Towards a DBpedia of Tourism: the case of Tourpedia

Stefano Cresci, Andrea D'Errico, Davide Gazzé, Angelica Lo Duca, Andrea Marchetti, Maurizio Tesconi

Institute of Informatics and Telematics, National Research Council, via Moruzzi 1, 56124 Italy
email: [name].[surname]@iit.cnr.it

Abstract. In this paper we illustrate Tourpedia, which would be the DBpedia of tourism. Tourpedia contains more than half a million places, divided in four categories: accommodations, restaurants, points of interests and attractions. They are related to eight locations: Amsterdam, Barcelona, Berlin, Dubai, London, Paris, Rome and Tuscany, but new locations are continuously added. Information about places were extracted from four social media: Facebook, Foursquare, GooglePlaces and Booking and were integrated in order to build a unique catalogue. Tourpedia provides also a Web API and a SPARQL endpoint to access data.

1 Introduction

The concept of Semantic Web was introduced by Tim Berners Lee in 2001[2]. His main idea consisted in migrating from the Web of documents to the Web of data. The purpose of the Web of data is to connect concepts and contents to each other, instead of simply connecting documents. Thus the Web of data has led to the conversion of existing documents to linked data [6], and to the creation of new datasets[1]. Among them, one of the most exploited datasets is DBpedia[2], which is the linked data version of Wikipedia[3].

DBpedia is available in different languages. Its English version contains about 4.0 million things, classified in different categories, including people, places, creative works, organizations, species and deseases. However, DBpedia, as well as Wikipedia, contains only a small number of things related to the tourism domain, such as accommodations and restaurants. In addition, to the best of our knowledge, only few linked datasets have been implemented in the field of tourism. Among them, the case of El Viajero[4], which provides information about more than 20.000 travel guides, pictures, videos and posts, and that of Accommodations in Tuscany[5], which contains the list of accommodations in

[1] For a list of shared datasets, please look at: http://datahub.io.
[2] http://dbpedia.org
[3] http://wikipedia.org
[4] http://datahub.io/dataset/elviajero
[5] http://datahub.io/dataset/grrt

Tuscany, Italy. For more details about datasets about tourism, please refer to: http://datahub.io/dataset?q=tourism.

In this paper we illustrate Tourpedia, which would be the DBpedia of Tourism. Tourpedia is reachable through its portal[6] and is available also in the datahub.io platform[7].

Tourpedia was developed within the OpeNER Project[8] (Open Polarity Enhanced Name Entity Recognition), whose main objective is to implement a pipeline to process natural language.

The usage of Tourpedia could be very various. For example, it could be used to perform named entity disambiguation in tourism domain, or to extract the most appreciated points of interest in a town.

2 Tourpedia

Figure 1 illustrates the Tourpedia architecture. The Data Extraction module consists of four ad-hoc scrapers, which extract data from four social media: Facebook[9], Foursquare[10], Google Places[11] and Booking[12]. We chose these social media firstly because they are very popular and secondly because they provide an easy way to extract data. The scrapers of Facebook, GooglePlaces and Foursquare exploit the RESTful APIs the social media provide, while the Booking scraper extracts information from each accommodation page.

The Named Entity repository contains two main datasets, which belong to the specific domain of tourism: Places and Reviews about places. The dataset of Places contains more than 500.000 places in Europe divided in four categories: accommodations, restaurants, points of interest and attractions[13]. At the moment the following locations are covered: Amsterdam, Barcelona, Berlin, Dubai, London, Paris, Rome and Tuscany. Places were elaborated and integrated through the Data Integration module in order to build a unique catalogue. Data Integration was performed by using a merging algorithm based on distance and string similarity.

The dataset of Reviews contains about 600.000 reviews about places. Reviews were analysed through the OpeNER pipeline in order to extract their sentiment.

2.1 Web application

Tourpedia provides also a Web application[14] [5], which shows the sentiment about places on an interactive map, which is Google Maps-like.

