Ollscoil na Gaillimhe
University of Galway

Provided by the author(s) and University of Galway in accordance with publisher policies. Please cite the published version when available.

Title	Invited Talk: Can We Deal with Emergent Knowledge Yet?
Author(s)	Nováek, Vít
Publication Date	2010
Publication Information	Vit Novacek, Invited Talk: Can We Deal with Emergent Knowledge Yet?", Proceedings of Znalosti 2010, VSE Prague, 2010.
Publisher	VSE Prague
Item record	http://hdl.handle.net/10379/1128

Downloaded 2024-03-13T07:27:57Z

Some rights reserved. For more information, please see the item record link above.

Can We Deal with Emergent Knowledge Yet?*

Vít Nováček

DERI, National University of Ireland, Galway
IDA Business Park, Galway, Ireland
E-mail: vit.novacek@deri.org

Abstract. This overview paper briefly describes problems we need to tackle if we want to meaningfully and efficiently process emergent knowledge. By this term we essentially mean knowledge continually emerging in a bottom-up manner from heterogeneous and possibly noisy resources in the context of the (Semantic) Web. We sketch a suggested solution for proper treatment of such knowledge, consisting of a novel light-weight knowledge representation framework. We also introduce current applications of our research, illustrating its practical applicability and promise for the future.

1 Introduction

In many practical scenarios encountered within the nascent Semantic Web, we have to deal with knowledge in the form of statements emerging in a bottom-up manner from multiple resources of varying relevance. The statements themselves may be noisy, uncertain (e.g., inconsistent, potentially incorrect or having an explicit certainty degree), and often have quite low expressivity from the formal point of view. Prominent examples of the scenarios we have in mind are:

- Ontology learning and population [2], where one has to reconcile relatively precise domain ontologies with their rather scruffy extensions being automatically extracted from text.
- Exploitation of social networks, i.e., mining of folksonomies from the collaborative tagging systems (see for instance [3]).
- Collaborative ontology development and knowledge engineering [4], where one has to integrate individual contributions of varying reliability into a common resulting model.

Each of these scenarios requires some means for appropriate representation, integration and processing of the emergent knowledge. There are certain substantial challenges these means should tackle, summarised along the following general aspects:

* Note that the rather informative content presented here largely stems from our more technical paper [1], which provides many details regarding both theoretical and implementation aspects of the outlined approach.

1. **Representation**: support for uncertainty and contextual features (e.g., provenance or time-stamp of emergent statements); extensible basic semantics for representation of the relatively simple structure of emergent knowledge, allowing also for gradual specification to incorporate more complex legacy models, though
2. **Processing**: inconsistency-tolerant aggregation of emergent statements, meaningful and scalable management of very large amounts of possibly noisy data
3. **Accessibility**: support for effortless involvement of lay users (i.e., domain, not AI or knowledge engineering experts) regarding editing or curating the emergent knowledge, constructing queries and interpreting their results
4. **Robustness**: intrinsic robustness – ability to *manage* sparse and noisy input, approximately *integrate* it with more precise and complex legacy and/or already processed content, and *exploit* it in a meaningful way; extrinsic robustness – ability to deal with incomplete, ambiguous or imprecise user queries in an efficient and meaningful manner

Approaches stemming from the current tradition in (applied) knowledge representation and reasoning, such as [5–13], provide particular solutions apt for coping with the challenges separately, however, to the best of our knowledge there is no off-the-shelf framework tackling all of them at once on a well-founded basis. Our long-term ambition is to provide an alternative light-weight, yet extensible solution, enabling emergent knowledge processing and evolution within a truly efficient continuous man-machine cooperation.

2 Solution Outline

In the following we informally outline the essential notions of the framework we propose in order to remedy the drawbacks of the traditional knowledge representation paradigms.

