Semantic Search in RealFoodTrade

Andrea Calì[1,4], Roberto De Virgilio[2], Tommaso Di Noia[3], Luca Menichetti[2], Roberto Mirizzi[5], Luca Nardini[2], Vito Claudio Ostuni[3], Fabrizio Rebecca[2], and Marco Ungania[2]

[1]Dept. of Computer Science and Inf. Systems, Birkbeck University of London, UK
[2]Dip. di Ingegneria, Università Roma Tre, Italy
[3]Dept of Electrical and Electronic Engineering, Politecnico di Bari, Italy
[4]Oxford-Man Institute of Quantitative Finance, University of Oxford, UK
[5]Hewlett-Packard Laboratories, USA

andrea@dcs.bbk.ac.uk
dvr@dia.uniroma3.it
t.dinoia@poliba.it, ostuni@deelab.poliba.it
{meniluca,lucamarionardini, fabri.rebecca,mar.ungania}@gmail.com
roberto.mirizzi@hp.com

Abstract

We present RealFoodTrade (RFT), a system that allows farmers and fishermen to sell their products directly to the end-buyer. RFT makes use of Linked Data sets, together with a domain ontology designed by experts, to perform semantic search over products on sale. RFT employs geo-location technology on mobile devices to match demand and supply according to the location. We sketch the semantic search techniques in RFT and illustrate a prototype tailored to the fishing industry.

1 Aim and Scope

In this paper we introduce the system *Real Food Trade* (RFT), which provides a marketplace for any food producer to sell their produce directly, freeing the producer both from the wholesaler and from the effort of retail. The main idea is that the producer uses the system to advertise *flash stands*, each of which consists of *(1)* a set of products, with their quantities, and *(2)* a delivery time and location. From the economic point of view, this is important as several farmers and fishermen do not have the possibility of advertising their produce, nor to reach the end buyer in a consistent way (e.g. by means of a shop or door-to-door delivery). The economic aspects of RFT will be presented elsewhere; here we concentrate on the issues regarding the search of products by potential buyers.

In RFT, buyers search for flash stands selling a certain product within a certain area. The wanted product is specified by keywords, and the system identifies the stands that sell products which are *semantically close* to the input keywords. To do this, we aim at employing novel techniques for keyword search, employing

as data sets both the ones available in the *Linked Open Data* cloud[1] and "hand-crafted" ontologies. We intend to empower the ontologies of the latter type with the less-polished, but richer, information available at Linked Data stores.

Related work. Several techniques have been proposed to automatically calculate the semantic relatedness between words, texts or concepts in a way corresponding closely to that of human subjects. Some of the commonly used methods derive statistical information from text corpora and combine that information with lexical sources [8,2]. The lack of domain-specific coverage of the resources used by these measures makes them ineffective for use in domain specific tasks where the context plays an important role. Most classical methods to compute semantic measures exploit particular lexical sources: corpus, dictionaries, or well structured taxonomies such as WordNet [7]. Some of these methods explore path lengths among nodes in taxonomies [10]; others exploit textual descriptions of concepts in dictionaries [8], while a last group relies on annotated corpora to compute information content [3,9]. Some recent research efforts has focused on using Wikipedia to improve coverage with respect to traditional thesauri-based methods [11,5]. In particular, the work in [5] proposes a method to represent the meaning of texts or words as weighted vectors of Wikipedia-based concepts, using machine learning techniques. Wikipedia seems to be unable to effectively discover and evaluate implicit relationships [6]. In order to guarantee maximum coverage, the work in [6] focuses on methods that exploit the Web as source of knowledge. In [1] the authors propose a similarity measure that combines various similarity scores based on page counts, with another one based on lexico-syntactic patterns extracted from text snippets.

In the rest of the paper we will describe the Real Food Trade system, and how it deals (and plans to deal) with the problem of keyword search on ontologies represented as Linked Data.

2 Overview

In this section we sketch how RFT works and overview its architecture, as shown in Fig. 1.

The system exploits the Google data cloud driven architecture[2]. In particular the essential components in the proposed mobile solution architecture are:

- Android mobile clients;
- Google Cloud Endpoints used for communications between the clients and the backend over REST API with OAuth2 authentication;
- A backend application code running on Google App Engine[3] and responsible for serving requests from the clients.

