Semantic Gossiping: Coping with Heterogeneous Semantic Knowledge Management Systems in the Large

Karl Aberer and Philippe Cudré-Mauroux
School Of Computer and Communication Sciences
EPFL, Lausanne, Switzerland
{karl.aberer, philippe.cudre-mauroux}@epfl.ch

Coping with heterogeneous systems in the large

With the creation and wide adoption of Semantic Web standards like RDF or OWL, a new breed of semantic networks is about to appear. For the first time, we can expect large numbers of knowledge management systems to interoperate using common languages to express the semantics of the data they share or seek. We however foresee semantic heterogeneity to surface once more as a key problem in information integration, given the scale and variety of the systems we are dealing with: Indeed, we realistically cannot expect common global upper-ontologies to capture with sufficient adequacy the requirements of all the different parties in our heterogeneous environment. Custom ontologies will be developed for various application needs, thus endangering global semantic interoperability by introducing local concepts and properties. Also, the situation is somewhat complicated by the fact that ontologies will not be static in such environments, but will tend to evolve, appear or disappear dynamically as systems join and leave the network.

Cleary, there is a need and potential to develop new semantic integration techniques here, since traditional approaches can neither cope with the scale, nor with the dynamicity of such environments. Observing that fostering semantic interoperability requires some form of agreement or consensus among information sharing parties we focus on mechanisms for establishing such agreements in the large. Unlike traditional approaches, like database schema integration or ontology engineering, we do not require the agreements to be static, global or even accurate. Rather, we consider the existence of mutual local agreements only, and expect global semantic properties to emerge from the continuous interactions of autonomous entities in a self-organizing manner. Establishing semantic interoperability in the large can be understood as studying the dynamics of a complex, self-organizing system. Local communication and decision-making of information sharing agents generate the dynamics of this system. When agents locally reach acceptable agreements that are as consistent as possible with the information they receive, the global system reaches a state that embodies what we call the global semantic agreement. Since this state emerges from the dynamics of a complex system, we call the resulting global semantic agreement also the *emergent semantics* of the semantic interoperability system.

Semantic Gossiping as a new semantic reconciliation technique

Following the principles outlined above, we have developed a concrete approach that we termed Semantic Gossiping [1] where we obtain semantic interoperability in a bottom-up, semi-automatic manner without relying on global semantic models. We assume that some local agreement (e.g., ontology mappings) exists between systems using different ontologies to model their data. Fig. 1 below represents a network where seven distinct semantic systems are related by various semantic translation links. We call such networks *semantic overlay networks*. Requiring some initial consensus between pairs of systems is not that stringent, given the recent developments on automatic and semi-automatic schema and ontology alignment techniques.

Semantic gossiping realizes now sharing of local knowledge on translations by using the principle of gossiping that has been successfully applied for creating useful global behaviors in decentralized environments: When different schemas are involved, local mappings are used to further distribute a search request into other semantic domains. Inferring agreements from transitive closures on local translation links, we can relate systems that would have been semantically disconnected otherwise. By comparing the initial search query with the search queries forwarded through series of translation links (*transformed* query), we characterize the quality of the agreements obtained in this manner in two ways:

- We iteratively determine a *syntactic similarity* measure that accounts for the net loss of information (e.g., attributes which cannot be mapped) between the original and the transformed query
- We determine *semantic similarity* values that reflect the degree of semantic agreement (e.g., precision of attribute mappings) that can be achieved by two systems given a translation link.

Whereas syntactic similarity values can be computed locally by comparing a given query with its transformed version, semantic similarity values have to be derived from the context provided by the semantic overlay network. Thus, we do not depend on any hypothetical local expertise to assess the correctness or precision of the mappings, but rather derive these values by aggregating evidences throughout the network. One possibility to realize this is to analyze cycles in the translation graph that allow us to compare any initial query with a syntactically similar query sent through series of translation links and returning to the originator of the query (e.g., a query issued by *A*, going through *B* and *C* and returning to *A* in the figure below). Another possibility is to analyze the results returned by the systems to which the query was forwarded to determine the end-to-end semantic gap between two semantically heterogeneous systems.

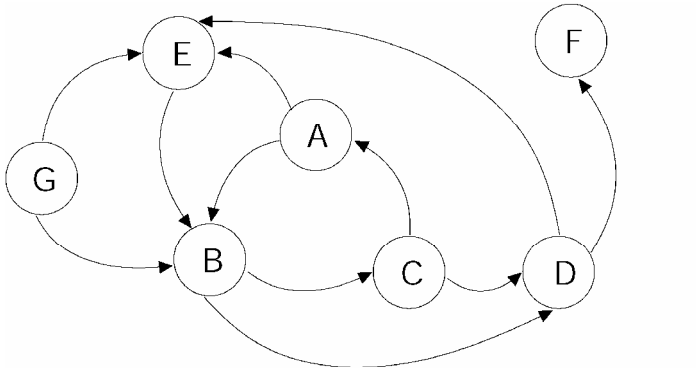

Fig. 1: A network of translation links mapping semantically heterogeneous systems

