Editor: Charles Petrie • petrie@stanford.edu

Unifying Reasoning and Search to Web Scale

Dieter Fensel • *University of Innsbruck, Austria*
Frank van Harmelen • *Vrije Universiteit Amsterdam*

We recently heard about a telecom project that required reasoning about 10 billion RDF triples (statements of the form <subject, relation, object>) in less than 100 ms. The use case was defined around generating revenue streams through new context-sensitive and personalized mobile services. Existing approaches can handle Resource Description Framework Schema (RDFS) queries for roughly 100 million triples in 100 ms, but this project required sophisticated reasoning with a set of triples that's two orders of magnitude larger — and the requirements will certainly grow. Indeed, scale requirements could increase much faster over time than any progress in reasoning algorithms, clever coding, and improved hardware can compensate.

Being forced to turn away potential customers led us to wonder why this problem even existed. Problems usually become intractable through improper conceptualization — asking intelligence to introduce assumptions that make the problem solvable, on the one hand, without restricting them on the other hand to ensure usefulness. So the question is: Why isn't reasoning scaling for the Web and how can this be fixed?

The Contradiction of Web and Reasoning

The Web and reasoning started to meet around 1996 with the first projects that added semantics to Web page descriptions in much the same way that HTML added formatting information. Despite the subsequent growth of these Semantic Web efforts into a dynamic and well-established research area,[1] serious doubt remains whether reasoning really adds something useful in the Web context.

Researchers have developed reasoning methods for rather small, closed, trustworthy, consistent, and static domains. They usually provide a small set of axioms (together with various facts — a special type of axiom); a proof engine can typically provide complete and correct inferences of the knowledge contained in them. Take natural numbers as an example: the seven so-called Peano axioms (one of which is an axiom schema that actually represents countably many axioms) can characterize all the relevant knowledge about them. Interestingly enough, as Gödel's famous incompleteness theorem proves, no complete and correct inference method can be developed for this simple logical theory. As a result, the past 50 years have witnessed serious efforts to find efficient inference methods for computationally less complex logics (that is, logics that can't code all the relevant knowledge about natural numbers). Description logic, for example, restricts the logical language in such a way that decidable procedures can be found and implemented for inferring deductive closure for a set of axioms. Another well-known example is logic programming, which takes the Horn fragment of first-order logic and reasons only about a specific model (a kind of minimal model) rather than all of them. Both approaches have their merits. DL reasoners can deal with 10^5 axioms (so-called concept definitions), but they scale poorly for large instance sets. Logic programming engines can deal with similar-sized rule sets as well as larger instance sets (say, 10^6), but they can draw only simple logical conclusions from these theories.

Both streams are highly interesting areas of research, and open topics such as how to combine them attract a lot of attention. Still, we doubt whether this is the actual path to reasoning on a Web scale, which involves working with arbitrarily vast numbers of triples — "frillions," to use a colloquialism. After all, what are 10 billion triples in the end? A conservative estimate would be that it would take 10,000 triples just to describe each human, which gives us 100 trillion. The mismatch

continued on p. 94

 1089-7801/07/$25.00

continued from p. 96

is deeper than efficient reasoning algorithms over restricted subsets of first-order logic can resolve. To clarify, let's revisit the underlying assumptions of logic:

- *Small set of axioms*. Given that describing the natural numbers already requires countably many axioms, the Web is quite unlikely to require much less. If the Web is to capture the entirety of human knowledge, the number of axioms could end up being very large.
- *Small number of facts*. Assuming the Google count of roughly 30 billion Web pages and a modest estimate of 100 facts per page, we're already in the space of a trillion facts.
- *Completeness of inference rules*. The Web is open, with no defined boundaries. Therefore, completeness is a rather strange requirement for an inference procedure in this context. Given that collecting all relevant information on the Web is neither possible nor often even desirable (usually, you want to read the first 10 Google hits but don't have the time for the remaining two million), requesting complete reasoning on top of such already heavily incomplete facts seems meaningless.
- *Trustworthiness, correctness of inference rules, and consistency*. Traditional logic takes axioms as reflecting truth and tries to infer the implicit knowledge they provide. This procedure's correctness ensures that the truth captured by the axioms is preserved. In a Web context, information is unreliable from the beginning, which means even a correct inference engine can't guard truth; worse yet, any proof engine will simply infer a contradiction because the Web provides the space for "knowledge" expressing different viewpoints.
- *Static domains*. The Web is a dynamic entity: the known facts will change during the process of acquiring them and using them for inference. Traditional notions of complete and correct reasoning are obviously based on a heavily simplified world view naively applied to reality. It's a well-known insight that any knowledge about the state of large and distributed systems is either incomplete or outdated (that is, incorrect).

Traditional notions of complete and correct reasoning are obviously based on a heavily simplified world view naively applied to reality.

Spoken cynically, current reasoning engines have inherited clumpy syntax from the Web (XML, RDF, and URIs), and in return, the Web has received toy engines that neither meet its requirements nor scale to its size. Basically, both sides have been aligned at a level too superficial to generate something useful. The basic underlying assumptions of pure logical reasoning don't seem to match the reality the Web provides. An analogous mismatch in a different area of science might provide a way to resolve this problem.

Reasoning with Limited Rationality is Truly Rational

Classical economic theory assumes completely rational agents — that is, that agents base their decisions on complete information of the market and infer optimal choices from it. One the one hand, this leads to interesting equation systems with certain mathematical properties, but on the other, it models a groom who must go on roughly four billion dates before seriously considering marriage. Despite being somewhat mathematically interesting because of their rigidness in terms of global optima, these theories have limited power to predict or model reality. Collecting information and reasoning with it is actually a process bounded by limited resources.

