Semantic Graph Mining for e-Science

Tong Yu and **Xiaohong Jiang** and **Yi Feng**
Zhejiang University
College of Computer Science, Zhejiang University, Hangzhou 310027,
People's Republic of China
{ytcs,jiangxh,fengyi}@zju.edu.cn

Abstract

In this paper, we present a methodology, called Semantic Graph Mining, for computer-aided extraction of actionable rules from consolidated semantic graphs of statements. First, generate semantic annotations of a set of heterogeneous knowledge/information resources in terms of domain ontology. Second, merge a semantic graph by means of semantic integration of the annotated resources. Third, discover and recognize patterns from the graph. Fourth, generate and evaluate a set of candidate rules, which are organized and indexed for interactive discovery of actionable rules. As initial implementation efforts of the methodology, a generic architecture of specialized knowledge discovery services is proposed, and an application in biomedicine is initiated.

Introduction

The Semantic e-Science phenomena, as utilizing the Semantic Web technology for publishing knowledge/infromation repositories in a variety of domains, has the potential of great impacts on how scientific activities are conducted. However, the underlying methodology of knowledge discovery based on the Semantic e-Science is still approaching its coherent and mature form. In this paper, we present *Semantic Graph Mining*, a methodology for computer-aided extraction of actionable rules, as an initial effort toward this goal.

Scientific community undertakes a series of mass collaboration efforts for constructing decentralized knowledge/information repositories. As a consequence, these repositories are published on the Web, in conformity with W3C Semantic Web Recommendations, for the usage of a variety of data-driven research projects. Standardization mechanisms affecting a wide range of areas such as domain knowledge representation, semantic integration of heterogeneous web resources, semantic discovery and on-demand composition of web services, are major envisioned achievements with cascading consequences of both benefits and challenges.

The foundation of the Semantic Web is a language specification named Resources Description Framework (RDF) (Beckett 2004). A knowledge base is composed of a set of documents. A document is a set of statements in the form of Subject-Property-Object triple. Subjects are in practice (though not restricted to) resources, and Objects can be resources or literals. A resource is any object that has a unique identity throughout the information space by obtaining a Uniform Resource Identifier (URI) (Berners-Lee *et al.* 2005). Properties define binary relations between two resources or between a resource and a literal. The intuitive meaning of a statement $\langle S, P, O\rangle$ is that the S has a property of the type P, and the property value is the O.

Semantic graph model of a document represents a statement with (1) a node for the subject, (2) a node for the object, and (3) an arc for the predicate, directed from the subject node to the object node. The merging of two semantic graphs is essentially the union of the two underlining sets of statements. This model gives an elegant solution to express complex inter-relationships between concepts in a large information space.

A knowledge base, in scientific domains such as biomedicine, is fundamentally different from a data warehouse in business context, in that it is made of statements instead of facts. Rector and Nowlan define a medical record as a faithful record of what clinicians have heard, seen, thought, and done. They further state that the other requirements for a medical record, e.g., that it be attributable and permanent, follow naturally from this view (Rector *et al.* 1991). The belief of a derived knowledge component, such as a rule or a fact, must be calculated based on, among other factors, the trust of authors that make the related statements from which the rule/fact is derived.

A consolidated semantic graph represents a collective intelligence of the scientific community, and thus serves as a potential source for knowledge discovery. The machine-generated candidate rules, after judgments of domain experts, can be a source for human-readable guidelines or machine-executable scripts. Take biomedicine for example, this rule extraction process can strengthen evidence-based medicine and aid clinical decision-making.

Semantic Graph Mining Methodology

Traditionally, data mining methods mainly deal with facts (as are captured in business transactions), and knowledge reasoning methods mainly deal with statements. *The introduction of Semantic Graph Mining makes the combination of data mining and knowledge reasoning a necessity.* A set

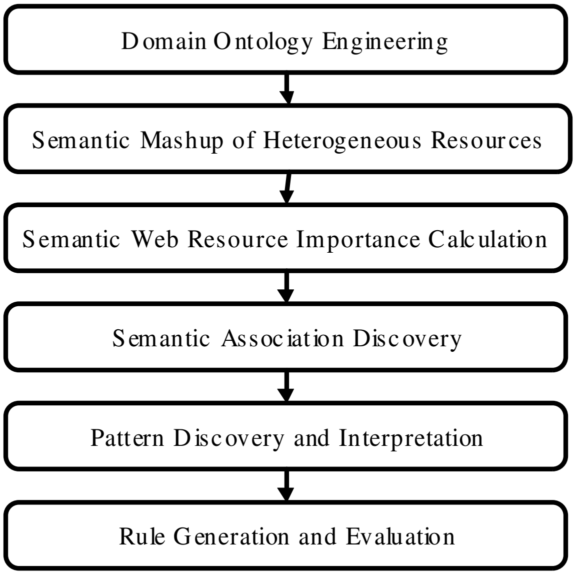
Domain Ontology Engineering
Semantic Mashup of Heterogeneous Resources
Semantic Web Resource Importance Calculation
Semantic Association Discovery
Pattern Discovery and Interpretation
Rule Generation and Evaluation

