Search Result Ontologies for Digital Libraries

Emanuel Reiterer

School of Information Systems, Curtin University, Perth, Western Australia, Australia
emanuel.reiterer@postgrad.curtin.edu.au

Abstract. This PhD investigates a novel architecture for digital libraries. This architecture should enable search processes to return instances of result core ontologies further on called result ontologies linked to documents found within a digital library. Such result ontologies would describe a search result more comprehensively, concisely and coherently. Other applications can then access these result ontologies via the web. This outcome should be achieved by introducing a modular ontology repository and an automatic ontology learning methodology for documents stored in a digital library. Current limitations in terms of automatic extraction of ontologies should be overcome with the help of seed ontologies, deep natural language processing techniques and weights applied to newly added concepts. The modular ontology repository will be comprised of a top-level ontology layer, a core ontology layer and a document and result ontology layer.

Keywords: ontology, ontology learning, ontology modularisation, digital library, semantic digital library, semantic data management, search result ontology

1 Motivation and Research Questions

The following motivators led to this research: Firstly, the semantic accessibility of documents within a digital library could contribute to the semantic web. This could be achieved by ontologies created automatically and triggered by a conceptual search. Secondly, an ontology repository within a digital library could enhance the search process within a digital library by enabling ontological search that includes more than a meta-data search and does not necessarily have to incorporate a full-text search. Thirdly, result ontologies, if rendered properly, could provide a concise, coherent, yet comprehensive search result to the user. This result will be concise because it will consist of a conceptualisation about a query and not only a set of documents, coherent because these concepts will be related meaningfully, and comprehensive because the whole result set of documents will be represented through one ontology.

These motivations lead to the following research question: Can a digital library be improved to enable more coherent, concise, yet comprehensive query result presentations by using ontologies?

2 State of the Art

This research covers two different research areas: digital libraries and ontologies. In computer science, the term ontology is used to mean "a formal, explicit specification of a shared conceptualisation" [1].

Cimiano [2] describes ontology learning as a reverse engineering process where an ontology reflects the author's point of view. Ontology learning includes several tasks: the extraction of terms, definition and hierarchical organisation of concepts, extraction of relations and attributes as well as the definition of axioms [2]. Deep natural language processing techniques, such as the use of lexico-syntactic patterns, are promising but not thoroughly investigated in current ontology learning processes [3]. This research proposes that patterns, which include the extraction of implicit information such as the train of reasoning, could improve the learning of expressive ontologies. Additionally, a standardised document core ontology could help to create consistent and reusable results.

There are an increasing number of ontology repositories available but current digital libraries could provide a wealth of new ontologies, although these ontologies have to be extracted first, which is part of this work.

A digital library is defined as "a focused collection of digital objects, including text, video, and audio, along with methods for access and retrieval, and for selection, organization, and maintenance" [4]. The goals of semantic digital libraries are to enhance information extraction, to connect information within a digital library, for query refinement, and also for recommendation services. Ontologies are used as bibliographic ontologies and community-aware ontologies [5]. Ontology repositories, built upon an ontology hierarchy, as well as implicit information, such as the extraction of the thesis statements could improve the search processes but also digital libraries in general. Additionally, a result ontology repository could combine information within a digital library, mentioned in a multitude of books or documents, by incorporating or referencing the actual document.

Open problems addressed in this research are (1) the learning of more expressive ontologies [6] by the use of deep natural language processing techniques, (2) the linkage between ontologies and unstructured documents, (3) the provision of standards for ontology repositories by strictly following a top-level ontology and the use of modularised ontologies within an ontology repository, and an (4) automatic ontology creation methodology.

3 Approach

This research proposes and will develop and evaluate five main new artefacts including (1) a generalised digital library architecture, which introduces (2) a modular ontology repository, (3) a search process, (4) an indexing process, and (5) an automatic ontology creation methodology. This work will extend a commonly used document repository system, a full-text search engine, as well as a natural language processing library. It also incorporates state-of-the-art ontology learning algorithms.

