2011 International Conference on e-Education, Entertainment and e-Management

Homology graph mining for social network analysis

Ford Lumban Gaol
Department of Graduate Program in Computer Science
Faculty of Computer Science
Bina Nusantara University
Jakarta, Indonesia

***Abstract*— In this paper, we present a methodology, called Homology Graph Mining, for computer-aided extraction of Social Network rules from consolidated homology graphs of statements. First, we will generate homology sources of a set of heterogeneous social networks resources in terms of relevant pathway. Second, combine a homology graph by means of homology integration of the social network resources. Third, Search and Analyze patterns from the graph. Fourth, generate and evaluate a set of candidate social network rules, which are maintained and indexed for interactive discovery of actionable rules. As part of implementation efforts of the methodology, framework architecture of specialized interrelated knowledge discovery services is proposed, and an application in biomedicine is initiated.**

Keywords-component; Homology; Graph Mining; Social Networks; interrelated knowledge discovery

I. INTRODUCTION

The Homology social network phenomena, as utilizing the Homology Internet technology for relation interrelated knowledge/information storages in a variety of domains, has the broad of great impacts on how social network activities are conducted. However, the underlying methodology of interrelated knowledge discovery based on the Homology social network is still approaching its coherent and mature form. In this paper, we present *Homology Graph Mining*, a methodology for computer-aided extraction of actionable rules, as an initial effort toward this goal.

Social Network community undertakes a sequence of rapid collaboration efforts for constructing dissemination of interrelated knowledge/ information storages. As a consequence, these storages are published on the Internet, in conformity with W3C Homology Internet Recommendations, for the usage of a variety of interrelated data-driven research projects. Standardization mechanisms affecting a wide range of areas such as domain interrelated knowledge representation, homology integration of heterogeneous internet resources, homology discovery and on-demand composition of Internet services, are major envisioned achievements with cascading consequences of both benefits and challenges.

The foundation of the Homology Internet is a language specification named Framework Resources Description (FRD) (Beckett 2004). An interrelated knowledge base is composed of a set of documents. A document is a set of commands 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 entities that have a unique identification 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 *(S, P,O)* is that the *S* has a property of the type *P*, and the property value is the *O*.

Homology 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 combination of two homology 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 interrelated knowledge base, in social network domains such as movie Star, is fundamentally different from a data warehouse in business context, in that it is made of written instead of image. Rector et al define a medicine record as a faithful record of what doctors have heard, seen, thought, and done. They further state that the other requirements for a medicine record, that it be attributable and permanent, follow from this view (Rector *et al*. 1991). The belief of a derived interrelated knowledge component, such as a rule or a fact, must be calculated based on influenced factors, the trust of authors that make the related statements from which the rule/fact is derived. A consolidated homology graph represents a collective intelligence of the social network community, and thus serves as potential source for interrelated 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 movie actors for example, this rule extraction process can strengthen evidence-based movie and actors preferences.

978-1-4577-1382-8/11/$26.00 ©2011 IEEE

II. HOMOLOGY GRAPH MINING METHODOLOGY

Fundamentally, data mining methods mainly deal with facts (as are captured in business transactions), and interrelated knowledge reasoning methods mainly deal with statements. *The introduction of Homology Graph Mining makes the combination of data mining and interrelated knowledge reasoning a necessity*. A set of intriguing problems are derived from this combination in the context of Homology Social Network.

In the methodology of Homology Graph Mining, we first treat a large homology 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 interrelated knowledge inference on discovered patterns and rules. Being self-contained and self-descriptive, a homology 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: FDR Schema (FDR-S) (Brickley & Guha 2004) and LWO (Bechhofer *et al*. 2004) can be used to explicitly represent the meaning of terms in libraries and the relationships between those terms. This representation of terms and their relationships is called ontology. In the efforts by us and by other teams, we discern the trend of publishing separately-engineered ontologies on the Homology Internet using FDRS/LWO, which provides a coherent social network ontology infrastructure. For example, Unified MCT Language System (UMCTS) (Feng *et al*. 2006) is a largescale 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.

• Native Compiler Implemented in SPARK: SPARK (Prud'hommeaux & Seaborne 2007) is an interrelated graph-matching query language for RDF Graphs. A Homology graph can be physically stored in (legacy) relational databases, and a SPARK-to-SQL rewriting middleware works on top of these databases to publish data on the Homology Internet.

The practice of developing data mining operators in Homology Internet languages such as FDR/XML, FDRS/OWL, and SPARK, 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.

• Homology Mashup of Heterogeneous Resources: DartGrid toolkit (Chen *et al*. 2006) provides an efficient and low cost solution for homology mashup of heterogeneous resources. DartGrid abstracts the global information space as a homology 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 homology view within the global information space, and the physical schema of an information source. The homology view, by virtue of being expressed in domain ontology, is practically the representation of homologys of the physical information source. Homology query rewriting is the translation of a query against a homology view to a series of queries against underlying physical schemas, together with the afterward integration and interpretation of the query results.

