An Overview of the Tourpedia Linked Dataset with a Focus on Relations Discovery among Places

Davide Gazzè, Angelica Lo Duca, Andrea Marchetti, Maurizio Tesconi
Institute of Informatics and Telematics
National Research Council
via Moruzzi 1, 56124 Italy
[name].[surname]@iit.cnr.it

ABSTRACT

Tourpedia (http://tour-pedia.org) is an open initiative which contains a linked dataset of tourism places, i.e. accommodations, attractions, points of interest (POIs) and restaurants. Tourpedia extracts and integrates information about places from four different social social media: Facebook, Foursquare, Google Places and Booking.com. The resulting knowledge base currently consists of more than 6M RDF triples and describes almost 500.000 places, each of which is identified by a globally unique identifier, which can be dereferenced over the Web into a RDF description. This paper gives an overview of the Tourpedia knowledge base and illustrates how new relations are discovered among places through Named Entity Recognition (NER) tools.

1. INTRODUCTION

Over the last years the Web has seen the growth of two big initiatives: linked data and social media. On the one hand, the linked data initiative [3] allows the creation of a global-scale interlinked data space, known as the Web of Data, by exposing datasets previously isolated as data graphs, which can be interlinked and integrated with other datasets. Thanks to this, an unprecedented amount of linked data sources were recently produced and continues growing fast. On the other hand, social media and mobile technologies [8] have been fostering crowdsourced databases, such as places, events, reviews, etc. that have huge potential of enrichment in mashup applications. Recently researches have used social media data for enrichment of spatial and spatiotemporal data. If combined, these two kinds of data sources can be very useful for creating mashup applications and resources.

This paper describes the Tourpedia[1] [6] initiative, a linked dataset, which combines and aggregates data extracted from four social media: Facebook[2], Foursquare[3], Google Places[4] and Booking.com[5]. Tourpedia contains almost half a million places, divided in four categories: accommodations, restaurants, points of interests (POIs) and attractions.

With respect to the state-of-the-art tourism datasets, Tourpedia provides the following features. Firstly, Tourpedia elaborates reviews of places extracted from social media to calculate their overall sentiment. This constitutes a novelty compared to existing datasets, because it provides the access to customer-satisfaction derived from social media. The availability of the sentiment about a place has been playing an important role over the last years. In fact, a recent study [1] shows that the percentage of consumers consulting reviews at social media prior to booking a hotel room has steadily increased over time, as well as the number of reviews they are reading prior to making their hotel reservation. Therefore, there is a clear need to manage and understand the knowledge conveyed by opinions and customer-generated online content [17]. Tourpedia satisfies this need by providing an easy way to access the sentiment about tourism places. Secondly, Tourpedia consolidates features of accommodations, by providing additional details. This aspect could strengthen research in the tourism industry. In fact, there are studies discussing which hotel characteristics travellers' interests require [11] or which kind of accommodations some categories of travellers prefer [13]. A resource like Tourpedia could encourage and facilitate these kinds of studies, by providing specific information about each accommodation. Finally, for each accommodation, Tourpedia provides also links to related attractions. By *related attraction* of an accommodation we mean an attraction that in some way is connected to that accommodation. For example, if the *Hotel Bologna* in Pisa is located near the *Leaning Tower of* Pisa, a possible relation can be established between the two places. Relations among accommodations and attractions could foster and promote new research in the services industry. In fact, many studies exist in the literature, showing how hotels choose their location in a city [18] and how tourists attractions change according to the position of the hotel [12, 14]. A resource like Tourpedia could foster and facilitate these kinds of studies, by providing specific relations between accommodations and attractions.

[1] http://tour-pedia.org

[2] http://www.facebook.com
[3] http://www.foursquare.com
[4] http://www.google.com/business/
[5] http://www.booking.com

This paper gives an overview of the Tourpedia knowledge base and then describes how the links between accommodations and attractions are built. The other aspects of Tourpedia, i.e. reviews analysis and features descriptions, will be described in a future work. The rest of the paper is organized as follows. Section 2 reviews related work; Section 3 gives an overview of Tourpedia and Section 4 explains how new relations are discovered among accommodations and attractions. Section 5 describes the evaluation process for relations discovery. Finally, Section 6 contains the conclusions and future work.

