BUILDING A CONCEPT SHELL: ONTOLOGY POPULATION WITH FACTS FROM WWW

Inna Novalija, Dunja Mladenić
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773144
e-mail: inna.koval@ijs.si, dunja.mladenic@ijs.si

ABSTRACT

This paper addresses the process of the ontology population with facts for a selected domain of interest extracted from the Web documents. We suggest an information extraction methodology based on the ontology structural and lexical features. The preliminary evaluation is performed in the financial domain using Cyc ontology.

1 INTRODUCTION

This paper presents an approach to ontology population with facts extracted from the Web documents. The usage of the extended ontology for textual information analysis constitutes the primary motivation for our research.

The suggested method is used for inserting the new financial knowledge into Cyc [7], which maintains one of the most extensive common-sense knowledge bases worldwide.

In [14] we have presented a methodology for semi-automatic ontology extension using ontology content and ontology structure information, where ontology content of a particular ontology concept represents the available textual background of the referred concept and ontology structure includes neighborhood concepts involved in the hierarchical and non-hierarchical relations with a referred concept. We have also defined ontology extension [14] as a process allowing for adding new concepts to the existing ontology or, augmentation of the existing textual representation of the relevant concepts with new available textual information.

Following a goal of inserting new knowledge into the existing ontology, in this paper we address a process of ontology-based fact extraction, incorporating previously developed methodology for ontology extension into the ontology population process.

We define a **Concept Shell** as a building block of the ontology extension process. Concept shell aggregates all available information about a candidate ontology concept. The information is described at two layers – specification layer and instantiation layer. While structural information is defined by finding existing related ontology concepts and their relationships, factual information is obtained during ontology population.

For example, Cyc concept *CommercialOrganization* represents a subclass of concept *Organization*, whose primary goal is to generate a profit for its owners, usually through buying and selling of goods or services. A specification layer of *CommercialOrganization* includes a number of related Cyc concepts, such as:

- *OrganizationTypeByProfitMotive,*
- *BusinessRelatedThing,*
- *TelecommunicationsCompany,*
- *BankingOrFinanceCompany,*
- *ManufacturingOrganization et al.,*

and a number of Cyc relations, such as:

- *companyIsInIndustry,*
- *executiveVicePresident,*
- *enterpriseValue,*
- *mainBusinessActivityOfOrgOccursAt,*
- *organizationGrantsFranchisesOfType,*
- *companyHasGeneralCounsel et.al.*

The instantiation layer for CommercialOrganization includes instances: *FileMaker-CommercialOrganization, Symantec-CommercialOrganization, EpinionsDotCom, Thunderstone-theCompany, Adaptec-CommercialOrganization, WiredDigital-CommercialOrganization, Snap-CommercialOrganization, Amiga-CommercialOrganization, LycosInc, ElectronicArtsInc, Cross-jones, Tascon, StJudeMedical, PanasonicInc, EgyptTrans-GasCompany, Jodco-Japan, DeloreanTheCompany, PeugeotTheCompany, BMWTheCompany, Mercedes-BenzTheCompany, LamborghiniTheCompany, SubaruTheCompany, ChevroletTheCompany et.al., etc.*

The relation instances at the instantiation layer of the *CommercialOrganization* are the following:

- *(companyIsInIndustry MicrosoftInc* **(***IndustryOfRegionFn ComputerHardwareIndustry UnitedStatesOfAmerica***)**
- *(companyIsInIndustry MicrosoftInc* **(***IndustryOfRegionFn SoftwareIndustry UnitedStatesOfAmerica)***)**
- *(mainBusinessActivityOfOrgOccursAt KrispyKremeCorporation UnitedStatesOfAmerica)*
- *(mainBusinessActivityOfOrgOccursAt CVSCorp UnitedStatesOfAmerica)*

- *Time Interval : (TimeIntervalInclusiveFn (MonthFn November (YearFn 2008)) Now) Time Parameter : TimePoint (executiveVicePresident Nokia EskoAho)et al.*

The paper is structured as follows: Section 2 presents the related work; the methodology for ontology population is discussed in Section 3, Sections 4 describes the preliminary experiments and the results, the conclusion is covered in Section 5.