Herbert Simon introduced the concept of *limited rationality* in 1957 to better model these processes.[2] In this view, agents make decisions based on incomplete knowledge and might lack the resources to draw all potential conclusions (the latter is of limited sense if the former is notoriously incomplete, anyway). This approach created a new research area around heuristic problem solving.

Our aim is to effect a similar paradigm shift by integrating reasoning and search at Web scale. Rather than abstract notions of completeness and correctness, our proposed paradigm, which we call "reasearch" for the moment, employs a more concrete idea of usability in the context of actual problem solving. It's also based on the idea that current reasoning, which is agnostic about properly reflecting the effort and resources required for acquiring and processing information, models truly irrational behavior.

Fusing Reasoning and Search

As explained before, 100 trillion triples and more will certainly require incomplete and incorrect reasoning based on prioritizing information and a combination of stochastics and logic. The basic idea is to select a random sample of any number of triples and reason with them. This method scales to any

size. A slightly more intelligent approach is to try to improve the triple selection through preprocessing to find the important ones for reasoning. We could describe such a research algorithm as follows:

```
do
  draw a sample,
  do the reasoning on the sample;
  if    you have more time,
        and/or if you don't
        trust the result,
  then draw a bigger sample,
repeat
```

We could then attempt to cleverly select the sample based on

- known distribution properties for the triples,
- their relationship to the query,
- provenance properties such as reputation or trust, and
- experiences with previous queries.

Such algorithms don't just scale better than classical algorithms, they scale to any order of magnitude by simply trading in quality (for example, by basing the reasoning on a proportionally smaller sample). In some of our early work in this area, we've had encouraging results using the normalized Google distance[3] — a function that measures how close word x is to word y on a zero-to-infinity scale — as a heuristic for drawing samples from a large (and globally inconsistent) knowledge base.

Logical reasoning is currently agnostic regarding where the facts and axioms stem from. We assume they provide truth and thus apply truth-preserving reasoning to them, but this neither fits nor scales in a Web context. The only way to bridge the divide is to interweave the reasoning process with the process of establishing the relevant facts and axioms through retrieval (ranking or selection) and abstraction (compressing information). That way, retrieval and reasoning become two sides of the same coin — a process that aims for useful information derived from data on the Web.

Areas of related work include data compression in fields such as computer graphics, machine learning, and data warehousing, as well as anytime and approximate reasoning.[4] Work toward such algorithms will have to provide answers on the following questions (among others):

- What are probabilistic notions of entailment, consistency, and so on?
- What are desirable properties of such inferencing? Some possibilities include repeatability (If you do it twice, do you get the same answer?), monotonicity (If you take a larger sample, do you get a better answer?), and anytime availability (What trade-offs exist between computation time and answer quality?).
- What kind of query language, and answers, are needed?
- Can the triples self-organize under the influence of past inference tasks, so that selecting relevant triples becomes easier for future tasks?

Working on these issues promises to actually integrate logic and the Web.

Rather than adding bizarre syntax to our languages or nonscalable logic to superficially align Web principles and reasoning, we seek to reflect on the underlying principles, exclude those that don't fit, and merge the remainder in something new that reflects proper unification. Tim Berners-Lee and colleagues might have had the same goal in mind when they discussed alternative ways of reasoning.[5]

This stream of research fits well with the European Union's new research framework program VII, which asks for projects related to "semantic foundations: probabilistic, temporal and modal modeling and approximate reasoning through objective-driven research moving beyond current formalisms. Theoretical results will be matched by robust and scalable reference implementations." (http://cordis.europa.eu/fp7/ict/) It's needed, and it will be funded!

Acknowledgments

We thank Michael Kiefer, Atanas Kiryakov, and Charles Petrie for very helpful discussions and look forward to elaborating our preliminary ideas with them.

References

1. *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*, D. Fensel et al., eds., MIT Press, 2003.
2. H. Simon, *Models of Man*, Wiley, 1957.
3. R. Cilibrasi and P. Vitanyi, "Automatic Meaning Discovery Using Google," tech. report, 2004; www.arxiv.org/abs/cs.CL/0412098.
4. S. Russell and E.H. Wefald, *Do the Right Thing: Studies in Limited Rationality*, MIT Press, 1991.
5. T. Berners-Lee et al., "A Framework for Web Science," *Foundations and Trends in Web Science*, vol. 1, no. 1, 2006, pp. 1–130.

Dieter Fensel is a professor at the University of Innsbruck, Austria. His research interests include the development of a semantic access layer to data and processes. Fensel has a PhD in artificial intelligence from the University of Innsbruck. He is coeditor of *Enabling Semantic Web Services: The Web Service Modeling Ontology*, (Springer, 2006) and author of *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*, (Springer-Verlag, 2001, 2003). Contact him at dieter.fensel@deri.org.

Frank van Harmelen is a professor at the Vrije Universiteit Amsterdam. His research interests include knowledge representation on the Web and approximate reasoning techniques. Van Harmelen has a PhD in artificial intelligence from the University of Edinburgh. He is coauthor of *The Semantic Web Primer* (MIT Press), the first textbook on the Semantic Web. Contact him at Frank.van.Harmelen@cs.vu.nl; www.cs.vu.nl/~frankh.