PerAdaMagazine
towards pervasive adaptation
10.2417/2200908.1766

Self-organizing ants find new paths to scalability

Robert Tolksdorf

The principles of locality and decentralization found in nature could be the key to managing Web data traffic, as shown in initial experiments with a semantic store.

A most astonishing observation about the Internet is its growth. The amount of e-mail exchanged, websites available and multimedia resources transmitted is seemingly growing at a huge rate without any noticeable deterioration of the user's experience. While the number of fibre channels installed provides sufficient bandwidth for traffic, organizing the massive data flow leads to serious problems that call current paradigms into question.

Our research focuses on the so-called Semantic Web. Here, information about other information is notated in standardized formats, the most basic being a 'triple'. It carries a statement such as "Prof. Miller authored a book" in the form of ⟨*subject, predicate, object*⟩. By referring to things of interest using Unique Resource Identifiers (URIs), the same example could be reformulated as ⟨*http://miller.org/me, http://preds.org/written, isbn:123-4-56789-123-4*⟩. Triples can number in the trillions. For example, the RDF (Resource Description Framework) representation of Wikipedia, DBpedia,[1] is currently approaching the 300 million triples mark.

The triples go into a 'triplestore', of which several implementations exist, including a small variety of commercial products. Aside from storing triples, triplestores allow queries on the information they contain. These are usually formulated in SPARQL, a query language designed for the Semantic Web. One can ask, for example, which books were authored by Prof. Miller.

But the task can become more complicated. Assume that the following additional triples exist: ⟨*isbn:123-4-56789-123-4, http://preds.org/isabout, http://science.org/Java*⟩ (the book is about Java) and ⟨*http://science.org/Java, rdfs:is-a, http://science.org/prog_lang*⟩ (Java is a programming language). If we now ask who wrote books on programming languages, 'Prof. Miller' will be part of the answer. A Semantic-Web triplestore can infer such results for a query because the 'is-a' relation is standardized.

What is happening inside the triplestore at this point? It executes some algorithms that look at triples and check whether they might trigger an inference rule. Since the is-a relation is predefined so that Java can be a result whenever the query is for a programming language, such a rule can enter the book on Java into a result set, in which books are then resolved to their authors.

To do so efficiently, the store must have access to all of the triples involved. If there is only one store that knows all the information, the task is simple. But in reality, the triples might be distributed over various locations. This makes sense since there is a strict bandwidth limit of the machine on which the store is running, which dictates an upper bound for its performance. To provide a scalable solution, we have adapted an existing[2] design for self-organizing data stores to the management of semantic information. We consider a triplestore as a huge and distributed collection of servers. Putting 'similar' triples on the same server facilitates fast retrieval of matches to a query.

Abandoning any kind of central organization scheme (such as a global hash function), we rely instead on principles from ant-colony optimization, especially the mechanism of brood sorting. Here, ants carry a triple and walk around in a landscape formed, in our case, by servers. At each halt (i.e., each stop in their walk around the 'server landscape') they check whether the location already contains similar triples. If so, the current load is dropped and the ant dies. If not, the ant decides which direction to take by looking at the environment. Similarly, an ant looking for a triple that matches a partial description—e.g., ⟨*isbn:123-4-56789-123-4, http://preds.org/isabout, ?o*⟩ to find out what the book is about—walks around and looks for matches at the current node. If none is found, it examines the surrounding nodes and chooses a path, leaving a scent on the current node. When it eventually finds a match, the scent trail serves as a guide for other ants.

We have experimented with several different notions of similarity. For example, we used only the syntactic structure relating all information whose URI begins with the identical fragment *http://science.org/*.[3] We also tried a scheme in which the ants have a partial view of relations, such as with the Java-is-a-programming-language example above, and can determine similarity based on such semantic relations.[4]

By implementing simulators for the algorithms and similarity configurations, we can measure the system's behaviour and

Continued on next page

show that its entropy is significantly reduced. This means that clusters with similar RDF information evolve, which in turn makes it easier to find triples. Because we introduce no centralization, we are able to build scalable, distributed semantic stores. We are currently beginning a real-world project that will apply the approach to a large-scale semantic storage service for a geoinformation system.

In summary, we have shown that principles of self-organization inspired by nature can lead to scalable systems in the field of Semantic-Web data management. The success of this approach depends on strictly avoiding any kind of global data or decision taking. We believe that these principles will be widely adopted in future Internet systems. In addition to real-world implementation of the concepts, we are extending them using a decentralized reasoning mechanism.

Author Information

Robert Tolksdorf
Networked Information Systems
Freie Universität Berlin
Berlin, Germany
http://www.ag-nbi.de

Robert Tolksdorf graduated in computer science from the Technische Universität Berlin and was appointed as a professor at the Freie Universität Berlin in 2002. His areas of work and interest are semantic technologies, coordination and self-organization.

References

1. http://wiki.dbpedia.org/About DBpedia overview.
2. Ronaldo Menezes and Robert Tolksdorf, *A new approach to scalable Linda systems based on swarms*, **Proc. ACM Symp. Appl. Comput.**, pp. 375–379, 2003.
3. Robert Tolksdorf and Anne Augustin, *Self-organisation in a store for semantic information*, **J. Software 4**. In press.
4. Sebastian Koske, *Swarm approaches for semantic triple clustering and retrieval in distributed RDF spaces* Tech. Rep. B-09-04B, FU Berlin, Institut für Informatik, 2009.