2 RELATED WORK

A number of approaches for automatic ontology population have proven themselves as effective tools of information extraction.

Natural language processing and unsupervised text mining are notably used for extending ontologies [15]. Extension of the existing ontology by automatically extending its relations was addressed by several researchers. The approaches include learning taxonomic [5]/non-taxonomic relations [12].

Described first by Hearst [10], the pattern based approach for instance and hyponym extraction uses a defined set of patterns while analyzing textual sources.

Etzioni et al. [8] developed a KnowItAll system for named entity classification. The approach performs pattern learning and can iteratively obtain new rules and new seeds.

The Open Information Extraction (OIE) paradigm for relation extraction from text was introduced by Banko et al. [3] and implemented in the TextRunner system. TextRunner performs self-supervised learning of a reliability classifier, single-pass extraction of tuples for all possible relations and redundancy-based assessing of probability for each trustworthy tuple.

Lexico-syntactic pattern-based ontology learning is handled by Text2Onto [6], a framework for ontology learning and data-driven change discovery.

SPRAT [13] is a tool for automatic semantic pattern-based ontology population. SPRAT system combines the name entity recognition, ontology-based information extraction and relation extraction in order to define patterns for the identification of a variety of entity types and relations between them.

Carlson et al. [4] present a method of coupling the semi supervised learning of category and relation instance extractors for ontology population with category and relation instances.

Several methods of the automatic ontology extension and population operate with enlarging of Cyc Knowledge Base (Cyc KB) [16], [17].

In our approach the available lexical and structural information of the large common-sense ontology is exploited for information extraction and validation. In particular, ontology-based patterns are used for extraction of relation instances from Web.

3 METHODOLOGY

In [14] we have proposed a new methodology for semi-automatic ontology extension, which combines text mining methods with user-oriented approach and supports the extension of multi-domain ontologies. Moreover, we have adapted the methodology in order to obtain an exhaustive specific methodology for Cyc knowledge base extension.

Experiments have been conducted in two domains – finances and fisheries & aquaculture. For the financial domain, we have used the Harvey [9] financial glossary, which contains around 6000 hyperlinked financial terms. The fisheries & aquaculture domain has been represented by the ASFA thesaurus [2], containing around 9900 terms involving several types of relationships: equivalence relationships, hierarchical relationships, associative relationships and notes.

As a part of the current research, we propose a method for ontology population with concept instances and relation instances extracted from the Web.

Basically, each ontology concept is represented as an information unit with all available information about it. The relationships the concept is involved in, the related existing ontology concepts, the concept instances and relation instances are combined in the concept shell.

The following steps are needed in order to define a concept specification and instantiation layers, which combine a concept shell:

1. *Identification of the Related Concepts*. When a new concept is added to the ontology, the methodology for semi-automatic ontology extension based on content and structure information [14] is used to find its hierarchically and non-hierarchically related concepts.

2. *Structural Inheritance.* From the superclasses the concept inherits its potential relationships.

3. *Concept Instances Search.* A number of patterns is used to search the Web for potential concept instances:
   "**c** such as **I**"
   "such **c** as **I**"
   "**c** including **I**"
   "**c**, especially **I**"
   "**c** like **I**"
   "**c** called **I**"
   "**I** is a **c**"
   "**I**, a **c**"

4. *Concept Instance Validation and Insertion.* Suggested concept instances are ontologically validated. In case no controversies are found, the new concept instances are inserted into the ontology.