Figure 1: Semantic Graph Mining Methodology. *The major elements are ontology engineering, semantic mashup of heterogeneous resources, semantic association discovery based on resource importance computing, and generating and interpreting patterns and rules.*

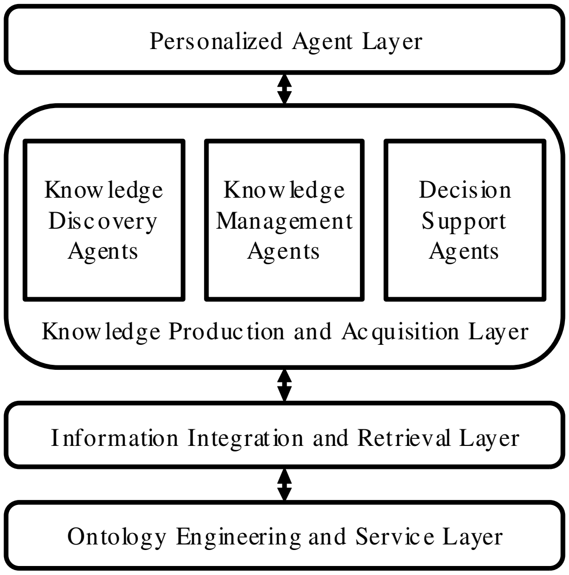
Personalized Agent Layer
Knowledge Discovery Agents
Knowledge Management Agents
Decision Support Agents
Knowledge Production and Acquisition Layer
Information Integration and Retrieval Layer
Ontology Engineering and Service Layer

Figure 2: A Reference Architecture for Semantic Graph Mining. *The four layers are personalized agent layer, knowledge layer, information layer, and ontology layer. The knowledge layer is composed of agents that are specialized in knowledge discovery, knowledge management, or decision support.*

of intriguing problems are derived from this combination in the context of Semantic e-scientific.

In the methodology of Semantic Graph Mining (Figure 1), we first treat a large semantic graph as a directed graph, and apply on that graph with existing graph mining algorithms such as mining frequent sub-graphs, and mining generalized association rules. Second, we perform knowledge inference on discovered patterns and rules. Being self-contained and self-descriptive, a semantic graph provides a reasoning context for interpretation and evaluation of graph mining results. Third, the candidate rules are presented to domain experts for inspection and usage. The methodology logically includes the following major elements:

- Domain Ontology Engineering: RDF Schema (RDF-S) (Brickley & Guha 2004) and OWL (Bechhofer *et al.* 2004) can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms. This representation of terms and their relationships is called an ontology. In the efforts by us and by other teams, we discern the trend of publishing separately-engineered ontologies on the Semantic Web using RDFS/OWL, which provides a coherent scientific ontology infrastructure. For example, Unified TCM Language System (UTCMLS) (Feng *et al.* 2006) is a large-scale ontology (including over 70 classes and 800 properties in 2006 according to (Chen *et al.* 2006)) that supports concept-based information retrieval and information integration. We store the UTCMLS in relational databases, and work towards publishing it on the Semantic Web using OWL.
- Generic Operators Implemented in SPARQL: SPARQL (Prud'hommeaux & Seaborne 2007) is a graph-matching query language for RDF Graphs. A Semantic graph can be physically stored in (legacy) relational databases, and a SPARQL-to-SQL rewriting middleware works on top of these databases to publish data on the Semantic Web. The practice of developing data mining operators in Semantic Web languages such as RDF/XML, RDFS/OWL, and SPARQL, instead of SQL, provides benefits such as addressing the complexity of domain conceptualization, providing the transparency of underlying data structures, and making operators generic and reusable.
- Semantic Mashup of Heterogeneous Resources: DartGrid toolkit (Chen *et al.* 2006) provides an efficient and low-cost solution for semantic mashup of heterogeneous resources. DartGrid abstracts the global information space as a semantic graph of statements expressed in domain ontology. Any physical information source is a materialized view of the global information space. In other words, there is a mapping between a semantic view within the global information space, and the physical schema of an information source. The semantic view, by virtue of being expressed in domain ontology, is practically the representation of semantics of the physical information source. Semantic query rewriting is the translation of a query against a semantic view to a series of queries against underlying physical schemas, together with the afterward integration and interpretation of the query results.
- Resource-Importance-Based Semantic Association Dis-

