Collective Knowledge Composition in a Peer-to-Peer Network

A survey for Peer-to-Peer applications

Boanerges Aleman-Meza, Chris Halaschek, I. Budak Arpinar
LSDIS Lab, Computer Science Department,
University of Georgia

INTRODUCTION

Today's data and information management tools enable massive accumulation and storage of knowledge that is produced through scientific advancements, personal and corporate experiences, communications, interactions, etc. In addition, the increase in the volume of this data and knowledge continues to accelerate. The willingness and the ability to share and use this information are key factors for realizing the full potential of this knowledge scattered over many distributed computing devices and human beings. By correlating these isolated islands of knowledge, individuals can gain new insights, discover new relations (Sheth, Arpinar, and Kashyap, 2003), and produce more knowledge. Despite the abundance of information, knowledge starvation still exists because most of the information cannot be used effectively for decision-making and problem solving purposes. This is in part due to the lack of easy to use knowledge sharing and collective discovery mechanisms. Thus, there is an emerging need for knowledge tools that will enable users to collectively create, share, browse and query their knowledge.

For example, many complex scientific problems increasingly require collaboration between teams of scientists who are distributed across space and time and who belong to diverse disciplines (Loser, Wolpers, Siberski, and Nejdl, 2003; Pike, Ahlqvist, Gahegan, and Oswal, 2003). Effective collaboration remains dependent, however, on how individual scientists (i.e., peers) can represent their meaningful knowledge, how they can query and browse each others' knowledge space (knowledge map), and, most importantly, how they can compose their local knowledge pieces together collectively to discover new insights that are not evident to each peer locally.

A common metaphor for knowledge is that it consists of separate little factoids, and that these knowledge "atoms" can be collected, stored, and passed along (Lakoff, and Johnson, 1983). Views like this are what underlie the notion that an important part of knowledge management is getting access to the "right knowledge."

While the state of the art is not at the point where we can duplicate the accomplishments of a Shakespeare or Einstein on demand, research developments allow us to craft technological and methodological support to increase the creation of new knowledge, both by individuals and by groups (Thomas, Kellogg, and Erickson, 2001).

A Peer-to-Peer (P2P) network can facilitate scalable composition of knowledge compared to a centralized architecture where local knowledge maps are extracted and collected in a server periodically to find possible compositions. This kind of vision can be realized by exploiting advances in various fields. Background and enabling technologies include (i) semantic metadata extraction and annotation, (ii) knowledge discovery and composition. Figure 1 shows these components for an ontology-based P2P query subsystem.

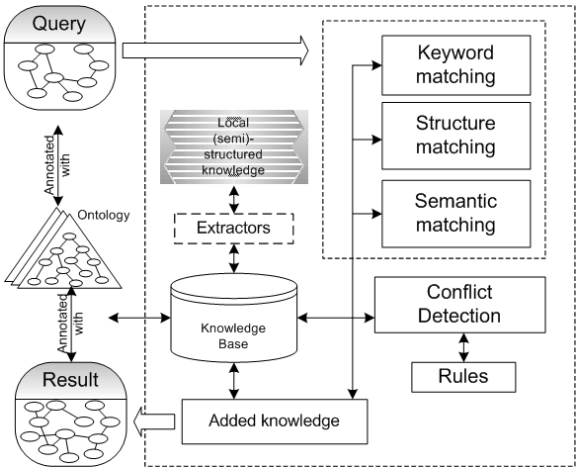

Figure 1. Part of an ontology-based P2P query subsystem

B A C K G R O U N D

Semantic Metadata Extraction and Annotation. A peer's local knowledge can be in various formats such as: Web pages (unstructured), text documents (unstructured), XML (semi-structured), RDF or OWL, etc. In the context of efficient collective knowledge composition, this data must be in a machine processable format, such as RDF or OWL. Thus, all data that is not in this format must be processed and converted (metadata extraction). Once this is completed, the knowledge will be suitable to be shared with other peers. The Semantic Web envisions making content machine processable, not just readable or consumable by the human beings (Berners-Lee, Hendler, and Lassila, 2001). This is accomplished by the use of ontologies which involve agreed terms and their relationships in different domains. Different peers can agree to use a common ontology to annotate their content and/or resolve their differences using ontology mapping techniques. Furthermore, peers' local knowledge will be represented in a machine processable format, with the goal of enabling the automatic composition of knowledge.