5. *Relation Instances Search.* Patterns formed from the lexical and ontological information available about concept relations and their arguments are used to search the Web for potential relation instances.
   In our approach we assume that each ontological relation is represented by the relation denotation (in natural language) and argument types.
   For the methodology testing with Cyc we selected a list of fact raw types taking to the account BBN's proposed answer categories for question answering [1] and assuming that relation arguments are related to one of these types:

   - Quantity
   - Date&Time
   - Location
   - Product
   - Money
   - Event
   - ConceptualWork
   - Rate
   - Agent

   Using the ontology structure, it is possible to identify the relevant raw type for each argument of the particular relation, and therefore, apply a more efficient search procedure to find the relation instances.
   We form a relation instance pattern taking the relation denotation and extracted concept instance name. Sentences from text found on the Web, in which both relation denotation and a particular concept instance occur, are then extracted and checked for facts occurrence. For each argument type a specified technique is used. For instance, for Agent type we search for the relation name in the text and analyze the preceding and subsequent words.

6. *Relation Instances Validation and Insertion.* The extracted arguments are ontologically validated for each particular relation. The relation argument types are compared to the extracted argument types and validated relation instances are inserted into the ontology.

4 EXPERIMENTS & RESULTS

In order to evaluate the proposed methodology we conducted a simple fact extraction experiment in the financial domain using Cyc ontology.

Since Cyc knowledge base contains common sense knowledge [11], we assume that Cyc KB includes some financial knowledge - a financial knowledge base (Cyc FKB).

Figure 1 presents an extract from the concept shell for the new concept *EmergingMarket* added to Cyc. The light ovals represent the existing Cyc concepts, related to the new Cyc concept *EmergingMarket*. The hexagons are the relations (Cyc predicates). The structural ontological relations (such as superclass-subclass relation), that can be obtained in the ontology extension process, relate the new ontology concept to the existing ontology concepts. The relations, which operate with the instances of a particular concept, can be inherited by the new concept through the concept hierarchy. The dark ovals represent the new concept instances and the dark diamonds are the relation instances of the new Cyc concept.

Using the content and structure methodology for ontology extension, we identify that the new concept can be a subclass of the existing Cyc concept *Market*. Hence, the candidate concept *EmergingMarket* inherits a number of Cyc relations from its superclass concept. The binary predicate *hasMonopolyInMarket* relates instances of the *CommercialOrganization* with the market in which it has a monopoly.

With Hearst patterns [10] a number of potential instances for a concept *EmergingMarket* are extracted. One potential instance is *ChineseMarket.*

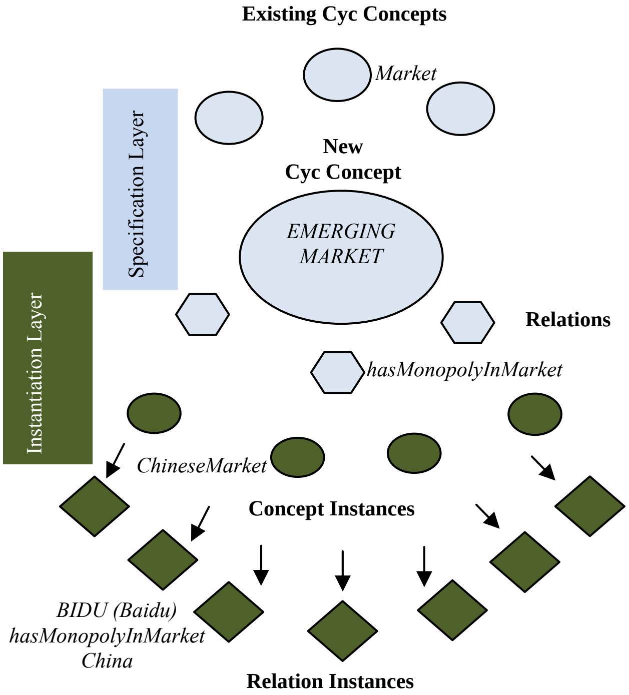

Figure 1: *Extract from Concept Shell for Concept "Emerging Market"*

With available information about the predicate *hasMonopolyInMarket*, we can form a relation instance pattern using lexical phrase *monopolies* from Cyc and instance name *Chinese market*. Looking at the argument types of the *hasMonopolyInMarket* predicate, we automatically define that we have to search for *CommercialOrganization* which is mapped to the raw type Agent.

