Using Open Information Extraction and Linked Open Data towards Ontology Enrichment and Alignment

Antonis Koukourikos[1,2], Pythagoras Karampiperis[2], George Vouros[1], and Vangelis Karkaletsis[2]

[1] University of Piraeus, Department of Digital Systems. 80, Karaoli and Dimitriou Str, Piraeus, 18534
[2] Software and Knowledge Engineering Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos"Agia Paraskevi Attikis, P.O.Box 60228, 15310 Athens, Greece

Abstract. The interlinking, maintenance and updating of different Linked Data repositories is steadily becoming a critical issue as the amount of published data increases. The wealth of information across the World Wide Web can be exploited in order to provide additional information about the way that an object is described in the real world. This paper proposes a method for discovering new concepts and examining the equivalence of properties in different LOD description schemas by using Open Information Extraction techniques on web resources. The method relies on constructing association graphs from the extracted information, proceeding to a transfer on the conceptual level using information previously known from the LOD repositories and examining the similarities and discrepancies between the produced graphs and the LOD descriptions, as well as between the graphs derived from different repositories.

Keywords: Ontology Enrichment, Ontology Alignment, Linked Open Data, Open Information Extraction.

1 Introduction

The need to describe in a consistent, machine-readable way the vast amount of information found on the Web has led to numerous initiatives related to the Semantic Web movement[1]. The Linked Data Initiative provides a set of guidelines and best practices for the publication and interlinking of different resources in RDF[2] format. The Linking Open Data community project[3] dedicates its efforts into collecting various datasets, publishing them as RDF triples and establishing semantic links between them [1]. Two interesting aspects that arise is the adequacy of the ontologies used for describing an entity in a dataset and the ease of identification of relations between description schemas. In the present paper, we describe a methodology for exploiting

[1] http://www.w3.org/2001/sw/
[2] http://www.w3.org/RDF/
[3] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

M. Bajec and J. Eder (Eds.): CAiSE 2012 Workshops, LNBIP 112, pp. 117–122, 2012.

information on the instance level found in the World Wide Web in order to propose additional meaningful concepts for describing a class of objects in a dataset. Furthermore, the method allows the identification of similarities between different schemas and thus provides the basis for aligning their underlying ontologies. The method relies on producing association graphs between the entities present in the relations found from the information extraction process, identifying entities corresponding to an entity known from a LOD dataset and examining the presence of additional properties not present in the description schema used by the dataset.

The paper is structured as follows. First, we briefly present related work. Then, we provide a description of the Linked Data datasets that we used. In section 4, we present the process for acquiring web documents and retrieving relation tuples from them. In section 5, we describe the preliminary method for constructing a graph corresponding to the aforementioned relations and some observations on the results of this process. The final section reports our conclusions so far as well as the next steps for improving the tools involved and assessing the effectiveness of the proposed method.

2 Related Work

Ontology alignment is a key-technology for the purpose of interlinking different datasets, by discovering equivalent and/or semantically similar classes and properties between the ontologies of distinct repositories. Correspondences can then drive further associations between data in distinct repositories. Some examples of recent alignment systems are SAMBO [2], ASMOV [3] and RIMOM [4]. Given the wealth of instance-level data available in the LOD repositories, it is important to consider techniques based on acquiring knowledge from instances. In this regard, in relation to LOD, certain approaches aim to connect datasets at the conceptual level using LOD information as it is. The BLOOMS ontology alignment system [5] is such a system. It uses information existing in the Wikipedia hierarchy in order to bootstrap the alignment of two input ontologies. However, the BLOOMS approach is limited to the categorization of Wikipedia to perform the contextual match, not taking into account the possible association with other available classification systems.

Ontology enrichment relies heavily on processes like object identification, synonym resolution and relation extraction in order to expand an existing ontology with additional and/or specialized terms. Depending on the used approach, ontology learning systems can enrich a given ontology with both concepts and relations [6, 7], solely with concepts [8] or solely with relations [9, 10]. The purpose of our experiments is to exploit the structured nature of the data available in the LOD repositories in order to improve on both the relation and concept extraction from raw web content by associating the discovered information with the LOD.

3 The Jamendo and Magnatune Data Sets

Jamendo is a repository of music licensed under Creative Commons. Jamendo hosts 55451 albums; however, that dataset includes 5786 of them. For the purposes of our experiment, we used the RDF dump for the dataset, since the available web service seems unstable and became frequently unavailable. The dataset is interlinked with

GeoNames[4], a geographical database. The database is accessible via daily dumps and Web Services. For obtaining lexical information on the GeoNames entities included in Jamendo, we used the Java client for GeoNames Web Services.

Magnatune is an independent label, which follows the pay-as-you-wish business model for the creations of its artists. Part of the label's roster is described at the Magnatune dataset, in similar fashion with that of Jamendo. The dataset includes descriptions for 318 artists, 706 records and 17203 music tracks.