2. RELATED WORK

Over the last years, many initiatives have been proposed in the field of tourism datasets. A complete list can be found in the DataHub.io Web site[6]. Here we give an overview only of the most important datasets related to the tourism domain. GeoNames[7] is a very popular geographical dataset with over 10 million geographical names. It contains also tourism places, such as hotels and restaurants. However, it does not provide any SPARQL endpoint to query data and, for each node, it gives only little information. For example, for the *Hotel Bologna* in Pisa, it provides only its geographical coordinates[8]. LinkedGeoData[9] [15] is an open initiative that exports information extracted from OpenStreetMap[10] as a RDF knowledge base. LinkedGeoData contains more than 1 billion nodes. Since it is derived from a collaborative platform, it is continuously updated. With respect to LinkedGeoData, we propose a smaller knowledge base, which contains more detailed information for each node. For example, the *Hotel Bologna* in Pisa contains more information in Tourpedia[11] than in LinkedGeoData[12]. In any case, it would be very interesting to map the two knowledge bases through sameAs links. We are planning to do it as future work, by adopting a flexible Linked Data Mashup view [16]. Other initiatives with minor impact are The Santillana Guide dataset[13], the Salzburgerland Tourismus dataset[14] and the Accommodations in Tuscany and Piedmont datasets, whose links are no longer available. To the best of our knowledge, Tourpedia is the only tourism dataset which provides sentiments about places and relations between accommodations and attractions.

3. TOURPEDIA

The Tourpedia knowledge base currently consists of 6.037.889 RDF triples and describes 492.888 entities, divided in four classes, as shown in Table 1. Tourpedia contains almost 500.000 links to DBpedia[15]. This is defined through the property `dbpedia-owl:location`, associated to every place. Other useful statistics about external links are contained in the Tourpedia VOID description[16]. The Tourpedia knowledge base is accessible over the Web through different mechanisms: a) SPARQL endpoint[17], b) direct access through the resource identifier (e.g. *Hotel Annalena*[18]), c) RDF dump[19], d) Web API[20], e) Web interface[21] [5]. Tourpedia is also registered to the DataHub.io platform[22] and provides a place for community discussion[23].

Class	Instances
Accommodation	32.554
Restaurant	158.325
Attraction	70.808
POI	231.193

Table 1: Tourpedia classes with the number of their instances.

3.1 Ontology

The Tourpedia knowledge base is represented by a unique ontology[24] (Table 2), which contains classes and properties derived information extracted from each social media. The prefix *marl:* refers to the MARL ontology[25], while the prefix *acco:* refers to the Acco ontology [9] and *h:* to the Hontology ontology [4], both related to the tourism domain. We defined some equivalences of classes with the Schema ontology[26]. We did not use schema.org directly as ontology, because our data model is conceptually different from that defined in schema.org, where the levels of hierarchy are more complicated. In Tourpedia concepts are simpler and subclasses are at the same level of hierarchy. Currently, Tourpedia contains only generic classes. However we extracted also subcategories from social media, e.g. type of accommodation. We are planning to integrate this aspect, although the mapping is not simple, because of the different classification mechanisms employed by different social media. More details about accommodation data modeling can be found in [2]. It is interesting to note how the Tourpedia ontology describes the sentiment about a place: each place has a property `marl:hasOpinion`, which is connected to a resource of class `marl:AggregatedOpinion`. The polarity value associated to each aggregated opinion ranges from 0 to 10.