As a result, we automatically extract and validate a relation instance: *BIDU (Baidu) hasMonopolyInMarket* in *Chinese Market.*

The results of the experiment confirm the applicability of the suggested methodology for ontology population to Cyc Knowledge Base augmentation.

6 CONCLUSION

This paper addresses the process of the ontology population with extracted facts for a selected domain of interest. We suggest an information extraction methodology based on the ontology structural and lexical features. The preliminary evaluation is performed in the financial domain using Cyc ontology.

The future work should include further extension and population of Cyc Knowledge Base and using it for sophisticated text analysis. Furthermore, the proposed methodology for ontology population should be tested on other domains.

6 ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under PASCAL2 (IST-NoE-216886) and ACTIVE (IST-2008-215040).

References

[1] Annotation guidelines for answer types: http://www.ldc.upenn.edu/Catalog/docs/LDC2005T33/BBN-Types-Subtypes.html

[2] ASFA thesaurus, http://www4.fao.org/asfa/asfa.htm

[3] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni. Open information extraction from the web. *Proc. of the 20th international joint conference on Artifical intelligence*, pages 2670–2676. 2007.

[4] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka Jr, T. M. Mitchell. Coupled Semi-Supervised Learning for Information Extraction. *Proc. of the Third ACM International Conference on Web Search and Data Mining (WSDM)*, volume 2, page 110. 2010.

[5] P. Cimiano, A. Pivk, L. Schmidt-Thieme, S. Staab. Learning Taxonomic Relations from Heterogeneous Evidence. *Proc. of ECAI Workshop on Ontology Learning and Population.* 2004.

[6] P. Cimiano, J. Völker. Text2Onto A Framework for Ontology Learning and Data-driven Change Discovery. *Proc. of NLDB 2005*, pp.227-238. 2005.

[7] Cycorp, Inc., http://www.cyc.com

[8] O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence*, 165(1):91-134. 2005.

[9] C. R. Harvey. Yahoo Financial Glossary, Fuqua School of Business, Duke University. 2003.

[10] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. *Proc of COLING*. 1992.

[11] D. Lenat. Cyc: A Large-Scale Investment in Knowledge Infrastructure. *Communic. of the ACM 38 (11).* 1995.

[12] A. Maedche, S. Staab. Discovering conceptual relations from text. *Proc. of ECAI 2000.*

[13] D. Maynard, A. Funk, W. Peters. SPRAT: a tool for automatic semantic pattern-based ontology population. *Proc. of International Conference for Digital Libraries and the Semantic Web*, Trento, Italy. 2009.

[14] I. Novalija, D. Mladenić. Content and Structure in the Aspect of Semi-Automatic Ontology Extension. *Proc. of the 32nd International Conference on Information Technology Interfaces.* 2010.

[15] T. Sabrina, A. Rosni, T. Enyakong. Extending Ontology Tree Using NLP Technique. *Proc. of National Conference on Research & Development in Computer Science REDECS.* 2001.

[16] P. Shah, D. Schneider, C. Matuszek, R. C. Kahlert, B. Aldag, D. Baxter, J. Cabral, M. Witbrock, J. Curtis. Automated population of Cyc: Extracting information about named-entities from the web. *Proc. of the Nineteenth International FLAIRS Conference*. 2006.

[17] M. Witbrock, D. Baxter, J. Curtis, D. Schneider, R. Kahlert, P. Miraglia, P. Wagner, K. Panton, G. Matthews, A. Vizedom. An Interactive Dialogue System for Knowledge Acquisition in Cyc. *Proc. of the Eighteenth International Joint Conference on Artificial Intelligence.* 2003.