covery: Semantic associations are complex relationships between resource entities (Anyanwu & Sheth 2003). Semantic associations discovery and ranking has potential applications in such fields as semantic search and social network analysis. Graph algorithms are useful for the computation of semantic associations on a graph representation of the RDF model. The implementation of semantic associations is essentially derivative of importance of Semantic Web resources, yet achieving high scalability is still an open research issue (Mukherjea *et al.* 2005).

- Pattern Discovery and Recognition: Discovery of frequent patterns from large databases is first introduced in (Agrawal & Srikant 1994). The extension of the problem to large graph datasets is to find the most common patterns in the graph structure (Chakrabarti & Faloutsos 2006). Interpretation of frequent patterns is an important yet unsolved issue. The problem of generating semantic annotations for frequent patterns with context analysis is first defined in (Mei *et al.* 2006). It can be addressed, in the context of a semantic graph, by performing knowledge reasoning methods on resulting patterns.
- Rule Generation and Evaluation: In a semantic graph, concept hierarchies over the resources are available, and users are interested in generating rules that span different levels of the concept hierarchies. Mining generalized association rules is first introduced in (Srikant & Agrawal 1995). Numerous techniques have been developed that seek to avoid false discoveries, but they are mostly based on statistical features. As we have mentioned above, one of the unique features of semantic graphs is that they are composed of attributable statements. The rule interestingness measurements can be combined with trust computing of authors to achieve more accurate resulting rule sets.

A Reference Architecture

In order to implement Semantic Graph Mining, we propose a generic agent-based architecture that (1) specifies the semantics of information/knowledge resources and Web Services using domain ontology, (2) attaches semantic annotations to exchanged documents and messages, (3) provides a container for wrapping knowledge discovery operators as semantically-explicit services, and (4) provides a service/resource composition mechanism for generating problem-solving experiments. As is illustrated in Figure 2, this architecture contains four layers:

- Ontology Engineering and Service Layer: This layer provides access to domain ontology, which is composed of concepts and their semantic relations. The parallelism of ontology engineering efforts is unavoidable in order to achieve efficiency in a virtual organization. Providing higher layers with a coherent ontology service by merging disparate ontologies becomes a critical issue.
- Information Integration and Retrieval Layer: Semantic annotating module is responsible for collecting and preprocessing documents and data. This process involves both (semi-)automated and manual tasks in mass collaboration across organizations, and therefore results into distributed and heterogeneous databases. The semantic integration module integrates these databases, and provides a coherent SPARQL query interface to agents in higher layers.
- Knowledge Production and Acquisition Layer: Knowledge discovery module wraps data mining operations as self-explained reusable components, which mine on the semantic graph provided by information layer, in order to discovery evidences and rules. Knowledge management module represents, stores, and indexes logic, evidences, and rules for retrieval, deleting, modification, and updating. Decision support module selects and evaluates evidences and rules for solving problems in a variety of applications.
- Personalized Agents Layer: A personalized intelligent agent translates a user-specified problem-solving requirement into a composition of requests for services and resources, which are provided by agents in lower layers. Navigational and visualization mechanisms are used to present a variety of inter-related objects such as evidences, patterns, and rules. The user can specify an experiment (for knowledge discovery or problem-solving) as an operator tree, which is then executed by the agent through discovering and interacting with other agents that provide demanded resources and services.

Towards a Biomedical Application

As is described in (Chen *et al.* 2006), the DartGrid platform provides access to heterogeneous databases with a coherent SPARQL interface in terms of domain shared ontology. However, the platform only provides information retrieval and searching services, and is unable to satisfy the requirements of knowledge discovery and decision making from biomedicine community.

We work towards a full-fledged platform that is able to provide specialized services for knowledge discovery and decision making. In our plan, every functional module in the architecture will be implemented with a corresponding services system. Some of the systems are built with DartGrid; the others are legacy and/or third party systems, all of which will be integrated within and managed by the DartGrid framework. Intelligent agents will be developed in Ajax and deployed as Rich Internet Applications. We will work with domain experts to achieve the adoption and tailoring of these services systems and tools in biomedical domain. Our major works include:

- Project Initiation: We initiate the project of Knowledge Discovery utilizing Biomedical Semantic Web. We work collaboratively with domain experts to articulate system requirements. We address the concerns of key stakeholders and explain the major aspects of the project such as the underlying methodology, technical/social challenges, and social benefits.
- Integrating and Tailoring of Information Technologies for Biomedicine: We adopt KDD methods and Semantic Web technical framework after integrating and tailoring, in order to address the uniqueness of biomedical requirements.