Ontology-driven extraction of domain-specific semantic metadata has been a highly researched area. Both semi-automatic (Handschuh, Staab, and Studer, 2002) and automatic (Hammond, Sheth, and Kochut, 2002) techniques and tools have been developed, and significant work continues in this area (Vargas-Vera, et al., 2002).

Knowledge Discovery and Composition. One of the approaches for knowledge discovery is to consider relations in the Semantic Web that are expressed semantically in languages like RDF(S). Anyanwu and Sheth (2003) have formally defined particular kinds of relations in the Semantic Web, namely, Semantic Associations. Discovery and ranking of these kinds of relations have been

addressed in a centralized system (Sheth, et al., 2004; Aleman-Meza, Halaschek, Arpinar, and Sheth, 2003). However, a P2P approach can be exploited to make the discovery of knowledge more dynamic, flexible, and scalable. Since different peers may have knowledge of related entities and relationships, they can be interconnected in order to provide a solution for a scientific problem and/or to discover new knowledge by means of composing knowledge of the otherwise isolated peers.

In order to exploit peers' knowledge, it is necessary to make use of knowledge query languages. A vast amount of research has been aimed at the development of query languages and mechanisms for a variety of knowledge representation models. However, there are additional special considerations to be addressed in distributed dynamic systems such as P2P.

P E E R - T O - P E E R N E T W O R K S

Recently there has been a substantial amount of research in P2P networks. For example, P2P network topology has been an area of much interest. Basic peer networks include random coupling of peers over a transport network such as Gnutella (http://www.gnutella.com) (discussed by Ripeanu, 2001) and centralized server networks such as that of Napster (http://www.napster.com) architecture. These networks suffer from drawbacks such as scalability, lack of search guarantees, and bottlenecks. Yang and Garcia-Molina (2003) discussed super-peer networks that introduce hierarchy into the network in which super-peers have additional capabilities and duties in the network that may include indexing the content of other peers. Queries are broadcasted among super-peers, and these queries are then forwarded to leaf peers. Schlosser, Sintek, Decker and Nejdl (2002) proposed HyperCup, a network in which a deterministic topology is maintained and known of by all nodes in the network. Therefore, nodes at least have an idea of what the network beyond their scope looks like. They can use this globally available information to reach locally optimal decisions while routing and broadcasting search messages. Content addressable networks (CAN) (Ratnasamy, Francis, Handley, Karp, and Shenker, 2001) have provided significant improvements for keyword search. If meta-information on a peer's content is available, this information can be used to organize the network in order to route queries more accurately and for more efficient searching. Similarly, ontologies can be used to bootstrap the P2P network organization: peers and the content that they provide can be classified by relating their content to concepts in an ontology or concept hierarchy. The classification determines, to a certain extent, a peer's location in the network. Peers routing queries can use their knowledge of this scheme to route and broadcast queries efficiently.

Peer network layouts have also combined multiple ideas briefly mentioned here. In addition, Nejdl et al. (2003) proposed a super-peer based layout for RDF-based P2P networks. Similar to content addressable networks, super-peers index the metadata context that the leaf peers have.