At a global level, we can view the problem as follows: The translations between domains of semantic homogeneity form a directed graph. Peers in the network compute syntactic and semantic similarity values whenever they want a forward a query to a heterogeneous semantic domain using a translation link and receive from forwarded queries feedback that in turn is used to improve the assessment of semantic similarity. The decision on whether or not it is useful to forward a query to a given system is dependent on the similarity values obtained (*per-hop* forwarding behaviour). In [2], we showed how such heuristics can be applied to maximize recall while limiting the total number of messages generated to answer a given query. Implicitly, this is a state where a global agreement on the semantics of the different schemas or ontologies has been reached. Furthermore, we devised techniques to automatically improve the quality of pre-existing translations based on the results of the similarity computations (*self-healing* semantic networks, see [3]).

A down-to-earth system: GridVine

To demonstrate our approach, we implemented Semantic Gossiping in a concrete system called *GridVine*[4]. In GridVine, we address the problem of building scalable semantic overlay networks by following the principle of data independence and separating the logical from the physical layer (see Fig. 2): At the logical layer, we support various operations necessary for the maintenance and use of a semantic overlay network, including attribute-based search, schema management, schema inheritance and schema mapping. We provide these mechanisms within the standard syntactic framework of RDF/OWL. At the physical layer we provide efficient realizations of the operations exploiting a structured DHT overlay network, namely P-Grid [5] which supports efficient location of resources in a network based on resource identifiers.

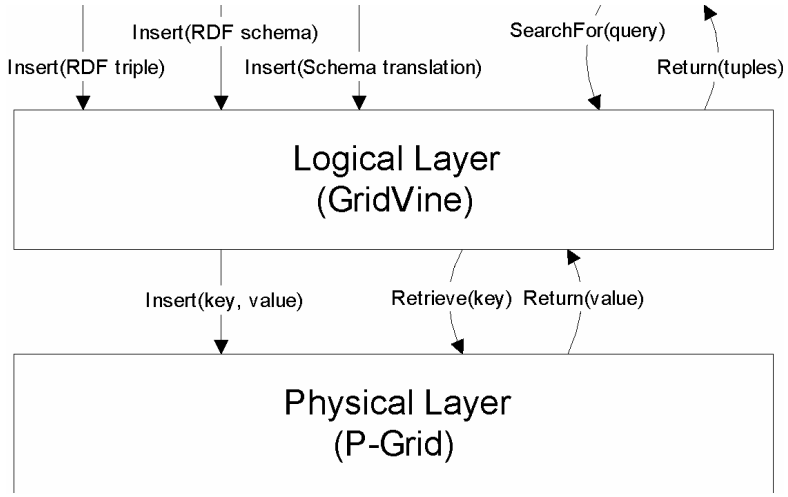

Fig. 2: GridVine: applying the principle of Data Independence

The separation of a physical from a logical layer allows us to process logical operations in the semantic overlay using different physical execution strategies. In particular we identify iterative and recursive strategies for the traversal of semantic overlay networks as two important alternatives. At the logical layer, we support semantic interoperability through schema inheritance and Semantic Gossiping. To the best of our knowledge, GridVine is the first semantic overlay network based on a scalable, efficient and totally decentralized access structure supporting the creation of local schemas while fostering global semantic interoperability.

Conclusion

Studying semantics in the context of large-scale systems will lead to a new breed of approaches and systems to tackle the inherently difficult problem of semantic interoperability. In particular, in these systems the "network" provides a rich source of knowledge that can be processed in a decentralized and automated manner. We can expect that tools for studying complex systems, such as graph theory and dynamical systems theory, will start to play an important role in better understanding the global behavior of such systems and will open exciting avenues for future research.

References:

[1] *A Framework for Semantic Gossiping,*
Karl Aberer, Philippe Cudré-Mauroux, Manfred Hauswirth
SIGMOD Record, 31(4), December 2002

[2] *The Chatty Web: Emergent Semantics Through Gossiping,*
Karl Aberer, Philippe Cudré-Mauroux, Manfred Hauswirth
Proceedings of the Twelfth International World Wide Web Conference (WWW2003), 20-24 May 2003, Budapest, Hungary.

[3] *Start making sense: The Chatty Web approach for global semantic agreements*,
Karl Aberer, Philippe Cudré-Mauroux, Manfred Hauswirth,
Journal of Web Semantics, 1 (1), December 2003

[4] *GridVine: Building Internet-Scale Semantic Overlay Networks,*
Karl Aberer, Philippe Cudré-Mauroux, Manfred Hauswirth, Tim van Pelt,
Third International Semantic Web Conference (ISWC), Hiroshima, Japan, 2004.

[5] *P-Grid: A Self-organizing Structured P2P System*
Karl Aberer, Philippe Cudré-Mauroux, Anwitaman Datta, Zoran Despotovic, Manfred Hauswirth, Magdalena Punceva, Roman Schmidt
SIGMOD Record, 32(2), September 2003