Methodology for Searching Entities on the Web*

Renaud Delbru

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

1 From a Web of Documents to a Web of Entities

The Semantic Web is driven by the idea of moving from a Web of documents, designed for human consumption, to a Web of data in order to "create a universal medium for the exchange of data where data can be shared and processed by automated tools as well as by people"[1].

Nowadays, more and more machine-readable annotations and meta-data are available on the Web. This data, typically codified using the Resource Description Framework (RDF) or Microformats, is accessible directly via HTTP. Microformat enables the annotation of an entity in a web page, whereas RDF enables the description of anything that can be named using a Uniform Resource Identifier (URI). By describing the relationships between resources, the Web moves from a Web of documents to a semantically interconnected Web of entities.

Although data are available, data consumers face a challenge due to the decentralised publishing infrastructure of the Web: they need to locate information about an entity and handle multiple, possibly discording, views of the entity.

Search engines are the primary method for accessing information on the Web, i.e. finding relevant documents given a keyword-based query. By leveraging the Web of entities, we can imagine an entity-centric search engine which, given a query, would support the user in obtaining an aggregated and balanced view of the data available on the Semantic Web. Given that the Semantic Web data are machine processable, the most interesting use of such an engine could be made by machines themselves: any application could use one such engine directly to find, interconnect and enrich information.

Searching information about a particular entity on the Web raises new challenges: (i) how to efficiently locate and retrieve Semantic Web data and (ii) how to integrate data on a decentralised and heterogeneous information space. We aims to propose a comprehensive methodology for searching entities on the Web along these requirements.

2 Research Problem

We plan to tackle such complex problems by exploring how existing and proven robust technologies can be advanced and specialized specifically to address the needs of an

* This material is based upon works supported by the European FP7 project *Okkam - Enabling a Web of Entities* (contract no. ICT-215032), and by Science Foundation Ireland under Grant No. SFI/02/CE1/I131.

[1] Semantic Web Activity Statement: http://www.w3.org/2001/sw/Activity.html

entity-centric search engine. In particular, our work will focus on the topics illustrated in the following sections.

2.1 Adapting Information Retrieval engines for Semantic Web Data

Standard Web search engines are intensively using Information Retrieval (IR) techniques for locating relevant information on the Web. Information Retrieval is a well studied field [1] and many optimisations have been developed for efficiently storing and querying large amount of information. Techniques such as inverted indexes [2] have proved to scale to the size of the Web (e.g. Google). The shortcoming of such systems is that they can only answer simple queries, e.g. a boolean combination of words, but are not really meant to query relationships between entities, e.g. a graph pattern.

On the contrary, entity-centric search engines such as SWSE [3] are built on a data structure which is more similar to relational databases than to IR engines. SWSE relies on YARS [4], a distributed RDF store, for storing and querying large amounts of graph-structured data. Such systems can typically answer complex conjunctive queries involving large joins, but they are in turn difficult to scale since they need clever indexes for query efficiency which are however computationally expensive to update.

Our intuition is that it is possible to construct a fast and scalable entity centric search engine based on a two-tier architecture: a modified IR engine to efficiently perform a preliminary semantic document selection, and an optimised triple level post-processing to answer complex queries. Our research will therefore focus on how to employ existing IR engines to perform useful queries over semantically structured documents.

Information Retrieval engines, however, are primarily designed for unstructured text information, and not for graph-structured information such as RDF. Information Retrieval engines for Semantic Web data have notable previous works with Semplore [5] and ESTER [6] which however were developed with different goals than those we consider. The developers of Swoogle [7] have also discussed the problem of introducing a new search paradigm for Semantic Web resources and emphasized the importance of combining knowledge inference with information retrieval methods.

2.2 Optimising Inference at Web Scale

Reasoning over semantically structured documents enables to make explicit what would otherwise be implicit knowledge: it adds value to the information and enables an entity-centric search engine to ultimately be much more competitive in terms of precision and recall [7]. The drawback is that inference can be computationally expensive, and therefore prevent efficient indexing.

The novel aspect that our work covers is how to reason over semantically structured documents that have been harvested from the Web. To reason on documents, we assume that ontologies, which are referenced explicitly with OWL:IMPORTS or implicitly by using properties and classes of a certain namespace, are also part of the Semantic Web as dereferenciable data, in accord with the W3C Best Practice[2]. As ontologies might

[2] Best Practice Recipes for Publishing RDF Vocabularies: http://www.w3.org/TR/swbp-vocab-pub/

refer to other ontologies, the web fetching process is recursive and should, in theory, be repeated for each harvested documents independently.

The proposed research will focus on how to maximally reuse the results of such "web closure reasoning", i.e finding and exploiting the referenced ontologies, that has been performed over previously indexed semantically structured documents in order to minimise the computational cost of indexing. We will also considers how to "keep in quarantine" reasoning tasks and inference results in order to prevent maliciously crafted web ontologies to alter the semantics of agreed ontologies published by third parties on a global level. For example, if an ontology states that FOAF:NAME is an inverse functional property, an inferencing agent should not consider this axiom outside the scope of the document that references this particular ontology.

The coordinate use of the features offered by the IR and inference engines will be demonstrated in the applications described in the following sections.

2.3 Identification, Coreference Resolution and Information Merging

Due to its decentralised publishing infrastructure, information about an entity are generally spread across the Web. The identification of an entity is fundamental for discovering complementary data sources. The use of URI makes easier the identification of an entity, but the Unique Name Assumption (UNA) does not hold. In theory, a single URI uniquely identifies a resource, but it is unrealistic to assume that data publishers can universally agree on a single identifier for each resource. Therefore, the identification of an entity among the Semantic Web becomes uncertain since two identifiers, apparently distinct, can refer to a unique entity.

