EUREEKA: Deepening the Semantic Web by More Efficient Emergent Knowledge Representation and Processing*

Vít Nováček
DERI, National University of Ireland, Galway
IDA Business Park, Galway, Ireland
vit.novacek@deri.org

ABSTRACT

One of the major Semantic Web challenges is the knowledge acquisition bottleneck. New content on the web is produced much faster than the respective machine readable annotations, while a scalable knowledge extraction from the legacy resources is still largely an open problem. This poster presents an ongoing research on an empirical knowledge representation and reasoning framework, which is tailored to robust and meaningful processing of emergent, automatically learned ontologies. According to the preliminary results of our EUREEKA[1] prototype, the proposed framework can substantially improve the applicability of the rather messy emergent knowledge and thus facilitate the knowledge acquisition in an unprecedented way.

Keywords

emergent knowledge, knowledge representation

1. INTRODUCTION

The amount of data available on the web grows day by day. In order to be able to dive into the web content instead of merely floating on the rather shallow waters of respective meta-data annotations, machines have to get the knowledge from the content first. However, a scalable acquisition of expressive machine-readable knowledge from unstructured resources is a major challenge [1], mainly due to the fact that the manual knowledge acquisition is tedious, expensive and error-prone in most practical settings.

Ontology learning [2] from text, possibly combined with the emergent semantics principles [3] is therefore often considered as a viable first step towards a really deep Semantic Web. The automatically extracted emergent knowledge is dynamic, often explicitly annotated by uncertain measures of confidence [2], potentially inconsistent and incorrect, though. Approaches handling these features within traditional logics-based Semantic Web KR&R have been proposed in the literature. However, a practical, robust and general-purpose modelling covering the knowledge emerging on and from the web content is deemed to be hardly possible using these paradigms [5]. The main reasons for that are severe theoretical challenges, intractability, restricted applicability and low comprehensibility (i.e., inaccessibility for most users). Moreover, there are substantial conceptual peculiarities hampering using the prevalent logical KR&R for the emergent knowledge. The shallow structure of learned ontologies does not typically allow for many non-trivial logical conclusions. Potential incorrectness of the emergent facts (i.e., the empirical nature of truth) is awkward to be modelled by any logical knowledge representation (which by definition requires a categorical notion of truth, or at least its measure or degree in the uncertain generalisations of logics).

This poster abstract presents an informative overview of an alternative non-logical and empirical KR&R framework, which allows for more robust and efficient emergent knowledge processing and exploitation than possible with the current approaches. Full context of our work, rigorous formalisation of our approach and a comprehensive report on a preliminary evaluation are given in a recent technical report [4].

2. PROPOSED KR&R PRINCIPLES

We intend to show that the learned ontologies can be more efficiently exploited after coining a more appropriate alternative formal notion of semantics. In particular, such semantics tailored to the emergent knowledge should naturally support: (1), continuous soft refinement and integration of emergent data, resulting in an incremental formation of more complex and useful (i.e., non-trivial and correct) concepts; (2), easy user involvement in providing domain-specific extensions and/or constraints of the basic semantics; (3), full-fledged reasoning services superseding the logical inference.

As can be observed for instance in [2], the output of ontology learning techniques can mostly be reduced to various types of binary relations between lexical entities (e.g., taxonomical/subsumption relationships or general relations bootstrapped from the subject-verb-object frames in natural language sentences). These relations may possibly occur in a negative form (e.g., disjointness as a lack of mutual subsumption between two classes, or relations extracted from negative grammatical constructions). Finally, the ontology learning results usually come with a heuristically computed confidence measure. We propose a respective compact and convenient formal representation of emergent entities as matrices of real values in $[-1, 1]$, associated with unique identifiers (i.e., *subjects* in the RDF terminology). The row and column indices of a matrix correspond to the *property* and *object* identifiers, respectively. The values of the particular matrix elements present the degrees of certainty about the fact that the corresponding subject-property-object state-

*This work has been supported by the EU IST 6th framework's project 'Nepomuk' (FP6-027705) and by Science Foundation Ireland under Grant No. SFI/02/CE1/I131.

[1]A permutated acronym for ***E**asy to **U**se **E**mpirical **R**easoning about **A**utomatically **E**xtracted **K**nowledge.*

ment holds or does not hold when the degree is higher or lower than zero, respectively.

We define concept[2] change and aggregation functions [4], building on the ordered weighted averaging operators [6]. Moreover, we define similarity using the notion of parametrised metrics on the set of all concepts. These features support the evolution of empirical knowledge bases and also several useful inference services.

The reasoning is largely based on one foundational service – query answering. A query is technically a concept matrix. The answering process consists of checking a knowledge base[3] for similar answer matrices and returning the statements complementing the information in the query. Other inference services—e.g., a soft analogical extension or blending of concepts—can be directly based on the query answering.