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

• Knowledge Data Discovery and Objects Recognition: Discovery from large databases is first introduced in (Agrawal & Srikant 1994). The contribution of this framework into the problem to collection of graph datasets is to find the most interesting patterns in the graph database (Chakrabarti & Faloutsos 2006). Interpretation of interesting patterns is an important yet unsolved issue. The problem of generating homology annotations for frequent patterns with context analysis is becomes the key issues for all of Social Network problems (Mei *et al*. 2006). It can be addressed, in the context of a homology graph, by performing interrelated knowledge reasoning methods on resulting patterns.

• Rule Generation and Evaluation: In a homology 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 homology 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.

III. A REFERENCE ARCHITECTURE

In order to implement Homology Graph Mining, we propose a interrelation agent-based architecture that (1) specifies

the homology of information/interrelated knowledge resources and Internet Services using domain ontology, (2) attaches homology annotations to exchanged documents and messages, (3) provides a repository for wrapping interrelated knowledge discovery operators as

homologically-explicit services, and (4) provides a service oriented mechanism for generating problem-solving experiments. This architecture contains 4 (four0 layers:

• Homology Engineering and Service Layer: This layer provides access to domain homology, which is composed of Concepts and their homology relations..

• Information Collaboration and Retrieval Layer: Homology annotating module is responsible for collecting and preprocessing documents and data. This process involves both automation and manual tasks in rapid collaboration across organizations, and therefore results into dispersed and sparse databases. The homology integration module integrates these databases, and provides a coherent SPARK query interface to agents in higher layers.

• Interrelated knowledge Production and Acquisition Layer: Interrelated knowledge discovery module wraps data mining operations as self-explained reusable components, which mine on the homology graph provided by information layer, in order to discovery evidences and rules. Interrelated 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.

• Privatization of Agents Layer: A pivatization 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 interrelated 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.

IV. TOWARDS A BIO-MOLECULAR APPLICATION

As is described in (Chen *et al*. 2006), the MolecularGrid platform provides access to heterogeneous databases with a coherent SPARK 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 interrelated knowledge discovery and decision making from biomedicine community.

We work towards an open platform that is able to provide specialized services for interrelated knowledge discovery and decision making. In our plan, every functional module in the infrastructure will be implemented with a corresponding services system. Some of the systems are built with Dart- Grid; the others are legacy and/or third party systems, all of which will be integrated within and managed by the Dart-Grid framework. Intelligent agents will be developed in Ajax and deployed as Rich Internet Applications. We will explore more in domain experts to get the merging of these services systems and tools in Bio-Molecular domain. Our major works include:

• Project Initiation: We initiate the project of Interrelated knowledge Discovery utilizing Bio-Molecular Homology Internet. 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 self-reconfiguration of Information Technologies for Bio-Medical: We adopt KDD methods and Homology Internet technical framework after integrating and tailoring, in order to address the uniqueness of Bio-Molecular requirements.

• Software Process as Delivering Model: We build a simulated model for concept demonstration and clarification of requirements, and submit to key stakeholders for contributive feedbacks.

V. CONCLUSION

The problem of how to utilize Homology social network for Interrelated knowledge discovery is of both theoretic and practical importance.

We propose Homology Graph Mining as mean and framework to cope the problem, with an agent-based architecture, and an initiated Bio-Molecular application as a validation effort. Our major works are as the follows:

• To underline that the nature of mining a consolidated homology graph, is not to discover the rules and hidden facts between objects, but to discover evidences and rules that capture a (perhaps hidden) common view of the social network community.

• Proposing Homology Graph Mining methodology that combines data mining and interrelated knowledge reasoning to extract actionable rules from homology graphs.

• Defining proven mathematical rules to implement the methodology, in conformity with recommendations/standards of the Homology Internet.

• Integrating the recent framework in Knowledge Data Discover methods with the emerging Homology Internet technical framework to address the unique requirements of Bio-Molecular domain.

REFERENCES

[1] 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.

[2] 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.

[3] Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., and Stein, L. A. 2004.

[4] OWL Web Ontology Language Reference. http://www.w3.org/TR/owl-ref/.
[5] Beckett, D. 2004. RDF/XML Syntax Specification (Revised). W3C Recommendation. http://www.w3.org/TR/rdf-syntax-grammar/.
[6] Berners-Lee, T., Fielding, R., and Masinter, L. 2005.
[7] Uniform Resource Identifier (URI). http://www.ietf. org/rfc/rfc3986.txt.
[8] Brickley, D. and Guha, R.V. 2004. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation. http://www.w3.org/TR/rdf-schema/.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] 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/.
[14] 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.
[15] 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.