Efficient searching in P2P networks is very important as well. Typically, a P2P node broadcasts a search request to its neighboring peers who propagate the request to their peers and so on. However, this can be dramatically improved. For example, Yang and Garcia-Molina (2003) have described techniques to increase search

effectiveness. These include iterative deepening, directed Breadth First Search, and local indices over the data contained within *r*-hops from itself. Ramanathan, Kalogeraki, and Pruyne (2001) proposed a mechanism in which peers monitor which other peers frequently respond successfully to their requests for information. When a peer is known to frequently provide good results, other peers attempt to move closer to it in the network by creating a new connection with that peer. This leads to clusters of peers with similar interests that allow to limit the depth of searches required to find good results. Nejdl et al. (2003) proposed using the semantic indices contained in super-peers to forward queries more efficiently. Yu and Singh (2003) proposed a vector-reputation scheme for query forwarding and reorganization of the network. Tang, Xu and Dwarkadas (2003) made use of data semantics in the pSearch project. In order to achieve efficient search, they rely on a distributed hash table to extend LSI and VSM algorithms for their use in P2P networks.

FUTURE TRENDS

Knowledge composition applications are fundamentally based on advances in research areas such as information retrieval, knowledge representation and databases. As the growth of the Web continues, knowledge composition will likely exploit pieces of knowledge from the multitude of heterogeneous sources of Web content available. The field of peer-to-peer networks is, as of now, an active research area with applicability as a framework for knowledge composition. Given our experiences, we believe that future research outcomes in peer-to-peer knowledge composition will make use of a variety of knowledge sources. Knowledge will be composed from structured data (such as relational databases), semi-structured data (such as XML feeds), semantically annotated data (using the RDF model or OWL), and necessary conversions will be done using knowledge extractors. Thus, knowledge will be composed from databases, XML, ontologies, and extracted data. However, the more valuable insights will probably be possible by combining knowledge sources with un-structured Web content. Large scale analysis and composition of knowledge exploiting massive amounts of Web content remain challenging and interesting topics.

CONCLUDING REMARKS

The problem of collectively composing knowledge can greatly benefit from research in the organization and discovery of information in P2P networks. Additionally, several capabilities in creating knowledge bases from heterogeneous sources provide the means for exploiting semantics in data and knowledge for knowledge composition purposes. In this respect, we have discussed the evolution of peer-to-peer systems from a knowledge composition perspective. Although challenging research problems remain, there is great potential for moving from centralized knowledge discovery systems towards a distributed environment. Thus, research in databases, information retrieval, semantic analytics, and P2P networks provides the basis of a framework in which applications for knowledge composition can be built.

R E F E R E N C E S

Aleman-Meza, B., Halaschek, C., Arpinar, I. B., & Sheth, A. (2003). *Context-Aware Semantic Association Ranking*. Paper presented at the First International Workshop on Semantic Web and Databases, Berlin, Germany.

Anyanwu, K., & Sheth, A. (2003). *r-Queries: Enabling Querying for Semantic Associations on the Semantic Web*. Paper presented at the Twelfth International World Wide Web Conference, Budapest, Hungary.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). *The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*. Scientific American, 284(5), 34-+.

Hammond, B., Sheth, A., & Kochut, K. (2002). Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content. In V. Kashyap & L. Shklar (Eds.), *Real World Semantic Web Applications* (pp. 29-49): Ios Pr Inc.

Handschuh, S., Staab, S., & Studer, R. (2003). Leveraging metadata creation for the semantic Web with CREAM. *KI 2003: Advances in Artificial Intelligence*, 2821, 19-33.

Lakoff, G. & Johnson, M. (1983). *Metaphors We Live By*, University of Chicago Press, Chicago, IL.

Loser, A., Wolpers, M., Siberski, W., & Nejdl, W. (2003). *Efficient Data Store and Discovery in a Scientific P2P network*. Paper presented at the ISWC 2003 Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, Sanibel Island, Florida.

Nejdl, W., Wolpers, M., Siberski, W., Schmitz, C., Schlosser, M., Brunkhorst, I., & Löser, A. (2003). *Super-peer-based routing and clustering strategies for RDF-based peer-to-peer networks*. Paper presented at the 12th International World Wide Web Conference, Budapest, Hungary.

Pike, W., Ahlqvist, O., Gahegan, M., & Oswal, S. (2003). *Supporting Collaborative Science through a Knowledge and Data Management Portal*. Paper presented at the ISWC 2003 Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, Sanibel Island, Florida.

