Approximation and self-organisation on the Web of Data

Christophe Guéret, Kathrin Dentler, Stefan Schlobach

VU University Amsterdam, 1081HV Amsterdam

Abstract

The Web of Data (WoD) is a network connecting billions of facts hosted by many different parties, represented using a variety of vocabularies with varying degrees of preciseness, and supplied in an inconsistent fashion, altogether yielding a high level of messiness. This WoD is growing at an amazing rate and it will no longer be feasible to deal with it in a global way, by centralising the data or reasoning processes making use of that data. Currently, techniques coming from research on databases are at the core of applications dealing with the WoD. However, considering its increasing size and messiness, there is a need to shift to approximated and decentralised algorithms. We believe that Computational Intelligence and collective intelligence techniques provides the approximation, adaptiveness, robustness and scalability that will be required to exploit the full value of ever growing amounts of dynamic data on the Web.

This short paper is a summary of the paper "Linked Data Meets Computational Intelligence", published at the AAAI spring symposium 2010 "Linked Data Meets Artificial Intelligence" [2]

1 The Web of Data

Using Semantic Web technologies such as RDF and SPARQL, people have lately been able to publish a massive amount of semantically enriched information. The published data is expressed as the association between a subject, a predicate and an object (a "triple"). By sharing subjects and objects, these triples gets connected to each other, thereby creating a knowledge network made of related facts. This Web of Data (WoD) is taking momentum with more than 15 Billions of triples already created with some contribution, for instance, from Facebook and the UK government. Currently, most of the techniques used to make use of the WoD comes from research done on databases. It often means that in order to be used, the data must be first gathered from different places to be stored into a single data store. Soon, this will no longer be a reasonable approach considering that:

- The number of triples and triple providers is growing. Data stores have become increasingly efficient and can now deal with billions of triples in a single store but the number of triple providers is also growing steadily. In plus, incoherence and conflicts likely to be found when combining heterogeneous data. Both the exponential increase of figures and the risk of conflicts when merging data poses a serious threat on centralised approaches.

- Centralising data rises privacy issues. Providing access to some data or giving away an copy of it are two different stories. While triple providers are, by definition, willing to provide access to their data some of them may not be keen on giving away entire copies of it. For instance, social network web sites provide access to part of the data they host but will typically not allow anyone to dump everything the know into one place.

- Centralisation implies consolidation. The predicates used to compose the triples are defined in user created vocabularies. Although some, like FOAF or Dublin Core, are very popular, two different data sets are likely not to use the same vocabulary. Before being merged and queried as a unique information base, all the triples have to be consolidated in order to use the same vocabulary.

- Data consumers are actually not interested in how the data is stored or processed. The recent developments on cloud computing highlighted the fact that users like to have the data somewhere and, more importantly, always accessible. There is less demand for single access point to data sets.

There is thus a need for the research community to work on techniques able to deal with decentralised, huge and incoherent amount of data and data providers. New algorithms designed to leverage knowledge from the WoD will have to be scalable, robust, anytime and adaptive. These properties are among the key features offered by computational intelligence algorithms and collective intelligence techniques.

2 Approximated and self-organising algorithms

Computational intelligence is a research field compromising work done on fuzzy systems, neural networks and evolutionary computation. While the usage of the former as already made it's way through databases, and by extension the Semantic Web, the later is still hardly considered. The self-organising behaviour of collective intelligence algorithms as also not been sufficiently investigated yet, although self-* properties would fit the decentralised context provided by the WoD. In more details, and considering the challenges posed by the WoD, these particular advantages are of interest:

- Simplicity and interactivity: techniques such as evolutionary algorithms and swarm intelligence only requires the (careful) design of interaction rules. The final result then emerges from the activity of the system. Besides, and because they can be run continuously, these algorithms can cope with changing parameters and stream back results as they are found. This breaks the classical loop of "query,wait,result,query" user actions.

- Learning, adaptation and approximation: programs are able to cope with changes and learn as they are executed. Approximated results can be returned as performances improve over time.

- Scalability, robustness and parallelism: population-based algorithms such as swarm intelligence and evolutionary algorithms are easier to parallelise as every individual is a unique, independent, entity. Robustness is also improved as the loose of a single entity doesn't rhyme with a complete failure of the system.

As to illustrate the relevancy of these techniques for the WoD and to give the reader an idea of "what's going on", here are a few examples of successful usage of CI and swarm techniques:

- Semantic gossiping is a technique making use of self-organisation to establish peer-to-peer semantic equivalences among vocabularies used by different triple providers [1]. This avoids the need of a central translation point.

- The brood sorting behaviour of ants can be used to design an efficient triple store that finds the most efficient way to distribute triples over a set of storage areas [3].

- In [4], an evolutionary algorithm used to solve a query answering problem turned into an optimisation problem has been proved to be able to stream "good enough" answers to queries.

The work on approximation algorithms and self-organisation for dealing with the WoD is an emergent research field with a nascent research community working on it. For more references, the interested reader is invited to read [2].

References

[1] Philippe Cudré-Mauroux. *Emergent semantics*. PhD thesis, EPFL, Lausanne, 2006.

[2] Christophe Guéret. Linked Data Meets Computational Intelligence - Position paper. In *Linked Data meets Artificial Intelligence, AAAI Spring Symposium*, Stanford, CA, 2010.

[3] Sebastian Koske. Swarm approaches for semantic triple clustering and retrieval in distributed rdf spaces. Technical report, Freie Universitat Berlin, February 2009.

[4] Eyal Oren, Christophe Guéret, and Stefan Schlobach. Anytime query answering in rdf through evolutionary algorithms. In *7th International Semantic Web Conference (ISWC2008)*, October 2008.