Entity Coreference Resolution Services in Sindice.com: Identification on the current Web of Data

Giovanni Tummarello and Renaud Delbru

Digital Enterprise Research Institute
National University of Ireland, Galway
firstname.lastname@deri.org

1 Introduction

Sindice [1] is a backend service that operates on semantically structured data harvested from the Web. Sindice uses both crawlers and Semantic Web Sitemaps [2] to find RDF sources as well as microformats[1] such as XFN, hcards, hvote and others. Sindice targets developers by offering the following a set of API to find, reuse and publish structured data on the Semantic Web.

Examples of low level APIs provided by Sindice include locating an RDF source given a URI, a string or a tuple property/value that should be contained in the source itself. High level APIs provided by Sindice include, for example, a SIOC specific API to interconnect messages posted by people in different settings.

In this paper we introduce the operating principles behind the Entity Coreference Resolution APIs, soon to be offered by Sindice. This new API addresses primarily two use cases:

Entity Coreference Resolution Given an entity description, locate existing URIs on the Web that correspond to an entity that matches, to some degree, the description.

Graph Coreference Resolution Given a graph of interconnected entities, expressed in RDF, provide a new RDF graph where as many entities as possible are associated with alternative URIs in the form of OWL:SAMEAS statements.

2 Entity Coreference Resolution Services

The two services share a common model of description for entities on the Semantic web. During the harvesting phase, several layers of information are collected and theyre used power the matching algorithm.

The first distinction is made between authoritative and non authoritative information. By authoritative we adopt the meaning that can be intended from [3], where it is said that "A URI owner may supply zero or more authoritative representations of the resource identified by that URI". Authoritative information is therefore given by the source itself and can exist only if the identifying URI is a resolvable URI, in line with

[1] Microformats: www.microformats.org

what suggested by the Linked Data[2] on the Semantic Web [4]. Based on this definition, we can define four classes of information that will be useful for our entity matching tasks.

Authoritative Entity Information In case of RDF, this information is composed by any triple which can be obtained by resolving directly the URI.

Non Authoritative Entity Information Using shared identifiers (URIs) or other methods such as Inverse Functional Properties in OWL, it is possible for external sources to state information about any entity. This information is in general non authoritative, but as it is precisely linked to the original entity, it is potentially very valuable.

Authoritative Contextual Information Authoritative contextual information is extracted from the context where the entity is listed. It can include items such as the return HTTP header (e.g. information about when the resource was last updated), or information about the entire collection in which the resource is listed (e.g. by locating the description of the dataset as specified in [2]).

Non Authoritative Contextual Information Information about the context can come from outside the context itself. For example, a PageRank rating of the website hosting the entity description will come from the evaluation of the links from other contexts.

2.1 Entity Coreference Resolution

The *Entity Coreference Resolution Service* (ECRS) makes use of all the above information to come up with potential match candidates. Within our current Sindice implementation, the service operates in 3 steps: first it will use the existing indexes and perform a preselection of relevant information sources based on some queries. The preselection queries are usually fuzzy text matching queries on textual literals and aim to reduce the number of entities to match from hundreds of millions (the current number of entities in Sindice) to a few hundreds.

Next, a state of the art record linkage analysis is performed on the available local and remote entity description. In particular, an enlarged record for the entity is constructed from the authoritative entity and contextual information and matched with external entity information. At this stage, a priori knowledge is applied in form of knowledge templates which weight properties (e.g. vocabulary terms) differently according to combination of context and entity descriptors.

Finally, the last stage of the service is performed recursively, when this is considered appropriate, e.g. in presence of chains of OWL:SAMEAS and strong positive indications from non authoritative contextual information. It is to be noticed that the *Entity Coreference Resolution Service* can also be used interactively, by simply allowing users to enter an entity description and provide feedbacks on the resulting matches.

2.2 Graph Coreference Resolution

The *Graph Coreference Resolution Service* (GCRS) could be implemented by iterating the ECRS service over every entity identifiers (URIs or IFPs) contained in a RDF graph.

[2] W3C SWEO Linking Open Data: `http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`

In practice, however, ECRS could make stronger use of the "context" or co-presence of same entities across different collections. For example, if URI_{O1} in the original graph G_O has been matched with URI_{E1} in an external graph G_E, then when trying to find matches for URI_{O2} in G_O, the GCRS algorithm will carefully investigate G_E more than other graphs to find possible matches. This has two effects: on the one hand it increases consistency between datasets linkage, on the other it fundamentally speed up the matching operations as compared to individual ECRS calls.

3 Related Works

At matching level, our approach leverages well known record linkage techniques [5] with variants which specifically consider Semantic Web or graph-based data [6, 7]. Probably the most relevant works that is currently being carried out in this field is the Okkam[3] project. We feel that the main difference between the Okkam approach and our is that Sindice wants to perform such tasks without requiring, nor offering, interaction from the data producer side. In other word, our API does not match, nor assign 'universal identifiers" as Okkam does, but rather only answers to direct questions such as "which existing identifiers should be connected?".

4 Conclusion

For the Semantic Web to prove its usefulness, there is the need to show convincing example of automatic or mostly automatic aggregation of information coming from diverse information sources. In this paper we illustrate the preliminary works inside Sindice to offer APIs that address this need based only on the available corpus of harvested Web data.

References

1. Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., Tummarello, G.: Sindice.com: A document-oriented lookup index for open linked data. International Journal of Metadata, Semantics and Ontologies **3** (2008)
2. Cyganiak, R., Delbru, R., Stenzhorn, H., Tummarello, G., Decker, S.: Semantic Sitemaps: Efficient and Flexible Access to Datasets on the Semantic Web. In: Proceedings of the 5th European Semantic Web Conference. (2008)
3. Group, W.T.A.: Architecture of the World Wide Web, Volume One. W3C Recommendation, W3C, http://www.w3.org/TR/webarch/ (2005)
4. Sauermann, L., Cyganiak, R.: Cool URIs for the Semantic Web. W3C Note, W3C, http://www.w3.org/TR/2008/NOTE-cooluris-20080331/ (2008)
5. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering **19** (2007) 1
6. Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S.E., Widom, J.: Swoosh: a generic approach to entity resolution. The VLDB Journal (2008)

[3] Okkam Project: `http://www.okkam.org`

7. Rocha, C., Schwabe, D., de Arago, M.P.: A hybrid approach for searching in the semantic web. In: Proceedings of the 13th conference on World Wide Web - WWW 04 WWW 04. (2004) 374