clustering techniques we propose to employ during the $1st$ step may be either i) co-clustering algorithms to get groups of related content with associated tags, which can be mapped to specific topics ([8]), ii) user-based clustering to extract topics of interests related to specific user communities inside the web 2.0 platform ([9]), or iii) time -aware clustering to track events or how the users are attracted by particular topics over time ([10]).

Having applied the clustering algorithm, the information that has been extracted in each cluster should be mapped to terms already in the LOD cloud. During this step, a semantic indexer like Sindice [18] may be employed, which takes input terms and returns related resources in DBPedia.

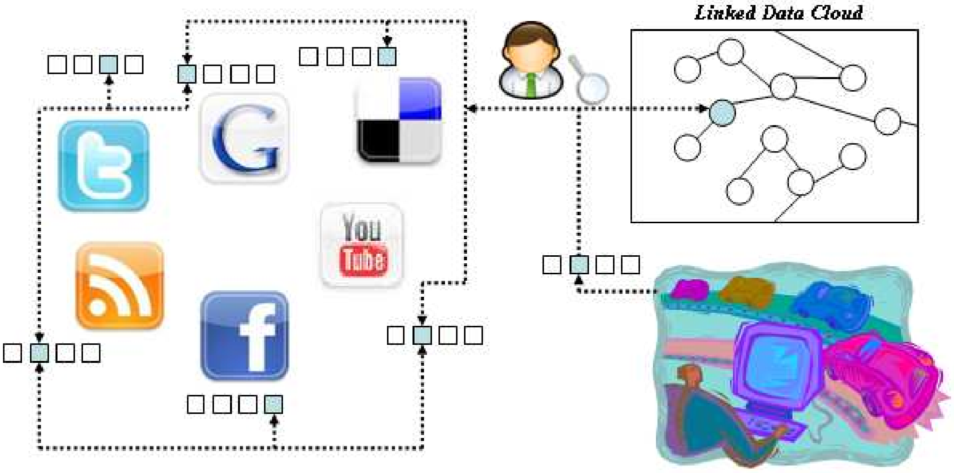

Fig. 3. A policy maker is able to analyze information regarding a topic X (which is depicted as light blue rectangle) from all web 2.0 sites that extract information related to topic X, as this information is published as linked data and, therefore, part of the LOD cloud. This information may be further combined with sensor data extracted from the smart devices of the city (cars, cameras, etc).

Finally, we propose publishing of this information as Linked Data, so that it can be further exploited. This involves representation of the data into RDF and creation of RDF links that connect the mined information with the mapped terms already in the LOC cloud, which were found in the previous step.

4 Vision

We envisage the employment of the proposed approach in the Future Internet, the so-called *Internet of things*, in a number of contexts that aim at employing jointly information coming from sensor data of "smart" devices and information coming from users' activities in web 2.0 sites. Publishing such information as Linked Data allows data that refer to the same topic, but come from different web 2.0 sites to be interlinked, so that the maximum span of available information regarding a particular topic will be exploited.

Figure 3 illustrates a sample scenario of use of the proposed approach in the context of a smart city environment. Specifically, we claim that the proposed approach can be used by automatic policy makers to improve a transportation system under special circumstances, or even to design a better mobility plan for the city. Smart cities will become a reality, when the objects around us become smart and begin to embody the functionality of our computers and mobile devices (Internet of Things). Thus, there will be internet-enabled cars, taxis, buses, cameras on the road and so-on. All such information will be forwarded to the policy maker agent to record the current status in the roads of the city. Additional information extracted from web 2.0 sites may signify events that are expected to cause disruption to normal city traffic (e.g. concerts, protests). As a result, the moving vehicles will be aware of such events and act accordingly (e.g. in case of a concert with high popularity, a number of taxis may decide to go to that direction to serve the music fans).

In the long run, the sensor data coming from devices may be combined with analysis of information coming from web 2.0 users, so that new mobility plans can be proposed (e.g. bicycle lanes, pedestrian roads). The novel aspect here is that the proposals will be driven not only by the actual needs of the citizens as those are recorded in the sensor data of the devices, but also by citizens' opinions and emerging trends, as those are expressed through citizens' activities in web 2.0 sites (e.g. if a lot of people criticize the parking problem in an area, upload photos that illustrate the problem, a policy maker could evaluate parking deficiencies in the specific area and recommend solutions such as additional parking, parking pricing and management strategies). The role of Linked Data in such scenarios is to link information that is mined by various web 2.0 sites that refer to the same topic.

