A Method for Querying Touristic Information Extracted from the Web

Ermelinda Oro(✉) and Massimo Ruffolo

National Research Council (CNR), Via P. Bucci 7/11C, 87036 Rende, CS, Italy
{linda.oro,massimo.ruffolo}@icar.cnr.it

Abstract. In this paper, we define a method that enables to: (i) exploit ontologies and terminology to represent concepts into specific domains, (ii) extract information from deep web pages, and opinions from social and community networks, (iii) semantically query information by using natural language interface. The method has been applied to query the touristic domain. This use case has shown the effectiveness of the method that can be easily extended to other domains.

Keywords: Natural language interface · Web data extraction · Ontology · Linked data · Question answering · Tourism domain

1 Introduction

Tourism is one of the most dynamic and web-based industries worldwide. It is an information-intensive market where tourists can use and leverage a plethora of sources having heterogeneous formats. To plan and organize touristic travels tourists navigate e-commerce web sites and social networks for accommodation and reservation, forums and blogs, holiday resort and points of interest web sites, comments on topics related to tourism. But searching and querying such a Big Data by standard search engines may result unsatisfactory, especially on mobile systems, whereas, by using formal languages results too complex for end-users. In this scenario, the usage of natural language interface for querying the web and existing knowledge bases offers opportunity to bridge the technological gap between end-users and systems that make use of formal query languages.

Several researchers and practitioners are working on the definition of algorithms and systems capable to translate natural language questions into formal languages able to query structured data. Question Answering systems (QAs) have attracted extensive attentions in both NLP [4,5] and database communities [6,7]. Different surveys are presented in literature [1,2]. But only few approaches worked on the tourism domain and defined a complete approach from the knowledge representation, data acquisition and information extraction, to the natural language processing and question answering.

In this paper, we present a method that enables to query, in natural language, knowledge bases populated by touristic information extracted from the web. This use case has shown the effectiveness of the proposed method that can be easily extended to other domains.

F. Frasincar et al. (Eds.): NLDB 2017, LNCS 10260, pp. 218–221, 2017.
DOI: 10.1007/978-3-319-59569-6_25

2 Proposed Method

This section describes a method for querying in natural language knowledge bases populated by using information extracted from the web. The proposed method is tailored to be flexible, scalable and general. It is constituted by three phases: (i) Definition of the ontology and terminology representing concepts into the specific domain. (ii) Extraction of information from deep web pages, from social and community networks. (iii) Semantic querying by using natural language processing, ontologies, and the Semantic Web tools. Figure 1 shows modules of the proposed method.

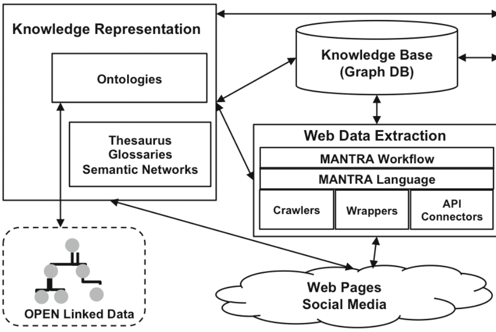


Fig. 1. Architecture of the proposed method

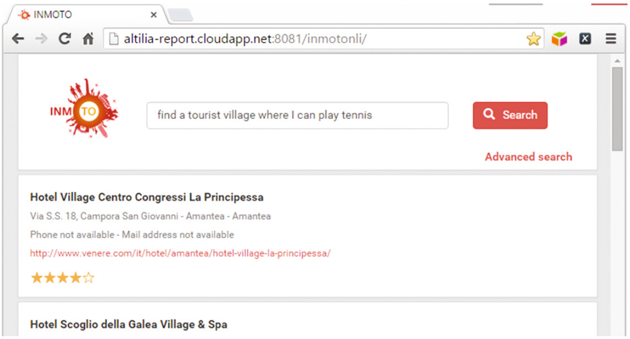
Knowledge Representation. The knowledge of the target domain can include ontologies, as well as thesauri, dictionaries and semantic networks. In our tourism domain, we created all such instruments by observing information existing on the Web and by including concepts from existing open linked data and by using top-level ontologies. The implemented ontology describes classes and properties of places (Places), objects and events (e.g. Facility, PointOfInterest, and Event). It includes extracted and computed reputation information (SocialObject). It maintains data about sources where the information has been extracted in order to be able to compute the reliability of the obtained information.