- Enabling Collaboration in an Open Computing Environment: We deliver our tools as free and open source, and maintain a website for introduction and download. We embrace W3C Recommendations such as RDF/XML, RDFS/OWL, and SPARQL to implement these tools.
- Rapid Prototyping as Delivering Model: We build a prototype for concept demonstration and clarification of requirements, and appeal to key stakeholders for contributive feedbacks.

Summary

The problem of how to utilize Semantic e-Science for knowledge discovery is of both theoretic and practical importance. We propose Semantic Graph Mining to address the problem, with an agent-based architecture, and an initiated biomedical application as a validation effort. Our major works are as the follows:

- Making it clear that the nature of mining a consolidated semantic graph, is not to discover objective rules and hidden facts, but to discover evidences and rules that capture a (perhaps hidden) common view of the scientific community.
- Proposing Semantic Graph Mining methodology that combines data mining and knowledge reasoning to extract actionable rules from semantic graphs.
- Defining a generic architecture to implement the methodology, in conformity with recommendations/standards of the Semantic Web.
- Integrating the state of the art KDD methods with the emerging Semantic Web technical framework to address the unique requirements of biomedical domain.

Acknowledgements

This work is supported by, among others, the China 973 subprogram named “Semantic Grid for Traditional Chinese Medicine” (NO.2003CB316906), the China NSF program (NO. NSFC60503018), and the EU-IST-027595 NeOn project. We would like to thank to our colleagues and partners for their insightful discussions and inspiring advises.

References

Agrawal, R. and Srikant, R. 1994. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th international Conference on Very Large Data Bases*, 487-499. San Francisco, CA.:Morgan Kaufmann Publishers.

Anyanwu, K. and Sheth, A. 2003. ρ-Queries: enabling querying for semantic associations on the semantic web. In *Proceedings of the 12th international Conference on World Wide Web*, 690-699. New York, NY: ACM Press.

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., and Stein, L. A. 2004. OWL Web Ontology Language Reference. http://www.w3.org/TR/owl-ref/.

Beckett, D. 2004. RDF/XML Syntax Specification (Revised). W3C Recommendation. http://www.w3.org/TR/rdf-syntax-grammar/.

Berners-Lee, T., Fielding, R., and Masinter, L. 2005. Uniform Resource Identifier (URI). http://www.ietf.org/rfc/rfc3986.txt.

Brickley, D. and Guha, R.V. 2004. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation. http://www.w3.org/TR/rdf-schema/.

Chen, H.J., Wang, Y.M., Wang, H, Mao Y.X., Tang, J.M., Zhou,C.Y., Yin, A.N., and Wu Z.H. 2006. From Legacy Relational Databases to the Semantic Web: an In-Use Application for Traditional Chinese Medicine. 5th International Semantic Web Conference, Athens, GA, USA, November 5-9, 2006, LNCS 4273.

Feng, Y., Wu, Z.H., Zhou, X.Z., Zhou, Z.M., and Fan, W.Y. 2006. Knowledge discovery in traditional Chinese medicine: State of the art and perspectives. *Artificial Intelligence in Medicine*, Volume 38, Issue 3, November 2006, Pages: 219-236.

Mei, Q., Xin, D., Cheng, H., Han, J., and Zhai, C. 2006. Generating semantic annotations for frequent patterns with context analysis. In *Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining,* 337-346. New York, NY:ACM Press.

Mukherjea, S., Bamba, B., and Kankar, P. 2005. Information Retrieval and Knowledge Discovery Utilizing a BioMedical Patent Semantic Web. *IEEE Transactions on Knowledge and Data Engineering*, Volume 17, Issue 8, August 2005, Pages: 1099-1110.

Prud’hommeaux, E., and Seaborne, A. 2007. SPARQL Query Language for RDF - W3C Working Draft 26 March 2007. http://www.w3.org/TR/rdf-sparql-query/.

Rector, A.L., Nolan, W.A., and Kay, S. 1991. Foundations for an Electronic Medical Record. *Methods of Information in Medicine*, Volume 30, Issue 3, Pages: 179-188.

Srikant, R., and Agrawal, R. 1995. Mining Generalized Association Rules. In *Proceedings of the 21st international Conference on Very Large Databases*,407-419. San Francisco, CA.: Morgan Kaufmann Publishers.