Figure 1 shows the initial architecture where a digital library is divided into a document repository, a document ontology repository, and a result ontology repository. The document ontology repository stores document ontologies about each document in the system. The result ontology repository consists of ontologies created on the fly if a search is not already represented by an ontology. This figure also depicts processes for search and indexing. The indexing process creates a document ontology out of an inserted document and updates affected result ontologies. The search process searches for existing result ontologies and combines document ontologies if no result ontologies are found. If no document ontology exists a full-text search will be initiated.

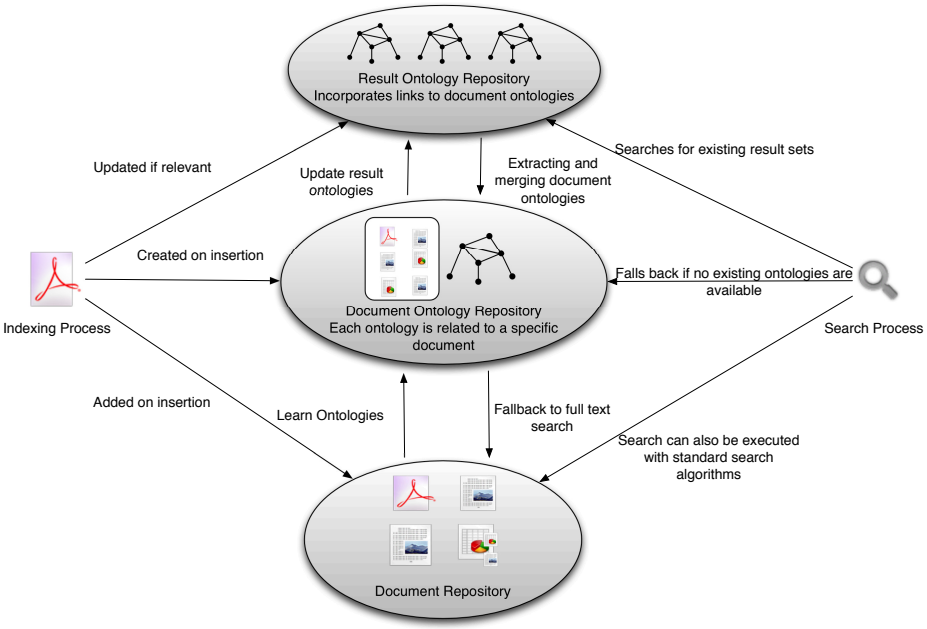

Fig. 1. Digital Library Processes and Repository Architecture

Figure 2 shows the ontology hierarchy used to integrate document ontologies and result ontologies. Although the main outcomes of this research are result ontologies, this hierarchy is essential to provide a consistent basis and is utilised to create such result ontologies. All ontologies are based on a common top-level ontology. The document core ontology describes four aspects: the structural aspect, the technical aspect, the syntactical aspect, and content. The result core ontology contains result based information. The reference ontology is comprised of contextual information. Each document is expressed by a document ontology. A result ontology is an instantiation of the result core ontology and incorporates subsets of a set of document ontologies. With this hierarchy it will be possible to

search for predefined concepts and relations in the document core ontology and the result core ontology but also more generally by using the reference ontology.

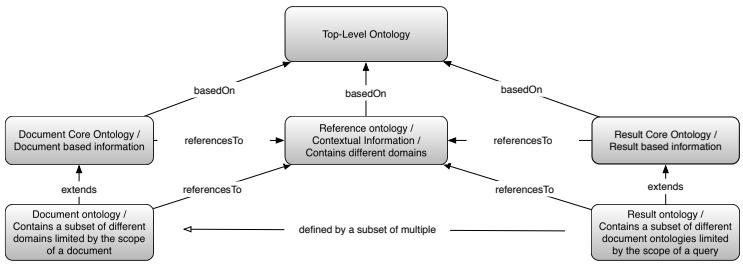

Fig. 2. Ontology Hierarchy

One difficulty of this approach is how to learn ontologies automatically. Well established seed ontologies, which are incorporated in a reference ontology, such as UMBEL (http://umbel.org), should mitigate this problem. Additionally, the use of lexico-syntactic patterns should make it possible to extract more valid and expressive ontologies. Also, weights for newly added concepts will be calculated, as proposed in Boese et al. [7], to minimise the influence of unimportant or false concepts. A limitation of this research is that it will not address the presentation of search result ontologies to the user.