4 Information Extraction from Web Search Results

The first step in the process of acquiring a corpus of Open Web resources is to retrieve pages relevant to the space covered by the Jamendo dataset, or a subset of it. For the task, we used the Bing Search API 2.0. The main entity that we used for our search queries was the artist's name, retrieved from the `foaf:name` property within the RDF description of a given entity. When the name contained less than two words (excluding stop-words like "the", "a" etc.) we appended the query string with the term "music", in order to avoid large amounts of irrelevant results from the search engine.

The names were selected in the following manner: The artists whose descriptions included a `foaf:based_near` property, that is the artists linked with a GeoNames feature, were given priority. From the set of 3505 in the RDF dump, this restriction held for 3244 of them. These artists were selected as candidates for performing queries. For the construction of the query set, we indexed the remaining artist names and selected random indexes. For the Magnatune dataset, we randomly selected 200 artists without any further restrictions, since the amount of available descriptions was significantly lower. The search queries were then executed and we reserved the 50 first search results returned by the Bing Search API for further processing in both cases.

In order to get a text segment suitable for information extraction, we processed each page retrieved from the search session in the following manner: The raw HTML page was stripped from irrelevant information. This included formative content like HTML tags and scripting sections, as well as content that was irrelevant to the main text of the page, like menus, ads, lists of previous articles etc. We based the module for removing such elements on the boilerpipe library[5]. A step that was deemed necessary was the resolution of co-references within the text. The absence of such an analysis led to the production of numerous relations that were not useful since they associated entities that could not be resolved. Use of pronouns and generic terms, like "the band", "the group" do not allow the direct expansion of the relation set for an entity. To overcome this issue, we use the co-reference resolution module of the OpenNLP Tools[6]. Finally, we enforced the OpenNLP name entity resolution module in order to identify Named Entities within the obtained text segments. Lexemes recognized as named entities were given greater priority during the graph construction phase, as the probability of them being distinct object is greater.

[4] http://www.geonames.org
[5] http://code.google.com/p/boilerpipe
[6] http://opennlp.sourceforge.net/projects.html

The documents obtained from the previous linguistic analysis were fed to the Open Information Extraction module. We used REVERB [11], a second generation OIE system. REVERB builds on the methods established in older systems, like TEXTRUNNER [12], by enforcing lexical and syntactic constrains in the extraction process. The triples that do not satisfy these constraints are not included in the returned relation set, thus the amount of irrelevant or erroneous relations is reduced. The corpora formed from the web search based on the two datasets were given as input to REVERB separately. Due to limitations of the boilerplate removal, some relations were not coherent, since they contained code snippets or HTML tags as arguments. We passed the initial results from a simple heuristic-based module in order to eliminate such relations. In the end, there were 506420 relations derived from the Jamendo search results and 438530 from the one for Magnatune.

5 Construction of the Object Graphs

Following the production of relations from the OIE module, the next step was to construct a graph denoting the associations between entities, as they occurred in the relation set. The central entities were the ones associated with the name of a musician, as retrieved from the datasets. Each of these entities constitutes a node in the graph. Going through the relation set, we retrieved every relation that involved the specific entity as either argument. These relations are the vertices from the given node to other entities. After repeating the step for every name, we constructed an unconnected graph that included every direct relation discovered for every musician.

The next step was to identify relations that are already in the RDF of the dataset. For each tuple (*E, rel, Arg*) or (*Arg, rel, E*), where *E* is the currently examined entity, *rel* is a lexicalization of an association and *Arg* is the other entity involved in the relation, we examined if *rel* is synonymous or similar to a property in the RDF. Specifically, we compared *rel* and the property names found within the LOD RDF using JWNL[7], a Java library for accessing WordNet[8]. If the two strings, after trivial manipulation like elimination of non-letter characters, are found to have a common sense in WordNet, the two relations are considered similar. If the argument *Arg* in the relation triple is also a lexicalization of the object linked with the entity *E* in the LOD dataset, we claim that the entity belongs to the dataset and the rest of the relations are possibly additional concepts for describing it.

An example of a sub-graph for a specific artist -with some indicative relations- is depicted in figure 1. The produced graph includes relations that exist in the RDF description from the Linked Data repository (e.g *(Beth Quist, recorded, Lucidity)* and meaningful relations that could be included in the description (is a vocalist and pianist, participated in the duo Ishwish). However, there are relations that have no concrete meaning (is an angel, includes perspective) and relations that are not sufficiently resolved (recorded her last solo album).