A typical requirement for a mobile solution with a backend is to store data outside of client devices. These data can be categorized into two groups: *(i)*

[1] http://lod-cloud.net/
[2] https://cloud.google.com/
[3] https://developers.google.com/appengine/

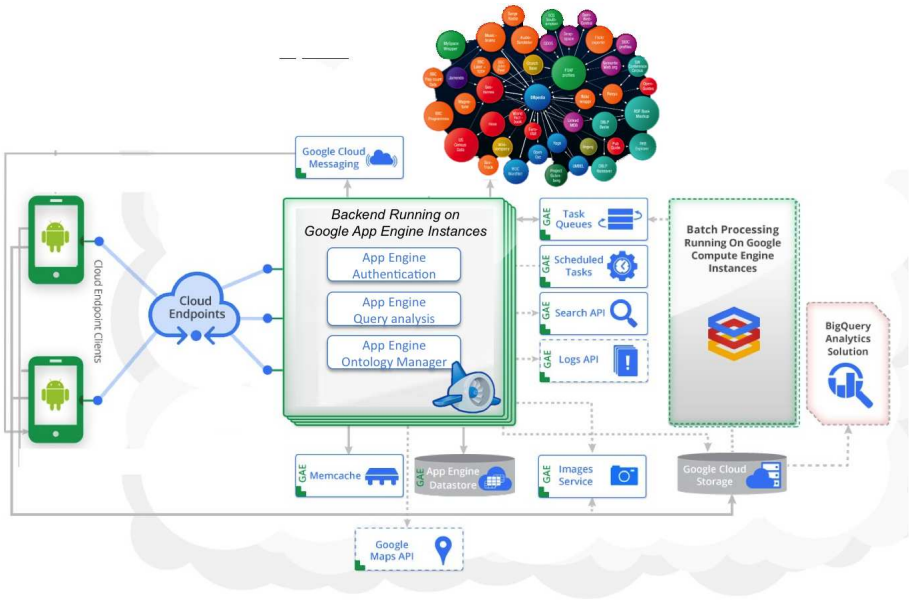

Fig. 1. An architecture of reference

large, and typically binary objects (e.g. images) and *(ii)* fine-grained properties and entities, which can also include a reference, for example, object name and optionally bucket name or URL, to objects stored in Google Cloud Storage.

Both groups of data are managed by two modules in RFT: *(i)* App Engine Authentication and *(ii)* the App Engine Query Analysis. The first application is responsible to store and manage all user profiles. In particular we embedded OpenID[4] technologies to exploit existing email or social network accounting (e.g. Gmail, Yahoo, Facebook, twitter and so on). Once the user is authenticated, the App Engine Query Analysis is responsible to analyze the user request and to invoke several instances of the application to retrieve all the data best fitting the user query. To this aim, the App Engine Ontology Manager is responsible to implement semantic matchmaking techniques exploiting the Linked Open Data cloud.

In particular, RFT employ different RDF taxonomies that characterize the domains of interest. Such taxonomies are indexed in the Google cloud and linked to the LOD network (i.e. DBPedia). Then by using different heuristics (i.e. user rating, usage statistics, and lexicographic analysis) we rank the similarities between concepts in the taxonomies. In this way RFT can exploit a weighted similarity search based on the rank. Intuitively depending on how many products the user would view, the similarity distance from a concept matching the

[4] http://openid.net/

required product can increase or decrease. In the next section we will discuss a case study for our system.

Finally, a natural place to store this kind of data is App Engine Datastore. It provides a NoSQL schemaless object data store, with a query engine and atomic transactions. These entities will often map to the Resources exposed over Google Cloud Endpoints API.

3 Case Study

As case study for RFT, we analyzed Fish & Seafood Markets in Chile. In particular we have two specific users: the producer (i.e. the fisherman) and the buyer (private or business). The producer is interested to spot the available produce by exploiting e-commerce technology (e.g. eBay model); in this case RFT provides real time points of interest (POIs) through Google Maps representing market announcements for all people interested in buying fish. Such POIs are invoked by producers and generated through RFT.

For instance, Fig. 2 shows our system at work: *(i)* the user compiles a keyword search query, *(ii)* the system retrieves all POIs on the map and *(iii)* the user selects and reserves items.

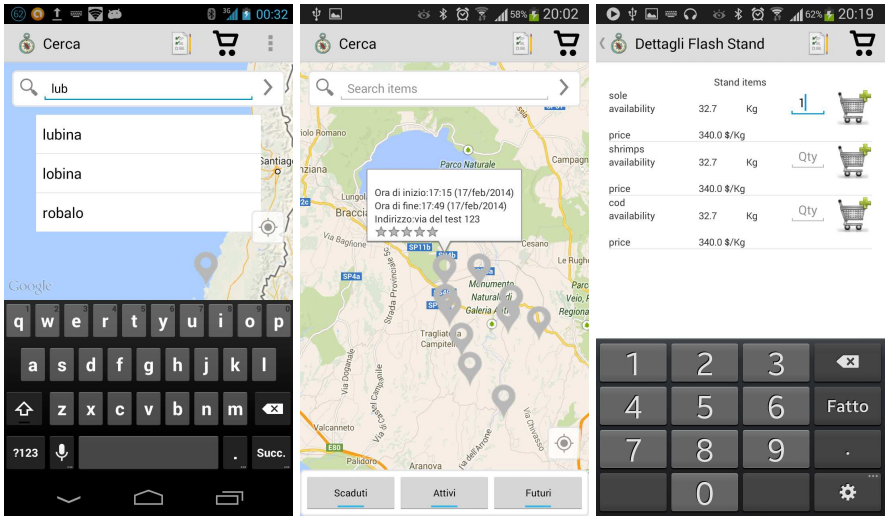

Fig. 2. RFT at work.