Web Data Extraction. Instances of the knowledge base are dynamically created by monitoring and extracting semi- and unstructured information (such as data, descriptions and comments about accommodations) by connecting to APIs of social networks, or by implementing smart crawlers and wrappers of deep web pages. To extract data from booking websites (e.g. booking.com, venere.it, and tripAdvisor), we modeled wrappers by using a visual interface[1] that records navigation actions performed by users into websites. In order to process comments and analyze opinions we implemented the sentiment analyzer writing rules in the MANTRA Language [3]. The MANTRA Language represents grammar-based

[1] MANTRA Web Extractor (MWE) http://mwe.altiliagroup.com/.

programs combined with logic predicates to identify concepts and their relations belonging to specific knowledge domains, and to compute sentiment polarities and opinion-targets. We can monitor websites and social networks scheduling workflows that populate the knowledge base (e.g. a graph database).

Natural Language Querying. We created a Natural Language Interface (NLI) that allows users to query knowledge bases through natural language expressions that are dynamically translated into formal queries (e.g. SPARQL and Cypher Query) to exploit various knowledge bases (i.e. ontologies and graph databases). Users write questions expressed in natural language by using a web-based GUI, shown in Fig. 2. The query in natural language is preprocessed to recognize linguistic constructs (such as, pos-tag, chunk, relations between chunk). The MANTRA Language Module takes as input MANTRA Language Programs to identify concepts, properties, and relationships that occur in the users questions. Patterns expressed in MANTRA Language allow for recognizing in a bottom-up fashion ontological information by exploiting natural language constructs, dictionaries, semantic networks, and thesauri. To match recognized concepts extracted from the natural language text to the knowledge base concepts, we use an external resource locator that maps ontological concepts to specific URI. The submodule Formal Query Builder creates a query in SPARQL or Cypher Language by using recognized objects, properties, and relationships. The formal query is then submitted to the knowledge base endpoint and results are visualized to users.



find a tourist village where I can play tennis

Hotel Village Centro Congressi La Principessa
Via S.S. 18, Campora San Giovanni - Amantea - Amantea
Phone not available - Mail address not available
http://www.venere.com/it/hotel/amantea/hotel-village-la-principessa/
★★★★☆

Hotel Scoglio della Galea Village & Spa

Fig. 2. Web-based graphical user interface for natural language querying

3 Discussion and Conclusion

In this paper, we presented a modular method that enables to query information extracted from the web. The main contribution of this paper consists in the definition of a proposed method that enables to: (i) exploit ontologies and terminology to represent concepts of specific domains; (ii) extract information and opinions from deep web pages, social and community networks; (iii) semantically query information by using a natural language interface. The modular method enables to easily create specific domain solutions. We tested our method for querying tourism information. More in detail, we focused on extracting and querying both descriptive data and opinions about accommodations. In particular, we collected descriptive data from booking websites, user ratings and comments from both booking websites and social networks. In our test, we extracted more than 3000 Italian hotels offering more than 20 different types of services from booking.com and venere.it. In addition, we monitored more than 5000 comments from around 1600 users. We collected 100 different questions written by a group of heterogeneous persons in Italian or English natural language. Questions adopted in the test considered services, nearby points of interest and opinions about the hotel. Example of questions are: "I would like to eat in a very good restaurant of an hotel in Cosenza" "I'm looking for a 4 star hotel in Amantea with free parking, swimming pool and elevator. Furthermore, I'd like to play tennis and read some books in a library." We compared obtained results with the computed Ground Truth obtaining precision 1. This approach works well for domain specific knowledge, like the considered touristic domain. Experimental applications, involving real user questions on touristic domain, demonstrate that our system provides high-quality results. In the future, we intend to evaluate our approach on a larger scale, and we plan to conduct an intensive usability study. In addition, our goal is to provide robust question answering interface that exploits some innovative algorithms based on deep neural networks.

References

1. Dwivedi, S.K., Singh, V.: Research and reviews in question answering system. Procedia Technol. **10**, 417–424 (2013)
2. Mishra, A., Jain, S.K.: A survey on question answering systems with classification. J. King Saud Univ. Comput. Inf. Sci. **28**(3), 345–361 (2016)
3. Oro, E., Ruffolo, M.: Using apps and rules in contextual workflows to semantically extract data from documents. In: Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services (2015)
4. Unger, C., Bühmann, L., Lehmann, J., Ngonga Ngomo, A.C., Gerber, D., Cimiano, P.: Template-based question answering over RDF data. In: Proceedings of the 21st International Conference on World Wide Web, pp. 639–648. ACM (2012)
5. Yahya, M., Berberich, K., Elbassuoni, S., Ramanath, M., Tresp, V., Weikum, G.: Natural language questions for the web of data. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 379–390. Association for Computational Linguistics (2012)
6. Yahya, M., Berberich, K., Elbassuoni, S., Weikum, G.: Robust question answering over the web of linked data. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 1107–1116. ACM (2013)
7. Zou, L., Huang, R., Wang, H., Yu, J.X., He, W., Zhao, D.: Natural language question answering over RDF: a graph data driven approach. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, pp. 313–324. ACM (2014)