Semantic Keyword Search in Linked Data

Andrea Calì[1,3], Leonardo Coaccioli[2], Mirko Michele Dimartino[1]
, Riccardo Frosini[1], and Federico Pastori[2]

[1]Dept. of Computer Science and Inf. Systems, Birkbeck University of London, UK
[2]Dip. di Ingegneria, Università Roma Tre, Italy
[3]Oxford-Man Institute of Quantitative Finance, University of Oxford, UK

andrea@dcs.bbk.ac.uk
{dimartinomirko, fedepast89, frosini.riccardo,
leo.coaccioli}@gmail.com

1 Motivation and scope

Linked Data sets available on the web constitute now a large corpus of (semi)structured information that is generally highly valuable. Though data in Linked Data sets are represented in in a flat data structure (usually RDF), from the semantic point of view there are several levels of abstractions: for instance, we can find information about classes of objects and inclusion (or subset, or is-a) relationship between classes. We therefore argue that *semantic search* is particularly suited to Linked Data sets.

In this short paper we outline our approach to semantic search in Linked Data. In particular, we focus on keyword search; the goal is to return results (in the form of concepts) that are semantically close to the keyword(s) entered by the user. We have applied our techniques to the system *Real Food Trade* (RFT), which provides a marketplace for producers to sell their produce directly to the end-buyer (but not necessarily). In RFT, buyers search for 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 employ novel techniques for keyword search, using as data (and meta-data) sets both the ones available in the *Linked Open Data* cloud[1] and "hand-crafted" ontologies built by experts. In the case study we are currently working on, we focus on the fish market.

Common methods for computing semantic similarity make use of text corpora combined with lexical sources [5]. Several works also make use of more structured taxonomies like WordNet [3], for instance exploring paths between words in the Wordnet graph. Other works gather information from the web: for instance [2], which uses Machine Learning techniques on vectors of concepts based on Wikipedia.

2 Semantic Search

Our approach is similar to the one of [4], which adopts technique of Information Retrieval for building a recommender system based on semantic information extracted

[1] http://lod-cloud.net/

from DBpedia, LinkedMDB and Freebase. Our marketplace is currently focussed on fish, and we rely on information extracted from two sources: *(1)* The FAO Network of Fisheries Ontology [1], which provides a taxonomy (in English only, with scientific names in Latin) of relevant varieties of fish, and *(2)* the DBpedia data set, from where we extract, through a suitable navigation, information on fish to be stored offline. To compute a similarity measure between fish varieties, we consider the attributes `class`, `family` and `species` and we adopt a Vector Space Model (VSM), associating to each fish variety a vector of the aforementioned attributes. The similarity value between two varieties is computed, similarly to what is done in [4], by computing the cosine of the angle between the two associated vectors. Notice that vectors contain *both* the rigid information of the FAO ontology and the less structured information from DBpedia.

Together with the semantic similarity, we also use *syntactic* similarity, with a Levenshtein distance, between keywords entered by the user and the names of fish varieties in the system. This is also used to suggest entries, in the mobile version of the system, when the user types the first characters. Roughly speaking, when the user enters the keywords, a set of fish varieties that syntactically matches the keywords is determined, and then varieties that are semantically close to them are returned in order of similarity. The system is then able to select the set of stands that sell semantically similar varities to the one entered by the user.

3 Discussion

We have sketched the main features of the semantic keyword search techniques in the RFT system, which provides a location-based infrastructure for the sale of food. We use Linked Data information merged with an ontology designed by experts to combine the wide coverage of the former with the high quality of the latter. We adopt an approacb based on Information Retrieval techniques to compute the degree of similarity between two varieties of fish; this is used to match demand and supply in the RFT marketplace. Our first experiments validate the effectiveness of the approach.

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 are already studying how to incorporate learning algorithms that derive knowledge from users' behaviour in order to enhance the knowledge base used in the system. Last but not least, we intend to study the economic impact of the high level of information provided by RFT on the markets.

References

1. Network of fisheries ontology. `http://aims.fao.org/node/3059/atom/feed`.
2. E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc. of IJCAI*, pages 1606–1611, 2007.
3. G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.
4. T. D. Noia, R. Mirizzi, V. C. Ostuni, D. Romito, and M. Zanker. Linked open data to support content-based recommender systems. In *Proc. of I-SEMANTICS*, pages 1–8, 2012.
5. S. Patwardhan, S. Banerjee, and T. Pedersen. Using measures of semantic relatedness for word sense disambiguation. In *Proc. of CICLing*, pages 241–257, 2003.