Central to our framework is a notion of *entities* that represent real and/or conceivable objects using unique identifiers and sets of positive or negative uncertain *relations* to other entities. To give an example, let us consider the d, a, c, t identifiers representing the `dog`, `animal`, `cat` concepts and the `type` relationship, respectively. The `dog` entity can be further specified by binary relations $t(d, a)$ and $t(d, c)$ with a positive and negative certainty, respectively, meaning dogs are animals different from cats. To support contextual features of entity relationships (e.g., provenance or time-stamp), the relations may generally have arbitrary arities. A direct correspondence of sets of n-ary certainty-valued relations to n-dimensional tensors (generalisations of the scalar, vector and matrix notions) provides for a compact computational representation of entities. An entity E is then represented as $(e, \mathbf{E})$, i.e., its unique identifier and the respective compact representation of uncertain relations to other entities. To ensure accessibility for lay users, we link the somewhat abstract representation to corresponding natural language referents via a set of *grounding* functions. These may map, for instance, the `dog` entity to a preferred "dog" expression with a high certainty, but also

to alternative synonyms like "doggy" or "hound", perhaps with a bit lower certainty. The other way around, a grounding would map the "mutt" word to the `dog` entity in the lexical domain of animals, but to a completely different entity in the domain of, say, humans. Thus the grounding provides a two-way bridge between the lexical (human-centric) and computational (machine-centric) aspects of the proposed lightweight semantics. The bridge is particularly important when answering user queries—formulated as mostly natural language statements—by means of a query answering service dealing with abstract entity representations.

Building on the compact computational representation of entities, we introduce the *aggregation* and *querying* services in order to tackle the remaining challenges specified in the introduction. Entity aggregation employs linear combinations that naturally model merging of possibly conflicting statements coming from sources with varying relevance. For instance, imagine a statement that dogs eat meat, coming from a highly relevant source, and an opposite, yet relatively irrelevant statement (vegetarian dogs actually exist, however, the respective rather exceptional sources are presumably less relevant). The sum of the corresponding representations, weighed by the relative source relevance, will result in a claim that dogs eat meat with a positive, but slightly lower certainty (as the knowledge from more relevant source prevails in the aggregation).

Query answering makes use of two notions of entity similarity. Let us imagine entities of `dog` and `cow`, eating and not eating meat, respectively. Evaluation of a query for meat-eating animals first checks for entities fitting to the context of the query, i.e., being animals and linked by an "eat" relation to meat. Both `dog` and `cow` entities fit the query within this coarse-grained approximation of similarity. A finer grained notion of similarity, taking the certainty degrees into account, can be naturally coined as dual to a distance defined on the set of entity representations. Utilising this type of similarity results into meat-eating `dog` being a much more certain answer to the query than `cow`, which is an animal, but does not eat meat. In more complex cases, we also sort the query results according to their relevance employing a generalised IR measure based on numbers of outgoing and incoming relations among stored entities.

3 Preliminary Implementations

The theoretical principles of emergent knowledge representation and processing we outlined in the previous section[1] have been recently reflected in EUREEKA, a prototype knowledge store and inference engine. So far it has been employed in two different practical scenarios, as summarised in the remainder of this part.

3.1 CORAAL

CORAAL (`http://coraal.deri.ie:8080/coraal/`) is a comprehensive life science publication search engine deployed on the data provided by Elsevier within

[1] Note that we expand the outline by a much more rigorous and explanatory description in [1].

their Grand Challenge contest (`http://www.elseviergrandchallenge.com/`). EUREEKA forms the engine's crucial back-end part, catering for the *representation*, *integration* and *exposure* tasks, thus enabling the knowledge-based search functionalities.

