Ontology Acquisition for Automatic Building of Scientific Portals

Pavel Smrž[1] and Vít Nováček[2⋆]

[1]Faculty of Information Technology, Brno University of Technology
Božetěchova 2, 612 66 Brno, Czech Republic
E-mail: smrz@fit.vutbr.cz
[2]Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: xnovacek@fi.muni.cz

Abstract. Ontologies are commonly considered as one of the essential parts of the Semantic Web vision, providing a theoretical basis and implementation framework for conceptual integration and information sharing among various domains. In this paper, we present the main principles of a new ontology acquisition framework applied for semi-automatic generation of scientific portals. Extracted ontological relations play a crucial role in the structuring of the information at the portal pages, automatic classification of the presented documents as well as for personalisation at the presentation level.

1 Introduction

Ontology acquisition framework described in this paper is a part of PortaGe — an ongoing project aiming at semi-automatic generation of scientific web portals. We would like to briefly introduce basic characteristics of the project that influenced our decisions in the area of ontology learning.

The generator of scientific web portals is meant as an extension of the existing tools such as Google Scholar (http://scholar.google.com) or CiteSeer (http://citeseer.ist.psu.edu/). A typical user is a young researcher or a PhD student that looks for relevant information (knowledge) in a subfield (s)he needs to fathom. The interest in the subject is supposed to be long-term, so the user would be notified about new publications, projects, events, calls, etc. in the field.

The current search engines employ user-specified keywords and phrases as the major means of their input. Digital libraries, such as ACM DL (http://portal.acm.org/dl.cfm) or Springer DL (http://arxiv.org/), add a detailed metainformation level and are able to find publications of a given author, from a given journal, conference proceedings etc. However, these services are not able to

⋆ This work was partially supported by the Ministry of Education of the Czech Republic, Research Plan MSM 6383917201, and by the Grant Agency of the Czech Academy of Sciences, Project T100300419.

relate the information to the context of the search. They cannot evaluate what "relevant" means in a particular case.

PortaGe builds a web portal for a domain given by initial data. In addition to the standard keywords, known authors, journals, conferences or projects characterising the subject field, the user can provide seed documents and conference/project web pages relevant for the current search and select apt nodes in the current ontology (automatically extracted from the given and retrieved documents). The tool combines responses from several information sources:

- search results from Google Scholar;
- articles and papers found in digital libraries (currently available ACM DL and Springer Link);
- information from freely accessible web services (arxiv.gov and ResearchIndex);
- metainformation about hard-copies (books, journals, proceedings) in the faculty library and other traditional repositories.

Besides the ontology acquisition by means of text mining which is tackled in the next sections, the essential components of PortaGe include: efficient local document classification and indexing, extraction of metainformation from the documents, citation analysis (from ResearchIndex), metasearch in digital libraries, analysis of "Publications" web pages, meta-data annotation of web resources, merging of information, continuous search and source-change analysis. The personalisation of the portal driven by ontologies is discussed in the next section.

The rest of the paper is organized as follows: The role ontologies play in PortaGe and the consequences in the form of requirement specification for the automatic acquisition system are presented in the next section. Section 3 describes fundamentals of OLE — a new ontology acquisition framework and OLITE — its essential part designed primarily for the extraction of detailed semantic relations from unstructured plain-text data. A brief comparative overview of other relevant approaches and related works is given in Section 4. We conclude the paper by proposing future directions for our research.

2 Ontologies in the Scope of PortaGe

Several components of PortaGe take advantage of domain-specific as well as general ontologies. This impacts the way the automatic ontology acquisition has been implemented. The particular needs have determined the methods and techniques that could be applied for the extraction of semantic relations. The following paragraphs briefly introduce the role of ontologies in PortaGe and summarise the defined requirements.

Ontologies found their place in a couple of areas within PortaGe:

1. The basic role consists in the definition of portal structures. The core ontology contains concepts of publishers, books and book series, journals and

their special issues, conferences, conference tracks workshops, projects, research teams, authors, papers, web pages, etc. PortaGe supposes that the most of this can be shared among various scientific fields (different disciplines slightly yet differ in the conceptualisation of their research areas). For a particular domain, it needs to be extended by individual instances of journals, conferences, etc. It is one of the tasks of an ontology extraction engine.