3.2 License and Maintenance

Tourpedia is served on the Web under the terms of the Creative and Commons CC0 1.0 License. We investigated carefully the issues related to the publication of data extracted from social media and we published only objective properties associated to each place (such as name, address, latitude, longitude and so on). Instead, sensible data produced

[6]http://datahub.io/dataset?q=tourism
[7]http://www.geonames.org
[8]http://www.geonames.org/6501002/hotel-bologna.html
[9]http://linkedgeodata.org/About
[10]https://www.openstreetmap.org/
[11]http://tour-pedia.org/page/acco-204042
[12]http://linkedgeodata.org/page/triplify/node2617740233
[13]http://webenemasuno.linkeddata.es
[14]http://data.salzburgerland.com/organization/salzburgerland-tourismus
[15]http://dbpedia.org
[16]http://tour-pedia.org/download/void.ttl
[17]http://tour-pedia.org/sparql
[18]http://tour-pedia.org/resource/acco-207756
[19]http://tour-pedia.org/download/tourpedia.rdf
[20]http://tour-pedia.org/api
[21]http://tour-pedia.org/gui/demo/
[22]http://datahub.io/dataset/tourpedia
[23]https://groups.google.com/forum/?tourpedia#!forum/tourpedia
[24]http://tour-pedia.org/download/tp.owl
[25]http://www.gsi.dit.upm.es/ontologies/marl/
[26]http://schema.org

tp:Place
+ vcard:fn + dbpedia-owl:address + vcard:hasTelephone + vcard:hasPhoto + wgs84_pos:lat + wgs84_pos:long + dbpedia-owl:location + dbpedia-owl:wikiPageExternalLink + marl:hasOpinion
tp:Accommodation *extends* tp:Place
+ acco:feature + h:InternalFeature + **skos:related**
tp:Restaurant *extends* tp:Place
tp:POI *extends* tp:Place
tp:Attraction *extends* tp:Place
+ **owl:sameAs**
marl:AggregatedOpinion
+ marl:polarityValue

Table 2: Tourpedia ontology.

by the social media (such as the content of a review) cannot be published by Tourpedia. In this case we provide the link to original resource on the social media. Beside this, in Tourpedia we publish only integrated data and not original data, as they were extracted from the social media. For this reason, in Tourpedia data provenance of each attribute is not associated to a specific social media source. Tourpedia maintains data provenance through the property `wikiPageExternalLink` accessible via SPARQL and through direct access to each resource[27]. In the Web search interface data provenance is preserved through a marker on the map associated to each place. A click on the marker opens an infobox, containing the direct links to social media.

The maintenance of Tourpedia is under deployment. It implements the incremental maintenance of Linked Data Mashup views described in [16]. Periodically, for each social media Tourpedia computes the incremental changes in terms of: INSERT, DELETE and UPDATE records and propagates them to the final mashup view, by considering also the preservation/change of same as links.

4. RELATIONS DISCOVERY AND LINKING

In this paper we describe how we discovered relations among accommodations and attractions, represented through the property `skos:related`. Besides this, we illustrate how we built the `owl:sameAs` links between attractions and other linked data sources. The description of how we extracted the sentiment will be further illustrated in a future work.

4.1 Relations discovery

In order to discover new relations, we implemented the Entity Discovery System (EDS), a tool which extracts all the Named Entities (NEs) from accommodations'descriptions (extracted from Booking.com). This is motivated by the fact that almost all descriptions contain a great number of NEs, i.e. attractions, located near the accommodation

[27]http://tour-pedia.org/page/acco-204042

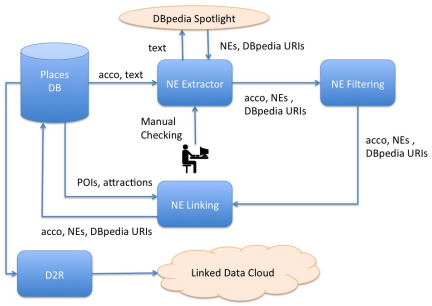

Figure 1: The architecture of EDS.