For the initial knowledge extraction in CORAAL, we used a NLP-based heuristics stemming from [14, 15] in order to process chunk-parsed texts into subject-predicate-object-score quads. The scores were derived from absolute and document frequencies of subject/object/predicate terms aggregated with subject/object co-occurrence measures. If a relation's score is not available for any reason (e.g., when importing legacy knowledge from crisp resources instead of extracting it from text), we simply set it to 1 (or -1) in the implementation. The extracted quads encoded three major types of ontological relations between concepts: (i) taxonomical—*type* or *same as*—relationships; (ii) concept difference (i.e., negative *type* relationships); and (iii) "facet" relations derived from verb frames in the input texts (e.g., *has part*, *involves* or *occurs in*). We imposed a taxonomy on the latter, considering the head verb of the respective phrase as a more generic relation (e.g., *involves expression of* was assumed to be a type of *involves*). Also, several artificial relation types were introduced to specify the semantics of some most frequent relations. Namely, (positive) *type* was considered transitive and anti-symmetric, and *same as* is set transitive and symmetric. Similarly, *part of* was assumed transitive and being inverse of *has part*.

After the initial knowledge extraction in CORAAL, EUREEKA comes into play in order to integrate the emergent statements, link them to precise domain thesauri and expose them to users via intuitive approximate querying. Example queries and selected top answer statements are (answer certainties in brackets):

- *Q: ? : type : breast cancer* $\rightsquigarrow$ ***cystosarcoma phylloides*** *TYPE breast cancer (1);*
- *Q: rapid antigen testing : part of : ? AND ? : type : clinical study* $\rightsquigarrow$ ***dicom study*** *USE protein info (0.8),* ***initial study*** *INVOLVED patients (0.9).*

The examples abstract from the result provenance, however, full-fledged presentation of answers to the above or any other queries can be tried live with CORAAL at `http://coraal.deri.ie:8080/coraal/`, using the *Knowledge* tab or the guided query builder. For a more comprehensive description of the CORAAL system, see our recent article [16].

3.2 TWEAKR

Another deployment of EUREEKA we are currently working on is TWEAKR – a *Text-to-WikipEdiA linKeR*. It is a simple application that links an input text (a newspaper article, scientific paper, blog entry or even an e-book) to related articles in Wikipedia. The main purpose is to provide a bit more context to readers by automatically linking texts to relevant content of the world's largest and most up-to-date encyclopedia. In addition to mere lists of related articles, TWEAKR offers also particular statements linking the input text and Wikipedia knowledge at more fine-grained entity level.

The linking is enabled by an underlying EUREEKA knowledge base containing statements extracted from the textual content of Wikipedia articles, which are incorporated into a relevant seed model – the YAGO ontology [17]. For each input text, significant noun phrases are identified. Entities corresponding to these noun phrases are then retrieved from the EUREEKA knowledge base and extended according to a set of rules (we are currently using slightly generalised RDFS entailment rules [18]). This results in a "local closure" of the entities present in the input text. The provenance information of the statements in the closure is then directly used for generating a ranked list of related Wikipedia articles, taking the statements' relevance and certainty into account. The service is supposed to be more useful for users than traditional methods applicable to the text linking problem, as it reflects the semantics of the actual texts (which is not the case of state of the art technologies like vector-space models or classifier-based approaches).

4 Conclusion and Outlook

We have briefly outlined a new way of dealing with emergent knowledge aimed at filling the gap in the current state of the art in the (applied) knowledge representation field. Besides giving a general overview of the problems in question and sketching an alternative approach, we have also introduced respective preliminary implementations. Recent experiments with our prototypes indicate a promising potential of the proposed solution. This has been demonstrated not only by the results presented for instance in [1, 16], but also by our recent successful participation in the Elsevier Grand Challenge contest (cf. `http://www.elseviergrandchallenge.com/winners.html`).

Regarding short-term future goals, we are going to extend the user-centric query language by contexts and release the extended EUREEKA implementation as an open source module. In a longer term perspective, we have to investigate import of more complex ontologies into EUREEKA – so far we have covered only rather simple RDFS-like semantics. We also intend to provide means for distributed implementation of the principles introduced here in order to scale the framework up to arbitrarily large data.