2. Ontologies also help to classify the content of documents in PortaGe. This is important especially for very narrow subfields with a limited number of documents that can be applied for training of the standard classifiers. The automatic classification process can base its decision on the knowledge extracted from other documents in a previous run, such as the fact that a particular method is used for machine learning in other fields.
3. As stated above, it is difficult to define a context of the search when using the standard search engines. Ontologies provide mechanisms for a comprehensive context specification. In PortaGe, the user can restrict the search for documents reflecting certain semantic relations based on the ontology, e.g. limit the output to the documents discussing "context-free grammars" as a "tool-for" "analysis of protein sequences". The OLE framework interlinks individual pieces of such knowledge with lexico-syntactic patterns able to identify the relations in the retrieved documents.
4. The discussion of the PortaGe system has assumed a single individual user of the generated portals so far. However, the multi-user environment is much more realistic in many circumstances. For example, imagine a typical scenario of a team leader that supervises several PhD students. He creates a general web portal that covers various subfields of the area in focus. Individual students work on their particular topics, interact with the system and extend its coverage in the given subfield. The last role of the ontologies in PortaGe that will be mentioned here deals with the personalisation of general portals. The system uses ontologies to evaluate what "relevant" information means for a particular user. Based on user profiles PortaGe defines rules to identify "the best" information for an individual user. A novice (in the given research domain) can ask for introductory documents, others prefer new information (the documents that appeared/were found in the last month), need a general summary of used methods (usually the most referenced documents), or focus on the relevance only. The user profiles and the ontologies also cover the availability of the resources for a particular user (e. g. a preference for a general introductory book from the local library available for loan this weekend), user-specified amount of documents that should be presented (e. g. two new documents every Friday) and processing time requirements (the detailed analysis of a new bunch of documents will not be available until tomorrow morning).

Taking into account the given functions of ontologies in PortaGe, the following basic requirements on the ontology acquisition must be considered:

– Ideally, the process of ontology acquisition should run without any need of human assistance. On the other hand, the user must be able to influence the

learning, refine the extracted, select relevant information and modify the stored data manually.
- In general, the amount of the processed resources can be very high (thousands of documents). The implementation of the ontology learning must be computationally efficient and robust.
- The produced ontologies must reflect the stepwise development of the PortaGe system. If there is no current need for a particular kind of knowledge, the extraction (which often needs detailed analysis and is therefore resource demanding) should be postponed to later phases.

3 Architecture of the Ontology Extraction Framework

type of the relation	subject	object	relevance
used_for	SCFG	RNA secondary structure prediction	0.66
described_in	CKY algorithm	Cocke-Kasami-Younger	0.81
is_a	ribosomal frameshifting	RNA function	0.73
abbr_means	HMM	Hidden Markov Models	0.69
abbr_means	SCFG	Stochastic Context-Free Grammars	0.62
is_a	RNA	molecule	0.45
is_a	protein	molecule	0.45

Table 1. A fragment of a miniontology extracted from bioinformatics texts

OLE — the ontology acquisition framework described in this section has been developed in order to support the PortaGe project with instant ontological background. PortaGe ontologies are supposed to grow continuously when processing new resources provided by external tools.

The framework comprises several modules and related system components:

- OLITE module is responsible for processing the plain text resources (e. g. articles and conference papers from a given domain) and creating very simple ontologies from the extracted information. Presently, the relations are extracted according to specific patterns. However, any other method of information extraction can be easily incorporated as an independent plug-in.
- PALEA is the module responsible for learning of new semantic relations' patterns; the patterns are induced from the same resources as those used by OLITE. This component employs the methods described in [1] and [2] for learning new patterns.
- OLEMAN is intended to merge the outputs of the OLITE module — miniontologies — and update the PortaGe domain ontology with the resulting one. The uncertain information representation techniques [3] are used in this phase. Crisp ontology merging and alignment is based on the algorithms described in [4], [5], [6], [7] or [8]. Moreover, fuzzy ontology representation and

alignment framework is currently one of the main subjects of our intensive research.

The OLE parts are implemented as stand-alone modules. However, a server version is supposed to be developed for the final integration within the PortaGe project.

The OLITE module forms a crucial part of the entire system. The following paragraphs characterise the main processing steps performed by this component. The resources are first preprocessed by the subsystem. The main reasons for this are:

- the amount of input data must be reduced to its relevant subset only in order to increase the computational efficiency;
- at least some shallow syntactic structure must be imposed upon the reduced data before trying to extract the semantic relations.