4 Research Methodology

Because this research is about the design and evaluation of the artefacts mentioned in section 3, design science research [8] has been chosen. The artefacts will be evaluated ex-ante and ex-post [9]. Ex-ante evaluations should demonstrate the feasibility of the generalised architecture and algorithms. The implemented artefacts will then be evaluated ex-post. Artificial methods [9] will be used to analyse the artefact in terms of functionality and efficacy and naturalistic methods will be applied by asking ontology experts to evaluate the created ontologies.

5 Current Status and Future Work

This research started with the creation of an initial version of a document core ontology design and a result core ontology design. Afterwards, a proposal for an initial automatic ontology creation methodology has been defined, which meets the needs for use in a digital library. This methodology relies on seed ontologies that are already available and heavily utilised. Such ontologies are either hand selected or well established ones that are selected automatically. To support automatic selection of such ontologies, it is planned to utilise the ontology usage analysis framework by Ashraf [10].

To evaluate the intended benefits of the result ontologies a result ontology will be created manually and presented to a small group of study participants. These participants will then be interviewed about the completeness of this ontology concerning the searched topic in terms of concepts, relations, and linkage to the actual documents and the improvement of such a result in contrast to a normal result list. The next step includes the automatic creation of document and result ontologies. Deep natural language processing for ontology learning by defining lexico-syntactic patterns will build the basis for learning ontologies. After that, a generalised architecture for digital libraries as well as indexing and search processes will be defined and evaluated ex-ante. Then the artefacts will be implemented and finally evaluated ex-post.

References

1. Studer, R., Benjamins, R., Fensel D.: Knowledge engineering: Principles and methods. In: Data & Knowledge Engineering. Vol. 25(1): 161-197 (1998)
2. Cimiano, P.: Ontology Learning from Text: Algorithms, Evaluation and Applications. (p20) Springer Science and Business Media (2006)
3. Zouaq, A.: An Overview of Shallow and Deep Natural Language Processing for Ontology Learning. In: Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances. (pp. 16-37). Hershey, PA (2011)
4. Witten, I. H., Bainbridge, D., and Nichols, D.M.: How to Build a Digital Library (Second Edition). (p xvi) Burlington, MA : Morgan Kaufmann Publishers. (2010)
5. Kruk, S. R., Mc Daniel, B.: Semantic Digital Libraries. (p 5 / 73) Springer-Verlag Berlin Heidelberg (2009)
6. Völker, J., Haase, P., Hitzler, P.: Learning Expressive Ontologies. In: Ontology Learning and Population: Bridging the Gap between Text and Knowledge. (pp. 45-69) IOS Press (2008)
7. Boese, S., Reiners, T. and Wood, L. C.: Concept-based indexing in the design and construction of semantic document networks to support concept retrieval. (In Press) In: Encyclopedia of Business Analytics and Optimization, Hershey, PA: IGI Global
8. Hevner, A, March, S., Park, J., Ram, S.: Design science in information systems research. In: MIS Quarterly. Vol. 28 (1): 75-105 (2004)
9. Venable, J, Pries-Heje J., Baskerville, R.: A comprehensive framework for evaluation in design science research. In: Proceedings: Design Science Research In Information Systems. (pp 423-438) Springer Berlin Heidelberg (2012).
10. Ashraf, J., Khadeer Hussain, O., Khadeer Hussain, F.: A Framework for Measuring Ontology Usage on the Web. (In Press) In: The Computer Journal. Oxford University Press (2012)