Our general vision of the future development along the outline presented here is about bringing well-founded, yet practically applicable, large-scale knowledge representation and automated reasoning to masses. We believe this is possible only if the knowledge acquisition will be as effortless as possible. Perhaps the most straightforward way how to achieve this is combining automated knowledge extraction with intuitive means for individual contributions of domain experts. An efficient and universal enough framework for representation and processing of knowledge emerging this way could eventually lead towards an unprecedented open-ended evolution of knowledge jointly acquired, processed and utilised by humanity and machines. And providing such a framework is essentially the ultimate goal of the basic research presented here.

Acknowledgments The work presented or referenced here has been supported by the 'Líon II' project funded by SFI (Science Foundation Ireland) under Grant No. SFI/08/CE/I1380. We would like to acknowledge the contributions of Tudor Groza and Siegfried Handschuh, who have played an indispensable role in the CORAAL prototype development. Last but not least, we are thankful to Elsevier, B.V., for an access to their data stores and for their general support related to our participation in the Elsevier Grand Challenge competition.

References

1. Nováček, V., Decker, S.: Towards lightweight and robust large scale emergent knowledge processing. In: Proceedings of ISWC'09, Springer (2009)
2. Buitelaar, P., Cimiano, P.: Ontology Learning and Population: Bridging the Gap between Text and Knowledge. IOS Press (2008)
3. Jäschke, R., Hotho, A., Schmitz, C., Ganter, B., Stumme, G.: Discovering shared conceptualizations in folksonomies. J. Web Sem. **6**(1) (2008) 38–53
4. Tudorache, T., Noy, N.F., Tu, S.W., Musen, M.A.: Supporting collaborative ontology development in protégé. In: International Semantic Web Conference. (2008) 17–32
5. Schueler, B., Sizov, S., Staab, S., Tran, D.T.: Querying for meta knowledge. In: Proceedings of WWW 2008, ACM (2008)
6. Mazzieri, M.: A fuzzy RDF semantics to represent trust metadata. In: Proceedings of SWAP'04. (2004)
7. Bobillo, F., Straccia, U.: fuzzyDL: An expressive fuzzy description logic reasoner. In: In Proceedings of FUZZ-08. (2008)
8. Hartig, O.: Querying Trust in RDF Data with tSPARQL. In: ESWC'09. (2009)
9. Kiefer, C., Bernstein, A., Stocker, M.: The fundamentals of isparql: A virtual triple approach for similarity-based semantic web tasks. In: ISWC/ASWC. (2007)
10. Udrea, O., Deng, Y., Ruckhaus, E., Subrahmanian, V.S.: A graph theoretical foundation for integrating RDF ontologies. In: Proceedings of AAAI'05. (2005)
11. Alani, H., Brewster, C., Shadbolt, N.: Ranking ontologies with AKTiveRank. In: Proceedings of ISWC'06. (2006)
12. Oren, E., Guéret, C., Schlobach, S.: Anytime query answering in RDF through evolutionary algorithms. In: Proceedings of ISWC'08. (2008)
13. Banko, M., Etzioni, O.: The tradeoffs between open and traditional relation extraction. In: Proceedings of ACL-08: HLT, ACL (2008) 28–36
14. Maedche, A., Staab, S.: Discovering conceptual relations from text. In: Proceedings of ECAI 2000, IOS Press (2000)
15. Voelker, J., Vrandecic, D., Sure, Y., Hotho, A.: Learning disjointness. In: Proceedings of ESWC'07, Springer (2007)
16. Nováček, V., Groza, T., Handschuh, S., Decker, S.: CORAAL – dive into publications, bathe in the knowledge. Journal of Web Semantics (2009) In press.
17. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A large ontology from wikipedia and wordnet. Journal of Web Semantics **6**(3) (2008) 203–217
18. Brickley, D., Guha, R.V.: RDF Vocabulary Description Language 1.0: RDF Schema. (2004) Available at (Feb 2006): `http://www.w3.org/TR/rdf-schema/`.