The preprocessing must be as fast as possible, so no sophisticated (and time consuming) linguistic techniques, such as deep syntactic analysis, cannot be used. The input data are preprocessed in the following steps: sentence splitting, reduction to the relevant sentences only, sentence tokenization, POS tagging and lemmatization of the tokenized sentence, and chunking of the tagged sentences. We use our own custom preprocessing tools developed with support of NLTK toolkit (see [9]) instead of ready-made platforms (such as GATE, see [10]). This approach allows us to port the system easily for different languages, not only English. After the successful preprocessing, the extraction patterns are applied.

The OLITE module structure is devised so that it is able to adopt any extraction algorithm independently in the form of a specific plug-in. Such a plug-in is responsible for the concept extraction then, precise (or fuzzy) annotation by some class or property and passing of gained information further to the other parts of the module in order to build an output miniontology.

A fragment of the miniontology resulting from a test run of the extraction module is presented in Table 1. The semantic relations have been learned from a testing set of documents from the bioinformatics field. The relevance measure is computed by an algorithm inspired by C-value/NC-value method described in [11].

The extracted information is stored in a universal internal format that can be passed to the alignment module in order to be merged with the current ontology (also loaded in this format). The format is extensively expressive and universal with respect to efficient encoding of various relations and uncertainty representation[1]. The updated ontology (or even the output miniontology) file can be directly produced by applying translation rules. These rules are implemented as an independent plug-in (likewise the extraction algorithm itself) responsible for producing the output file in a desired format. Currently, the OWL DL format is supported, but OLITE is able to produce any other format this way (such as BayesOWL, see [3]).

[1] The research behind proposal and implementation of this format will be presented in another paper.

4 Related Work

The OLE project dissociates from the frameworks concentrated on facilitation of the manual (or expert-guided) ontology engineering activities, such as Protégé [12], WebODE [13] or OntoEdit [14]. The main reason is the infeasibility of the development and management of many different domain-specific ontologies needed for the full function of PortaGe.

Several automatic ontology acquisition systems have been developed in the last decade. One of them is OntoLT [15] implemented as a plug-in for the Protégé ontology editor. Its focus on the linguistic analysis for knowledge extraction is shared by our tool. However, our approach is able to extract deep semantic relations that seem to be out of scope of OntoLT. In this respect, PortaGe also differs from another ontology learning system — the Mo'K Workbench [16] based on clustering techniques for concept taxonomy building.

The OntoLearn [17, 18] and KnowItAll [1] systems incorporate the extraction of semantic relations, as well as we do. In KnowItAll, there is a notion of uncertainty introduced in the form of so called web-scale probability assessment to the extractions made, although it is not included in the ontology structure itself. On the contrary, the system proposed by T. T. Quan et al. in [19] deals with uncertain information implicitly and on the well defined fuzzy-logic basis. Their system is oriented to meta-information representation, which is supposed to be helpful when building scholarly semantic web. Our system attempts to represent the whole conceptual structure of a domain in an uncertain ontology. Such an ontology can be used for improvement of full-text search in the PortaGe portal documents, relevance measuring, resource categorisation and even for domain meta-information representation.

A different perspective of the uncertain information is present in Text2Onto [20] — a successor of the former TextToOnto [21, 22] system. The learned knowledge is represented at a meta-level within Probabilistic Ontology Model (POM). The independence brought by the use of POM is not necessary in our case as the output to other knowledge representation formalisms can be easily added in the form of plug-ins.

The OLE tools are designed as an open platform, which is easy to be amended by different extraction techniques or output modes of creation of ontologies. The pattern-based extraction of semantic relations is described in [23], [1], or [2]. The concept clustering techniques are introduced in the terascale knowledge acquisition efforts ([2]) and in [19] (fuzzy concept clustering). All these techniques can be easily adopted by the OLITE module to supplement the dynamic pattern learning and application (being under research within the PALEA module).

5 Conclusions and Future Directions

The ontology acquisition framework is presented in the context of automatic creation of web portals by means of the PortaGe system. The paper discussed the importance of ontologies for scientific portals. The preliminary results indicate

that the ontologies automatically extracted by the OLE system provide valuable resource of semantic data that are necessary for the function of PortaGe.