[7] http://sourceforge.net/projects/jwordnet/
[8] http:// wordnet.princeton.edu/

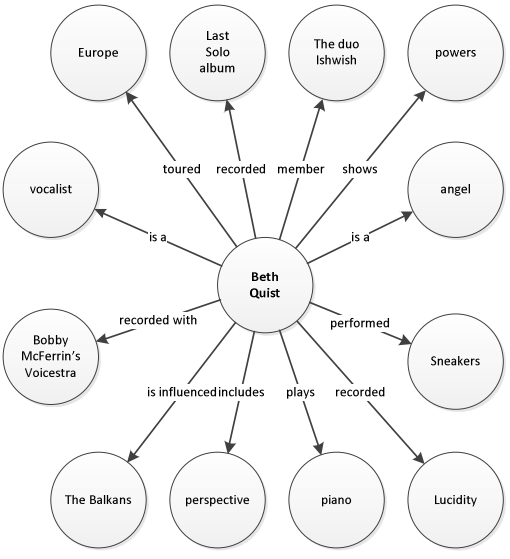

Fig. 1. Part of the association graph for the artist Beth Quist (http://dbtune.org/magnatune/artist/beth_quist)

The existence of a relation from or to an entity of known classification in multiple entities within the same dataset and between the two distinct datasets used for extracting web information is a good quantifier for increasing the probability that the property holds as a characteristic of the class and therefore it could be added in the ontology. For example, relations stating participation in a music group are frequent, so it is highly possible that membership is a meaningful property. On the contrary, a relation stating that musician A "is an angel" does not appear for multiple artists, therefore the probability of such a relation being meaningful is reduced.

6 Conclusions and Future Work

The purpose of our experiments at this stage was to examine the possibilities of acquiring knowledge in the conceptual level by combining web information and data available from LOD repositories in the instance level.

The preliminary results in a restricted domain (works of music) indicated that there is significant room for the introduction of additional properties not taken into account in the descriptions provided by the examined schemas. Furthermore, the equivalent properties between the two ontologies are revealed in a straightforward manner, though it is important to observe the inherent similarity of the schemas in the first place. To obtain more conclusive results, we will test the alignment opportunities in datasets with greater degree of differentiation.

For this run of the system, the hypothesis was that a lexicalization is considered as equivalent to a single entity. This held true in most of the cases, as the domain of the original search results was restricted from the beginning. However, there were exam-

ples where entity disambiguation should take place. For example, the music artist with the name "Cicadas" was associated with irrelevant results, as the insect with the same name produces "music" so the query did not differentiate the results for the band and the insect. For more generic terms the veracity of this statement was extremely low, as expected. The term "members" was a typical problematic word, since there were membership relations for multiple bands and these relations were included in the same sub-graph, connected with a single entity. In order to resolve this issue, we will expand the OIE system used in order to keep track of the relations derived from each web page and associate occurrences of relations including the term with a specific artist, the one for which the page is relevant. After the improvements in the infrastructure of our system and the distinct processes involved, we will apply a strict, formal graph similarity method in order to quantify the distance between the graphs derived from the relation triples and the LOD description graph.

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked data – the story so far. Intl J. on Semantic Web and Information Systems 5(3), 1–22 (2009)
2. Lambrix, P., Tan, H.: SAMBO – a system for aligning and merging biomedical ontologies. Journal of Web Semantics 49(1), 196–206 (2006)
3. Jean-Mary, Y.R., Shironoshita, E.P., Kabuka, M.R.: Ontology matching with semantic verification. Journal of Web Semantics 7(3), 235–251 (2009)
4. Li, J., Tang, J., Li, Y., Luo, Q.: Rimom: A dynamic multistrategy ontology alignment framework. IEEE Transactions on Knowledge and Data Engineering 21(8), 1218–1232 (2009)
5. Jain, P., Hitzler, P., Sheth, A.P., Verma, K., Yeh, P.Z.: Ontology Alignment for Linked Open Data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 402–417. Springer, Heidelberg (2010)
6. Alani, H., Sanghee, K., Millard, E.D., Weal, J.M., Lewis, P.H., Hall, W., Shadbolt, N.: Automatic Extraction of Knowledge from Web Documents. In: Proc. of HLT (2003)
7. Kim, S.-S., Son, J.-W., Park, S.-B., Park, S.-Y., Lee, C., Wang, J.-H., Jang, M.-G., Park, H.-G.: OPTIMA: An Ontology Population System. In: 3rd Workshop on Ontology Learning and Population (2008)
8. Etzioni, O., Kok, S., Soderland, S., Cagarella, M., Popescu, A.M., Weld, D.S., Downey, D., Shaker, T., Yates, A.: Unsupervised named-entity extraction from the Web: An experimental Study. Artificial Intelligence 165, 91–134 (2005)
9. Brewster, C., Ciravegna, F., Wilks, Y.: User-Centred Ontology Learning for Knowledge Management. In: Andersson, B., Bergholtz, M., Johannesson, P. (eds.) NLDB 2002. LNCS, vol. 2553, pp. 203–207. Springer, Heidelberg (2002)
10. Suchanek, F.M., Ifrim, G., Weikum, G.: LEILA: Learning to Extract Information by Linguistic Analysis. In: Proc. of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, Sydney, Australia, pp. 18–25 (2006)
11. Etzioni, O., Fader, A., Christensen, J., Soderland, S.: Mausam: Open Information Extraction: the Second Generation. In: Intl. Joint Conference on Artificial Intelligence (2011)
12. Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., Soderland, S.: TextRunner: Open Information Extraction on the Web. Computational Linguistics 42, 25–26 (2007)