Ramanathan, M. K., Kalogeraki, V., & Pruyne, J. (2001). *Finding Good Peers in Peer-to-Peer Networks*. HP Labs, Technical Report HPL-2001-271.

Ratnasamy, S., Francis, P., Handley, M., Karp, R., & Shenker, S. (2001). *A scalable content addressable network*. Paper presented in *ACM SIGCOMM*.

Ripeanu, M. (2001). *Peer-to-Peer Architecture Case Study: Gnutella Network*. Paper presented at the International Conference on Peer-to-Peer Computing, Linkoping, Sweden.

Schlosser, M., Sintek, M., Decker, S., & Nejdl, W. (2003). HyperCuP–Hypercubes, Ontologies and Efficient Search on P2P Networks. *Lecture Notes in Artificial Intelligence*, 2530, pp. 112–124.

Sheth, A., Aleman-Meza, B., Arpinar, I. B., Halaschek, C., Ramakrishnan, C., Bertram, C., Warke, Y., Avant, D., Arpinar, F. S., Anyanwu, K., & Kochut, K. (2004). Semantic Association Identification and Knowledge Discovery for National Security Applications. In: L. Zhou, & W. Kim (Eds.). *Special Issue of Journal of Database Management on Database Technology for Enhancing National Security*, (to appear).

Sheth, A., Arpinar, I. B., & Kashyap, V. (2003). Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic

Relationships. In: M. Nikravesh, B. Azvin, R. Yager & L. A. Zadeh (Eds.), *Enhancing the Power of the Internet Studies in Fuzziness and Soft Computing*: Springer-Verlag.

Tang, C., Xu, Z., & Dwarkadas, S. (2003). *Peer-to-Peer Information Retrieval Using Self-Organizing Semantic Overlay Networks*. Paper presented at the ACM SIGCOMM 2003, Karlsruhe, Germany.

Thomas, J. C., Kellogg, W. A., & Erickson, T. (2001). The knowledge management puzzle: Human and social factors in knowledge management. *IBM Systems Journal*, 40(4).

Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., & Ciravegna, F. (2002). *MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup*. Paper presented at the 13th International Conference on Knowledge Engineering and Management, Sigüenza, Spain.

Yang, B., & Garcia-Molina, H. (2003). *Designing a Super-peer Network*. Paper presented at the 19th International Conference on Data Engineering, Bangalore, India.

Yu, B., & Singh, M. P. (2003). *Searching social networks*. Paper presented at the Second International Joint Conference on Autonomous Agents and Multiagent Systems, Melbourne, Australia.

Terms and Definitions

RDF(S): The Resource Description Framework is a language intended for representation and description of 'resources'. RDF makes use of a Vocabulary Description Language (commonly referred as RDFS or RDF Schema) to describe classes and relations among resources. With RDF(S), we refer to both RDF and its accompanying vocabulary description language. (RDF Primer, W3C Recommendation, February 2004).

Metadata: In general terms, metadata are data about data. Examples are size of a file, topic of a news article, etc.

Semantic metadata: We refer to 'semantic metadata' as that data about data that describes the content of the data. A representative example of semantic metadata is relating data with classes of an ontology. That is, the use of ontology for describing data.

Ontology: From a practical perspective, ontologies define a vocabulary to describe how *things* are related. Relationships of type "is-a" are very basic, yet taxonomies are built with is-a relationships. The value of ontologies is in the *agreement* they are intended to provide (for humans, and/or machines).

OWL: The OWL Web Ontology Language is designed for use by applications that need to process the content of information instead of just presenting information to humans. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDF-S) by providing additional vocabulary along with a formal semantics (OWL Web Ontology Language Overview, W3C Recommendation, February 2004).

Knowledge Composition: Knowledge composition involves assembling knowledge atoms (such as triples in RDF and OWL) to build more complex knowledge maps.

Semantic Web: The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners (W3C). The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation (Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001).