A lot of work still needs to be done on both the tools, the PortaGe system and the OLE tool. Our future research will focus on the design and implementation of advanced mechanism covering uncertainty in the acquired ontologies. We will also work on a qualitative evaluation of the scientific portals generated by PortaGe. They would be employed for example for e-learning of PhD students at our universities.

References

1. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Web-scale information extraction in knowitall: (preliminary results). In: WWW '04: Proceedings of the 13th international conference on World Wide Web, New York, NY, USA, ACM Press (2004) 100–110
2. P. Pantel, D. Ravichandran, E.H.: Towards terascale knowledge acquisition. In: Proceedings of Conference on Computational Linguistics (COLING-04). (2004) 771–777
3. Y. Peng, Z. Ding, R.P.: Bayesowl: A probabilistic framework for uncertainty in semantic web. In: Proceedings of Nineteenth International Joint Conference on Artificial Intelligence (IJCAI05). (2005)
4. Bouquet, P., Serafini, L., Zanobini, S.: Semantic coordination: a new approach and an application. In: ISWC 2003: Second International Semantic Web Conference. Proceedings, Springer-Verlag Berlin Heidelberg (2003) 130–145
5. Doan, A., Madhavan, J., Dhamankar, R., Domingos, P., Halevy, A.: Learning to match ontologies on the semantic web. The VLDB Journal **12** (2003) 303–319
6. Ehrig, M., Staab, S.: Qom - quick ontology mapping. In: ISWC 2004: Third International Semantic Web Conference. Proceedings. (2004) 683–697
7. Euzenat, J.: An api for ontology alignment. In: ISWC 2004: Third International Semantic Web Conference. Proceedings. (2004) 698–712
8. Widhalm, R., Mueck, T.A.: Merging topics in well-formed xml topic maps. In: ISWC 2003: Second International Semantic Web Conference. Proceedings, Springer-Verlag Berlin Heidelberg (2003) 64–79
9. : NLTK: Natural Language Toolkit – Technical Reports. (2005) Available at: http://nltk.sourceforge.net/tech/index.html.
10. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. (2002)
11. Frantzi, K.T., Ananiadou, S., Tsujii, J.: The c-value/nc-value method of automatic recognition for multi-word terms. In: Proceedings of Second European Conference ECDL '98, Springer-Verlag Berlin Heidelberg (1998) 585–604
12. Knublauch, H.: Ontology driven software development in the context of the semantic web: An example, scenario with protégé/owl. In: Proceedings of 1st International Workshop on the Model-Driven Semantic Web (MDSW2004). (2004)
13. Arpirez, J.C., Corcho, O., Fernandez-Lopez, M., Gomez-Perez, A.: Webode in a nutshell. AI Magazine **24** (2003) 37–47

14. Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., Wenke, D.: Ontoedit: Collaborative ontology development for the semantic web. In: ISWC 2002: First International Semantic Web Conference. Proceedings, Springer-Verlag Berlin Heidelberg (2002) 221–235
15. Buitelaar, P., Olejnik, D., Sintek, M.: OntoLT: A protégé plug-in for ontology extraction from text. In: Proceedings of the International Semantic Web Conference (ISWC). (2003)
16. Bisson, G., Nedellec, C., Canamero, L.: Designing clustering methods for ontology building - The Mo'K workbench. In: Proceedings of the ECAI Ontology Learning Workshop. (2000) 13–19
17. Gangemi, A., Navigli, R., Velardi, P.: Corpus driven ontology learning: a method and its application to automated terminology translation. IEEE Intelligent Systems (2003) 22–31
18. Velardi, P., Navigli, R., Cuchiarelli, A., Neri, F.: Evaluation of OntoLearn, a methodology for automatic population of domain ontologies. In Buitelaar, P., Cimiano, P., Magnini, B., eds.: Ontology Learning from Text: Methods, Applications and Evaluation. IOS Press (2005)
19. T. T. Quan, S. C. Hui, T.H.C.: Automatic generation of ontology for scholarly semantic web. In: ISWC 2004: Third International Semantic Web Conference. Proceedings, Springer-Verlag Berlin Heidelberg (2004) 726–740
20. Cimiano, P., Voelker, J.: Text2Onto — a framework for ontology learning and data-driven change discovery. In: Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB'05). (2005)
21. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intelligent Systems **16** (2001) 72–79
22. Maedche, A., Staab, S.: Ontology learning. In Staab, S., Studer, R., eds.: Handbook on Ontologies. Springer (2004) 173–189
23. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1992) 539–545