5 Conclusions

Here we suggested a way to publish and consume linked data out of tagging data; the presented approach can be applied on tagging data of any web 2.0 application, provided that it offers API access to such data. The benefit of such a process is that dynamic information about a web community (such as trends, topics of interest of user groups, etc.) becomes available to all. This information is of particular importance to people/professions that are interested in listening the public pulse, such as businessmen and public administration bodies.

References

1. M. Hausenblas, Linked Data Applications - The Genesis and the Challenges of Using Linked Data on the Web, DERI Technical Report, (2009).
2. M. Hausenblas: Exploiting Linked Data to Build Web Applications, IEEE Internet Computing (vol. 13 no. 4) pp. 68-73 (2009)
3. S. Softic, M. Hausenblas, Towards Opinion Mining Through Tracing Discussions on the Web, In: Social Data on the Web SDoW 2008 Workshop at the 7^{th} International Semantic Web Conference Karlsruhe, Germany: (2008).
4. G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, R. Lee: Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Connections. In 6th European Semantic Web Conference, ESWC 2009,pages 723-737. Springer, (2009).
5. C. Becker, C. Bizer: DBpedia Mobile: A Location-Enabled Linked Data Browser. In WWW 2008 Workshop: Linked Data on the Web (LDOW2008), Beijing, China, (2008).

6. T. Heath and E. Motta. Revyu: Linking reviews and ratings into the Web of Data. Journal of Web Semantics, 6(4):266-273, (2008).
7. Faviki - Social bookmarking tool using smart semantic Wikipedia (DBPedia) tags, `http://www.faviki.com`
8. E. Giannakidou, V. Koutsonikola, A. Vakali, I. Kompatsiaris: "Co-clustering Tags and Social Data Sources", In Proc. 9th International Conference On Web-Age Information Management (WAIM' 2008), IEEE Computer Society, pp 317-324, Zhangjiajie, China (2008).
9. V. Koutsonikola, A. Vakali, E. Giannakidou, I. Kompatsiaris: "Clustering Users of a Social Tagging System: A Topic and Time Based Approach" In Proc. of the 10th International Conference on Web Information Systems Engineering (WISE 2009), Springer Verlag, pp 75-86, Poznan, Poland (2009).
10. E. Giannakidou, V. Koutsonikola, A. Vakali and I. Kompatsiaris: "Exploring Temporal Aspects in User-Tag Co-Clustering", In Proc. of the 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2010), Desenzano del Garda, Italy (2010).
11. A. Passant, P. Laublet: Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data, in Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China, (2008)
12. Mike Graves, Adam Constabaris, Dan Brickley: FOAF: Connecting People on the Semantic Web, doi = 10.1300/J104v43n03 11, In: Cataloging Classification Quarterly, Vol. 43 (2007)
13. Alistair Miles, Brian Matthews, Michael Wilson, and Dan Brickley: SKOS core: simple knowledge organisation for the web. In Proceedings of the 2005 international conference on Dublin Core and metadata applications: vocabularies in practice (DCMI '05). Dublin Core Metadata Initiative , Article 1 (2005)
14. John G. Breslin, Stefan Decker, Andreas Harth, and Uldis Bojars. SIOC: an approach to connect web-based communities. Int. J. Web Based Communities 2, 2 133-142. DOI=10.1504/IJWBC.2006.010305 http://dx.doi.org/10.1504/IJWBC.2006.010305 (2006)
15. David Recordon and Drummond Reed. OpenID 2.0: a platform for user-centric identity management. In Proceedings of the second ACM workshop on Digital identity management (DIM '06). ACM, New York, NY, USA, 11-16. DOI=10.1145/1179529.1179532 http://doi.acm.org/10.1145/1179529.1179532 (2006)
16. Andreas Hotho, Robert Ja"schke, Christoph Schmitz, Gerd Stumme: Trend Detection in Folksonomies. In Proceedings of the First International Conference on Semantics And Digital Media Technology (SAMT), Vol. 4306, pp. 56-70 (2006).
17. Peter Mika: Ontologies Are Us: A Unified Model of Social Networks and Semantics. International Semantic Web Conference, pp: 522-536 (2005).
18. Eyal Oren, Renaud Delbru, Michele Catasta, Richard Cyganiak, Holger Stenzhorn, and Giovanni Tummarello: Sindice.com: a document-oriented lookup index for open linked data. Int. J. Metadata Semant. Ontologies 3, 1 37-52. (2008).