To constrain the basic emergent semantics, it is possible to import a precise domain ontology as a trusted seed model into a knowledge base. This defines domain semantics refining the emergent knowledge while being incorporated using the concept change. However, one may wish to further specify the domain semantics on the fly, not only by using imported legacy ontologies. We enable this by introducing simple, yet quite expressive conjunctive IF-THEN rules. Both rule antecedents and consequents can be naturally translated into the respective concepts, which are to be unified according to the content of a knowledge base in order to instantiate the rule variables. The rule consequents are then combined with the unified antecedent content then, using the concept aggregation and change operators. Note that details on the inference services can be found in [4].

3. PRELIMINARY RESULTS

We have designed EUREEKA, a proof-of-concept Python library, which realises all the notions and services outlined here and properly elaborated in [4]. For an automated incorporation of learned ontologies (K_L) making use of a master legacy model (K_M) and a rule set (R), we implemented the following emergent knowledge processing pipeline:

1. for each concept C in K_L, incorporate C into K_M using the concept change operation, with parameters set according to the relative relevance of the K_L source to the master K_M ontology

2. compute the closure of updated K_M w.r.t. R

3. after fully updating/refining K_M using K_L, create a set $e(K_L)$, containing analogical extensions of the K_L content according to the updated K_M content; note that in the previous steps, the new and current knowledge has been provisionally linked and augmented, which allows for more productive analogy retrieval – therefore it is applied now and not already within the initial incorporation of the new concepts

4. repeat the steps 1. and 2. for the concepts in $e(K_L)$ in order to update K_M using the extensions

To test the pipeline, we conducted an experiment aimed at automated extension of the Gene Ontology (GO; see `http://www.geneontology.org/`). Into this seed legacy model, we integrated emergent knowledge extracted by the Text2Onto ontology learning tool (see `http://ontoware.org/projects/text2onto/`) from the GO natural language definitions (explanatory excerpts from relevant scientific resources). Rules specifying the semantics of the *isA*, *sameAs* and *partOf* relations—namely transitivity and (anti)symmetry—together with sample single-variable rules refining the learned data were employed.

We evaluated several types of query results on the integrated knowledge base. The results were compared with a similar baseline experiment employing RDFS-based KR&R. The comparison of our approach to the baseline results showed an average 291% improvement of the query answer quality. Moreover, the integrated knowledge contained relatively low number, approx. 5%, of correct, but trivial statements (the ratio was about 25% in the originally learned ontologies). The new knowledge volume in the eventual empirically integrated ontology was four times higher when compared to the baseline (about 456.000 to 113.000 statements). The results (see [4] for a comprehensive report) clearly show the promising potential of our approach even with the few rather simple proof-of-concept rules for the knowledge refinement and extension employed.

4. FUTURE WORK OUTLINE

Recently, we finished a preliminary proposal of the empirical KR&R framework and implemented the EUREEKA proof-of-concept prototype [4]. We have started to investigate rule expressivity amendments (primarily by introduction of intervals instead of single degree values in the conjuncts) and more complex inference services. We will also continue with reasoning optimisations and identify formal relationships between our approach and traditional KR&R paradigms. Another important step is to deliver mature, publicly available API and user interfaces, allowing for larger scale deployment and evaluation of the framework among real users.

5. REFERENCES

[1] S. Bechhofer et al. Tackling the ontology acquisition bottleneck: An experiment in ontology re-engineering, 2003. Retrieved at `http://citeseer.ist.psu.edu/bechhofer03tackling.html`, Apr 2008.

[2] P. Buitelaar and P. Cimiano. *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*. IOS Press, 2008.

[3] A. Maedche. Emergent semantics for ontologies. In S. Staab, editor, *Emergent Semantics*, IEEE Intelligent Systems, pages 85–86. IEEE Press, 2002.

[4] V. Nováček. Empirical KR&R in action: A new framework for the emergent knowledge. Technical Report DERI-TR-2008-04-18, DERI, NUIG, 2008. Available at `http://140.203.154.209/~vit/resources/2008/pubs/aerTR0408.pdf`.

[5] A. Sheth, C. Ramakrishnan, and C. Thomas. Semantics for the semantic web: The implicit, the formal and the powerful. *International Journal on Semantic Web & Information Systems*, 1(1):1–18, 2005.

[6] R. R. Yager. On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man and Cybernetics*, 18:183–190, 1988.

[2]We understand concepts in a bit less restricted way than usual – any entity is a concept, its classification as a a class, instance, relation, etc., is generally dependent on the empirical data and may change in time.

[3]Essentially a tuple comprising sets of concepts, identifiers and lexical expressions with mutual mapping functions.