[6] http://tour-pedia.org
[7] http://datahub.io/dataset/tourpedia
[8] http://www.opener-project.eu
[9] http://www.facebook.com
[10] http://foursquare.com
[11] https://plus.google.com/u/0/local
[12] http://www.booking.com
[13] http://tour-pedia.org/about/statistics.html
[14] http://tour-pedia.org/gui/demo/

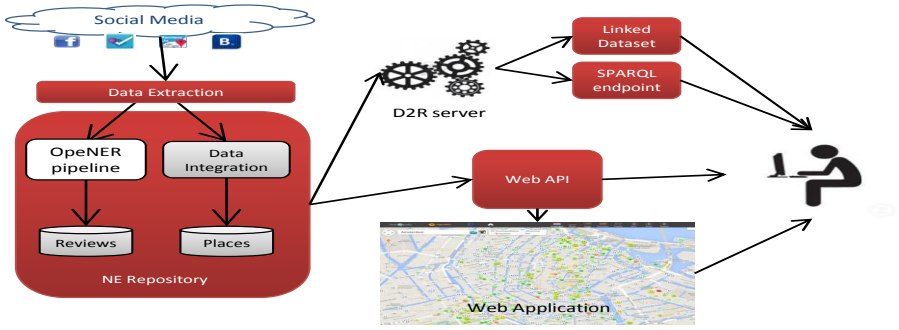

Fig. 1. The architecture of Tourpedia.

The sentiment of a place is calculated as a function of all the sentiments of the reviews about that place. In order to retrieve the sentiment of a review, the OpeNER pipeline was used. In particular, each place is associated to zero or more reviews extracted from social media (i.e. Facebook, Foursquare and Google Places). Each review is processed through the OpeNER pipeline and is associated to a rate, which expresses its specific sentiment.

2.2 Linked Data

Tourpedia is exposed as a linked data node and provides a SPARQL endpoint[15]. The service is implemented through the use of a D2R server[16]. For each place, the following ontologies are used to represent it: VCARD [9] and DBpedia OWL[17], for generic properties; Acco [8], Hontology [4] and GoodRelations [7] for domain-specific properties. In a previous work [1], we illustrated the employed ontologies and structures of accommodations as linked data. In order to fulfill the principles of linked data [3], each location is linked to the same location in DBpedia.

2.3 Web API

Tourpedia provides a RESTful API[18] to access places and statistics. The output of each request can be JSON, CSV and XML. For example, a search request about Places is an HTTP URL of the following form:

`http://tour-pedia.org/api/getPlaces?parameters`

where `parameters` must be at least one one of the following: *location* (the location of the places), *category* (the type of the places such as accomodation), attraction, restaurant, poi), and *name* (the keyword to be searched).

[15] http://tour-pedia.org/sparql
[16] http://d2rq.org/
[17] http://wiki.dbpedia.org/Ontology
[18] http://tour-pedia.org/api

3 Conclusions and Future Work

In this paper we have illustrated Tourpedia, which would be the DBpedia of Tourism. It could be interesting a deeper connection between Tourpedia and DBpedia. At the moment, in fact, only locations are connected to DBpedia. As future work, we are going to align also attractions and points of interest contained in Tourpedia to DBpedia.

Tourpedia could be exploited both by tourism stackholders to get the sentiment about touristic places and by common users.

At the moment, the procedure to update datasets is manual. As future work, we are going to define a semi-automatic procedure to update them and to add new locations.

Acknowledgements

This work has been carried out within OpeNER project, co-funded by the European Commission under the FP7 (7th Framework Programs Grant Agreement n. 296451).

References

1. Bacciu, C., Lo Duca, A., Marchetti, A., Tesconi, M.: Accommodations in Tuscany as Linked Data. In: Proceedings of The 9th edition of the Language Resources and Evaluation Conference (LREC 2014). pp. 3542–3545 (May, 26-31 2014)
2. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American 284(5), 34–43 (May 2001), http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. J. Semantic Web Inf. Syst. 5(3), 1–22 (2009)
4. Chaves, M.S., de Freitas, L.A., Vieira, R.: Hontology: A multilingual ontology for the accommodation sector in the tourism industry. In: Filipe, J., Dietz, J.L.G. (eds.) KEOD. pp. 149–154. SciTePress (2012)
5. Cresci, S., D'Errico, A., Gazzé, D., Lo Duca, A., Marchetti, A., Tesconi, M.: Tourpedia: a Web Application for Sentiment Visualization in Tourism Domain. In: Proceedings of The OpeNER Workshop in The 9th edition of the Language Resources and Evaluation Conference (LREC 2014). pp. 18–21 (May, 26 2014)
6. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool, 1st edn. (2011), http://linkeddatabook.com/
7. Hepp, M.: Goodrelations language reference. Tech. rep., Hepp Research GmbH, Innsbruck (2011)
8. Hepp, M.: Accommodation ontology language reference. Tech. rep., Hepp Research GmbH, Innsbruck (2013)
9. Iannella, R., McKinney, J.: VCARD ontology. Available at: http://www.w3.org/TR/vcard-rdf/. Tech. rep. (2013)