which the description refers to. Therefore, EDS builds a relation between each extracted NE and the accommodation. In details, EDS exploits the DBpedia Spotlight Service[28], a Named Entity Recognition and Disambiguation (NERD) tool, which provides also the links of disambiguated NEs to DBpedia. Other NERD services exist, both commercial and open source, such as OpenCalais[29], TagMe[30] and OpenNLP[31]. We chose DBpedia Spotlight for EDS because of the highest quality of extracted NEs. Due to space limitation, we do not illustrate the results of comparison among the different NERD tools. Figure 1 shows the architecture of the EDS: it is composed of three main modules: the *NE Extractor*, the *NE Filtering* and the *NE Linking*. The NE Extractor takes accommodations as well as their descriptions from the Places Database (DB) of Tourpedia and extracts NEs and their associated URIs from the DBpedia Spotlight service. In order to validate the results, a manual check is done. This phase consists in correcting wrong links to DBpedia and performing manual coreference, such as disambiguating *Central Station* in *Pisa Central Station*. Obviously, for a large amount of data this is not feasible. However, a manual check can be always done on a subset of data in order to verify the quality of extracted NEs. This process is very awkward because it is strongly related to the quality of established relations. We are aware that this process should be done automatically or semi-automatically. We are planning to move to this direction as future work, by also taking into account the difficulties related to it. The NE Filtering module deletes all the non-geographical NEs by applying a filter based on geographical coordinates. This module asks DBpedia for geographical coordinates. If they do not exist, the NE is not considered a place thus it is discarded. Otherwise, it is added to the list of discovered NEs. Once all the NEs have been extracted and corrected, a new relation is built between each pair (accommodation, discovered NE). The relation between the two records is defined through the property `skos:related`.

4.2 Linking

The NE Linking module links the discovered NEs to the places contained in the Places DB through the `owl:sameAs` property. This is done through a matching algorithm, which was implemented as follows. Let (i) p a place in the Places

[28]https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki
[29]http://new.opencalais.com
[30]http://tagme.di.unipi.it
[31]https://opennlp.apache.org

NEs	Nr. of entities	%
Entities Extracted	9.814	-
Entities Filtered	1.064	10,8 %
Entities Linked	809	76,03 %
Entities Added	255	23,97 %

Table 3: The output of EDS.

DB and q a discovered NE; (ii) $T_p = \{t_{p_1}, ..., t_{p_m}\}$ and $T_q = \{t_{q_1}, ..., t_{q_n}\}$ the set of tokens in the names of p and q, respectively; (iii) $G_p = (x_p, y_p)$ and $G_q = (x_q, y_q)$ the geographical coordinates of p and q, respectively. Intuitively, p and q are matching candidates if they are geographically close and their names are quite similar. The matching function computes the following steps: (i) calculate *distance(G_p, G_q)*, as the great circle distance [7] between G_p and G_q (ii) *iff distance(G_p, G_q)* $\leq \theta$, calculate

$$sim(T_p, T_q) = 2\frac{\sum_{i=1}^{\min(m,n)} \sum_{t1 \in T_p} \sum_{t2 \in T_q} w_i(t1, t2)}{m + n} \quad (1)$$

We used $\theta = 200m$. The function $w_i(.)$ represents the Levenshtein distance [10] between two tokens in T_p and T_q. (iii) p and q are considered matching candidates *iff* $\text{sim}(T_p, T_q) \geq \alpha$. We set $\alpha = 0.8$. If $\text{sim}(T_p, T_q) \geq \alpha$ holds, then a new `owl:sameAs` link is built between p and q, otherwise q is added to the Places DB as a new place with category *attraction* (e.g. *Pitti Palace* in Florence[32]).

5. EVALUATION OF EDS

EDS was applied on 10.338 descriptions, associated to just as many accommodations, derived from Booking.com. Table 3 illustrates the output of EDS. The NE Extractor module extracted 9.814 NEs, of which only 1.064 corresponded to geographical NEs (10,8%). The 76,03 % of filtered NEs were already available in the Tourpedia knowledge base, and the remaining 23,97% were added to Tourpedia as attractions. Besides this, EDS discovered 13.294 relations among accommodations and places, thus adding the same number of RDF triples to the Tourpedia knowledge base. EDS added 1,29 relations per accommodation on average.

