Nature Inspired Self-Organizing Semantic Storage

Robert Tolksdorf, Hannes Mühleisen, Kia Teymourian, Marco Harasic, Anne Augustin

Freie Universität Berlin
Department of Computer Science - AG Networked Information Systems
Königin-Luise-Str. 24/26, 14195 Berlin, Germany
tolk@ag-nbi.de, {muehleis,kia,harasic,aagusti}@inf.fu-berlin.de
http://www.ag-nbi.de

Abstract. Traditional approaches for semantic storage and analysis are facing their limits on the handling of enormous data amounts of today's applications. We believe that a more radical departure from contemporary architectures of stores is necessary to satisfy that central scalability requirement. One of the most promising new schools of thought in system design are swarm intelligent and swarm-based approaches for data distribution and organization. In this paper, we describe our current work on a swarm-based storage service for Semantic Web data.

1 Introduction and Related Work

Semantic applications share the need for an efficient and scalable storage service which can handle huge amounts of semantic data. The performance and scalability level of the storage services will be defined by use case scenarios, e.g. the future Semantic Web applications need to scale to the size of the Web and the Internet network, respectively.

Conventional approaches for distributed storage services are facing complex problems in scaling and their adaptivity to changes in network infrastructure, both requirements for large-scale semantic applications. Thus new concepts and architectures for distributed storage have to be developed, and a more radical departure from contemporary architectures of storage might be the key to the realization of scalable storage systems. One of the most promising approaches for data distribution are swarm intelligent and swarm-based algorithms.

While some central storage systems support replication of their stored data, they rely on a central instance orchestrating the execution of storage and query requests. Distributed storage systems on the other hand should not rely on central nodes, as they pose single points. In contrast, distributed semantic storage systems have been proposed such as RDFPeers [1] or GridVine [3].

2 The Self-Organized Semantic Storage Service

In SwarmLinda [4], a distributed coordination system based on swarm algorithms was proposed. This coordination model makes use of a global tuple space. In SwarmLinda this tuple space is distributed to a network of nodes. The different operations are realized by using swarm algorithms to reach a high level of scalability and adaptivity to network changes which are both important properties for open distributed systems. SwarmLinda clusters tuples that match the same template and trails of virtual pheromones are left in the system to make these clusters traceable. Our SwarmLinda implementation has been extended in [5] adopting ant colony algorithms to realize a distributed storage for RDF triples. Both approaches aimed at clustering semantically related RDF triples.

Our concept is to build a Self-Organized Semantic Storage Service (S4) which uses swarm-based algorithms to store the user provided semantic data into diverse clusters potentially spanning multiple nodes. This structures can later be exploited for efficient data retrieval. We have adapted the SwarmLinda concept and its basic algorithms in order to enable it to store Semantic Web meta data in the data model RDF. Consistent with SwarmLinda, virtual ants move over a landscape consisting of a number of nodes (servers) which are interconnected.

One of the basic requirements of the ant algorithms is a similarity metric used to calculate the relative similarity between triples. Previous approaches have for example either employed string-based distance measures or more complex metrics based on ontologies. Any metric is required to yield a relative value for any pair of triples, thus making data organization and efficient clustering possible. However, our approach was deliberately designed to allow an easy replacement of the similarity metric, so this specific challenge is not detailed further.

The S4 concept based on the described algorithms offers a wide range of benefits compared to other approaches: Swarming individuals (ants) are controlled by simple algorithms, they do not require a complex rule set to perform well, additionally, all decisions can be made using only a local view. Still, individuals are able to dynamically adapt to their possibly changing environment. Network organization thus is decentralized and robust to changes in the network topology. In contrast to hash-based approaches such as DHT or B-Tree our approach does not require costly network reconstruction (achieved through communication) in the case of an error. Every node has sufficient information in the form of present scents to execute every possible query on its own, hence further eliminating single points of failure. Feasible solutions exist for issues like over-clustering, where a skewed data distribution leads unfair distribution of data storage [2]. The same is true for hot point avoidance: if a node stores triples used in a large number of requests, it can simply move parts of them to another node or reject storage of similar triples in the future.

3 Conclusion and Outlook

We have described our motivation to extend the current state of storage systems for Semantic Web applications. The differences between centralized and distributed semantic storage services have been shown, as well as the current research in storage systems. We have outlined our design of a Self-Organized Semantic Storage Service (S4), and advantages over conventional data distribution algorithms.

While reasoning on central storage systems can be performed using logic engines, distributed reasoning is far more complex. As a general approach, achieving reasoning completeness is sacrificed in favor of scalability for distributed reasoning. Consequently, the next level for S4 is to design and implement a concept for a simple reasoning during retrieval along an is-a hierarchy and with best-effort guarantees. We then plan to add application-specific reasoning capabilities. We will continue working on the refinement and implementation of our S4 concept within our current joint research project “DigiPolis”, where this system will form the basis for a indoor navigation system possibly covering entire cities. Example use cases include a semantic search for points of interest in a vicinity, in house routing with semantic restrictions, and semantic annotation of map entries.

Acknowledgments This work has been supported by the German Federal Ministry of Education and Research under grant number 03WKP07B.

References

1. Min Cai and Martin Frank. RDFPeers: A scalable distributed RDF repository, February 27 2004.
2. Matteo Casadei, Ronaldo Menezes, Robert Tolksdorf, and Mirko Viroli. On the problem of over-clustering in tuple-based coordination systems. In *SASO*, pages 303–306. IEEE Computer Society, 2007.
3. Philippe Cudré-Mauroux, Suchit Agarwal, and Karl Aberer. GridVine: An infrastructure for peer information management. *IEEE Internet Computing*, 11(5):36–44, 2007.
4. Ronaldo Menezes and Robert Tolksdorf. A new approach to scalable linda-systems based on swarms. In *Proceedings of ACM SAC 2003*, pages 375–379, 2003.
5. Robert Tolksdorf and Anne Augustin. Selforganisation in a storage for semantic information. *Journal of Software*, 4, 2009.