The coreference problem is well known across various research communities with a variety of different names, such as record linkage [8], entity resolution [9], reference reconciliation [10] or object consolidation [11]. A wide variety of algorithms has been developed for resolving the coreference problem, but these are generally not designed for Web scale and semi-structured data. Recent initiatives amongst the Semantic Web community addressed the problem of resource identification: [12] described the phenomenon of the proliferation and coreference of URIs and the OKKAM project[3] proposed to research an infrastructure for assigning global identifiers at Web scale.

The problem of identification and coreference resolution will be a natural testbed for the IR and inference engines that we described previously. The IR engine will enable to perform a blocking pass [13] before executing complex coreference resolution and, coupled with the inference engine, will permit more advanced reasoning than what was possible in the Semantic Web object consolidation work described in [11].

Clearly, coreference resolution is an important enabler for information merging. More factors, however, have to be taken into consideration before aggregating diverse information sources. Entity descriptions are generally produced under a certain context (provenance, time, etc.). The descriptive information is usually a subjective view of the entity with a certain level of reliability. Merging these descriptions can result in inconsistent and contradictory information. In order to enable a proper data integration, we

[3] http://www.okkam.org/

have to keep information in its context which is naturally supported by the document-centric storage we adopt.

3 Methodology

We will evaluate the methodology for searching entity by implementing a solution for each identified problem, by integrating them into a single software platform and by performing a qualitative evaluation of the resulting platform. In addition, we will perform an evaluation of each solution with a dedicated corpus, as described below.

Information Retrieval Engine for Semantic Web Data A benchmark, including index size and query response time, against other systems is planned.

Semantic Web Inference Engine The evaluation of the inference engine will include an analysis of its complexity in term of size and response time.

Entity Identification and Coreference Resolution The evaluation of the coreference resolution system requires a gold standard dataset for analysing the precision of the different algorithms.

4 Achievements and Work Plan [4]

Information Retrieval Engine for Semantic Web Data We achieved, as part of the Sindice project [14], a first prototype of the Information Retrieval engine. The current system has currently indexed more than 2 billions of triples. The system enables fast lookup of URIs, keywords and Inverse Functional Properties (IFP) through a human interface or a HTTP API for machine access. We are currently finishing a second prototype that enables queries of increased complexity and semantic meaning, i.e combining URIs and keywords and adding triple-structure. We foresee a third and final prototype capable of answering more complex queries involving simple joins.

Semantic Web Inference Engine We have developed a prototype of an optimised inference engine that enables inference of a subset of OWL at indexing time. Preliminary results of this work has been published in [14]. We have formalised an advanced inference engine that avoid malicious users from "infecting" cached data on a global scale. Its development is in progress.

Entity Identification and Coreference Resolution As a next step we will tackle a coreference resolution system, based on the IR and inference engines. The first task will be to implement a prototype for identifying entities with the help of OWL:SAMEAS statement and IFPs, e.g. the e-mail of a person. The prototype will be able to return an aggregated view of the entity information available on the Web. The second task will be to improve the system with "pair-wise" matching algorithms.

[4] The work on the thesis has formally started in February, 2007.

References

1. Baeza-Yates, R.A., Ribeiro-Neto, B.A.: Modern Information Retrieval. ACM Press / Addison-Wesley (1999)
2. Zobel, J., Moffat, A.: Inverted files for text search engines. ACM Computing Surveys **38** (2006) 6
3. Harth, A., Hogan, A., Delbru, R., Umbrich, J., Ó'Riain, S., Decker, S.: SWSE: Answers before links! In: Proceedings of the Semantic Web Challenge, 6th International Semantic Web Conference. (2007)
4. Harth, A., Umbrich, J., Hogan, A., Decker, S.: YARS2: A federated repository for querying graph structured data from the web. In: Proceedings of the 6th International Semantic Web Conference. (2007) 211–224
5. Zhang, L., Liu, Q., Zhang, J., Wang, H., Pan, Y., Yu, Y.: Semplore: An IR approach to scalable hybrid query of semantic web data. In: Proceedings of the 6th International Semantic Web Conference. (2007) 652–665
6. Bast, H., Chitea, A., Suchanek, F., Weber, I.: ESTER: efficient search on text, entities, and relations. In: SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2007) 671–678
7. Mayfield, J., Finin, T.: Information retrieval on the Semantic Web: Integrating inference and retrieval. In: Proceedings of the SIGIR Workshop on the Semantic Web. (2003)
8. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. Journal of the American Statistical Association **64** (1969) 1183–1210
9. Benjelloun, O., Garcia-Molina, H., Jonas, J., Su, Q., Widom, J.: Swoosh: A generic approach to entity resolution. Technical report, Stanford University (2006)
10. Dong, X., Halevy, A.Y., Madhavan, J.: Reference reconciliation in complex information spaces. In Özcan, F., ed.: SIGMOD Conference, ACM (2005) 85–96
11. Hogan, A., Harth, A., Decker, S.: Performing object consolidation on the semantic web data graph. In: Proceedings of the WWW2007 Workshop I3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web. (2007)
12. Jaffri, A., Glaser, H., Millard, I.: URI identity management for semantic web data integration and linkage. In: 3rd International Workshop On Scalable Semantic Web Knowledge Base Systems, Springer (2007)
13. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering **19** (2007) 1–16
14. 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)