In this scenario the keyword search query processing involves different semantic tasks. To this aim, we adopted the Network of Fisheries Ontology[5] provided

[5] http://aims.fao.org/network-fisheries-ontologies

by the Food and Agriculture Organization of the United Nations (FAO). Such ontology is a fine-grain classification of species: the classification is made by both biological nature and regional proximity. Whenever a seller (fisherman) enters a product, typing its name, a concept (corresponding to a particular type of fish) in the ontology is associated to the product; such a node is the semantically closest to the entered one. The association between the input and the concept in the ontology is established in a fashion similar to that of [4], where ontological information extracted off-line from Linked Data sets is also employed. Similarly, whenever a buyer searches for a particular kind of fish, the semantically closest one in the ontology is returned, as well as other ones within the same species (alternative names for the same fish, or sub-species). The match is also performed during the input (when only a partial input is entered), so that the user can receive suggestions from the interface in order to save input time — this is especially useful when entering input on mobile devices.

4 Discussion

We have sketched the main features of the RFT system, which provides a location-based infrastructure for the sale of food. After having briefly illustrated the architecture of the system and the technologies employed in it, we have presented a case study, where we employ RFT to fisheries in Chile, in particular small fishing boats. We have shown how we integrate the FAO's Network of Fisheries Ontology with linked data sources in order to match entities in the ontology with fish names entered by users (fishermen or buyers).

We plan to extend the field of application of RFT to other markets, in particular that of agriculture. This will require the automated extraction of high-quality ontologies from Linked Data sets. Moreover, we intend to incorporate learning algorithms that incorporate knowledge from users' behaviour in order to enhance the knowledge base used in the system. Last but not least, we intend to carry out extensive experiments on real fish markets and report on the results, both from the system point of view and from the socio-economic point of view.

Acknowledgments. Andrea Calì acknowledges support by the EPSRC project "Logic-based Integration and Querying of Unindexed Data" (EP/E010865/1).

References

1. Bollegala, D., Matsuo, Yutaka, M.: Measuring semantic similarity between words using web search engines. In: Proceedings of the 16th international conference on World Wide Web. pp. 757–766. WWW '07, ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1242572.1242675
2. Budanitsky, A., Hirst, G.: Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In: In Workshop on WordNet and other lexical resources, second meeting of the North American chapter of the association for computational linguistics (2001)

3. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Comput. Linguist. 32(1), 13–47 (Mar 2006), http://dx.doi.org/10.1162/coli.2006.32.1.13
4. Calì, A., Capuzzi, S., Dimartino, M.M., Frosini, R.: Recommendation of text tags in social applications using linked data. In: 2nd Int. Workshop on Data Management in the Social Semantic Web (DMSSW 2013). pp. 187–191 (2013)
5. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th international joint conference on Artifical intelligence. pp. 1606–1611. IJCAI'07, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2007), http://dl.acm.org/citation.cfm?id=1625275.1625535
6. Gracia, J., Mena, E.: Web-based measure of semantic relatedness. In: In Proc. of 9th International Conference on Web Information Systems Engineering (WISE 2008), Auckland (New Zealand. pp. 136–150. Springer (2008)
7. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM 38(11), 39–41 (Nov 1995), http://doi.acm.org/10.1145/219717.219748
8. Patwardhan, S., Banerjee, S., Pedersen, T.: Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the 4th international conference on Computational linguistics and intelligent text processing. pp. 241–257. CICLing'03, Springer-Verlag, Berlin, Heidelberg (2003), http://dl.acm.org/citation.cfm?id=1791562.1791592
9. Pedersen, T., Banerjee, S., Patwardhan, S.: Maximizing Semantic Relatedness to Perform Word Sense Disambiguation. Research Report UMSI 2005/25, University of Minnesota Supercomputing Institute (March 2005), http://www.patwardhans.net/papers/PedersenBP05.pdf
10. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics 19(1), 17–30 (1989)
11. Strube, M., Ponzetto, S.P.: Wikirelate! computing semantic relatedness using wikipedia. In: proceedings of the 21st national conference on Artificial intelligence - Volume 2. pp. 1419–1424. AAAI'06, AAAI Press (2006), http://dl.acm.org/citation.cfm?id=1597348.1597414