6. CONCLUSION AND FUTURE WORK

In this paper we have given an overview of the Tourpedia knowledge base, which contains tourism places extracted from social media. Besides this, we have illustrated how new relations among accommodations and attractions were built in Tourpedia. The described procedures could be generalized and adopted also by other knowledge bases. For example the abstracts of places contained in DBpedia could be processed through EDS in order to discover new relations among entities. As future work, we are planning to improve EDS by studying and implementing a strategy to reduce the human participation during the NE extraction phase and to release EDS as a public framework.

7. REFERENCES

[1] C. K. Anderson. The impact of social media on lodging performance. *Cornell Hospitality Report*, 12(15), 2012.

[2] C. Bacciu, A. Lo Duca, A. Marchetti, and M. Tesconi. Accommodations in Tuscany as Linked Data. In *LREC 2014*, pages 3542–3545, May, 26-31 2014.

[3] T. Berners-Lee. Linked Data - Design Issues, July 2006.

[4] M. S. Chaves, L. A. de Freitas, and R. Vieira. Hontology: A multilingual ontology for the accommodation sector in the tourism industry. In J. Filipe and J. L. G. Dietz, editors, *KEOD*, pages 149–154. SciTePress, 2012.

[5] S. Cresci, A. D'Errico, D. Gazzé, A. Lo Duca, A. Marchetti, and M. Tesconi. Tourpedia: a web application for sentiment visualization in tourism domain. In *The OpeNER Workshop in LREC 2014*, pages 18–21, 2014.

[6] S. Cresci, A. D'Errico, D. Gazzè, A. Lo Duca, A. Marchetti, and M. Tesconi. Towards a DBpedia of Tourism: the case of Tourpedia. In *ISWC 2014 (Poster and Demo Track)*, ISWC2014, pages 129–132, 2014.

[7] K. Gade. A non-singular horizontal position representation. *Journal of Navigation*, 63:395–417, 7 2010.

[8] J. Heidemann, M. Klier, and F. Probst. Online social networks: A survey of a global phenomenon. *Computer Networks*, 56(18):3866 – 3878, 2012.

[9] M. Hepp. Accommodation ontology language reference. Technical report, Hepp Research GmbH, Innsbruck, 2013.

[10] V. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, 1966.

[11] A. Marzuki, A. A. Razak, and T. L. Chin. *What Women Want: Hotel Characteristics Preferences of Women Travellers*. INTECH Open Access Publisher, 2012.

[12] J. L. Nicolau and F. J. Más. The influence of distance and prices on the choice of tourist destinations: The moderating role of motivations. *Tourism Management*, 27(5):982 – 996, 2006.

[13] G. Sammons, P. Moreo, L. F. Benson, and F. Demicco. Analysis of female business travelers' selection of lodging accommodations. *Jour. of Travel & Tourism Marketing*, 8(1):65–83, 1999.

[14] N. Shoval, B. McKercher, E. Ng, and A. Birenboim. Hotel location and tourist activity in cities. *Annals of Tourism Research*, 38(4):1594 – 1612, 2011.

[15] C. Stadler, J. Lehmann, K. Höffner, and S. Auer. Linkedgeodata: A core for a web of spatial open data. *Semantic Web Journal*, 3(4):333–354, 2012.

[16] V. M. P. Vidal, M. A. Casanova, and D. S. Cardoso. Incremental Maintenance of RDF Views of Relational Data. In *OTM 2013*, pages 572–587, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[17] H. Werthner, A. Alzua-Sorzabal, L. Cantoni, A. Dickinger, U. Gretzel, D. Jannach, J. Neidhardt, B. Pröll, F. Ricci, M. Scaglione, et al. Future research issues in it and tourism. *Information Technology & Tourism*, 15(1):1–15, 2015.

[18] Y. Yang, K. K. Wong, and T. Wang. How do hotels choose their location? evidence from hotels in beijing. *Int. Jour. of Hospitality Management*, 31(3):675 – 685, 2012.

[32] http://tour-pedia.org/resource/attr-198013