WUJNS
Wuhan University Journal of Natural Sciences

Vol.11 No.1 2006 051-056

Article ID:1007-1202(2006)01-0051-06

Data Mining Ontology Development for High User Usability

□LI Yu-hua, LU Zheng-ding[†],
SUN Xiao-lin, WEN Kun-mei,
LI Rui-xuan
College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China

Abstract: This paper mainly introduces the development and implementation of the user-centered data mining service ontology on Universal Knowledge Grid (UKG). UKG is an ontology-based grid architecture model to build large-scale distributed knowledge discovery system on the grid. The data mining ontology services are the main service offering by UKG. It can meet the user requirements of knowledge discovery in different domains and different hierarchies and make the system exoteric, extensible and high usable. A data mining solution for money laundering is introduced.

Key words: data mining; ontology; usability; universal knowledge grid
CLC number: TP 302

0 Introduction

Data mining is one process of mining knowledge from large amounts of data using various analysis tools. The models and relations can be used for prediction and decision supporting. Massive data collections need to be analyzed in many scientific and business areas. Moreover, in many cases data sets must be shared by large communities of users that pool their resources from different sites of a single organization or from a large number of institutions.

So one kind of architecture is needed to facilitate database resource integration, data mining, knowledge sharing and knowledge integration for solving large-scale distributed knowledge discovery and knowledge integration.

An ontology is an explicit specification of a conceptualization where definitions associate concepts, taxonomies and relationships with human-readable text and formal, machine-readable axioms[1].

Universal Knowledge Grid (UKG) is an ontology-based grid architecture for building large-scale distributed knowledge system on the grid. UKG emphasizes geographically distributed high-performance knowledge discovery applications and integration services. Ontology server is the center module. Data mining ontology services are the main services offering by ontology server[2].

The user-centered design is viewed as a process that emphasizes especially on making products usable. The usability of a product is defined on ISO 9241-II as "the extent to which a product can be used by specified user to achieve specified goals with effectiveness, efficiency and satisfaction in a specified content of use"[3].

This paper mainly discusses the design and the imple-

Received date: 2005-04-26
Foundation item: Supported by the National Natural Science Foundation of China (60403027), the National Key Technologies R&D Program of China during the 10th Five-Year Plan Period (2002BA103A04, 2001BA102A06-11).

Biography: LI Yu-hua (1968-), female, Ph.D. candidate, Associate professor, research direction: data mining, knowledge management, semantic Web and ontology and artificial intelligence. E-mail: yuhua_yy@163.com.
† To whom correspondence should be addressed. E-mail: zdlu4409@public.wh.hb.cn

mentation of the user-centered data mining service ontology on Universal Knowledge Grid.

The rest of the paper is organized as follows. Section 1 describes the Universal Knowledge Grid general architecture. Section 2 introduces the building of data mining ontology. Section 3 discusses related work and Section 4 gives a conclusion of the paper.

1 The UKG Architecture

The UKG architecture (Fig. 1) is defined on top of Grid toolkits and services[2].

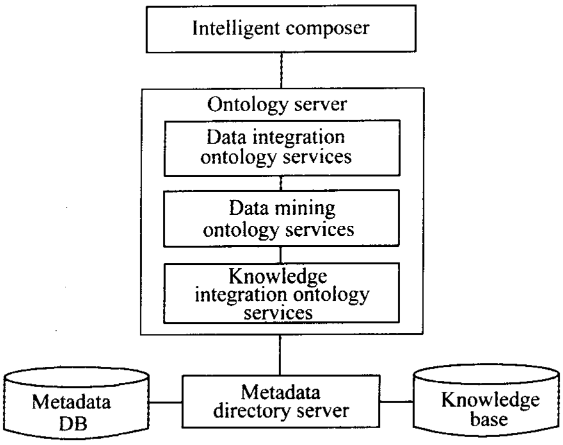
Intelligent composer
Ontology server
Data integration ontology services
Data mining ontology services
Knowledge integration ontology services
Metadata DB
Metadata directory server
Knowledge base

Fig. 1 The architecture of the UKG

1.1 Intelligent Composer

The intelligent composer offers a set of graphical tools. The user can form a graphic representation of her/his knowledge requirements using common visual facilities or natural language. It displays the machine understandable semantics and gives an explanation of the display. Users can operate resources according to semantics.

1.2 Ontology Server

The ontology server is the center module. It is responsible for the management and query of explicit, declaratively represented ontologies. An ontology server offers part or all of the following services, data integration ontology services, data mining ontology services, knowledge integration ontology services.

1.2.1 Data integration ontology services

Data integration ontology services describe the semantic of web documents, bridge the gap between semi-structured and structured databases, facilitate data cleaning and data preparation, heterogeneous data sources integration.

1.2.2 Data mining ontology services

Data Mining Ontology Services offer a semantic model of user's application needs, solution and data mining algorithm resources.

1.2.3 Knowledge integration ontology services

Knowledge integration ontology services maintain a set of public available ontologies, facilitating the communication and knowledge interchange among different knowledge bases, scoring the pattern of data mining ontology and appending the effective knowledge to knowledge bases and integrating knowledge resources of different levels to support problem analysis and solution.

1.3 Metadata Directory Server

This service deals with maintaining the metadata to describe all the data, tools and knowledge used in the Universal Knowledge Grid.

The metadata information is represented by XML (eXtensible Markup Language) documents and is stored in Metadata database (Metadata DB) and Knowledge base.

1.4 Metadata DB

Metadata DB stores metadata of heterogeneous data sources, tools and algorithms used for data integration and data mining.

1.5 Knowledge Base

Knowledge base stores knowledge obtained from results of the data mining process, i.e. learned models and discovered patterns and the knowledge gathered from application domain.

2 The Building of Data Mining Ontology

It is must be developed deeply and widely that data mining technique, application actuality and developmental trends in order to build the data mining ontology and identify its range.

2.1 The Design of Data Mining Ontology

There are many researches on data mining technique and system. Robert Grossman put forward the concept of four ages of data mining[4]. Gregory Piatetsky-Shapiro summarized the development of data mining system in KDD2000[5], which can be concluded three phases: independent data mining system, transverse data mining algorithm tools gather and lengthways data mining application resolving scheme. Lu Hongjun developed the integration trends of data mining, database and data warehouse in PAKDD'01[6]. Han Jiawei put forward the trends of developing lengthways data mining system combining data mining with its application in 2001[7]. Haseltine gave an alternate "user-centered" approach in KDD2004 that can produce KDD solutions with shorter development cycles, lower costs, and much better usability[8].

During their initial development, KDD solutions often focus heavily on algorithms, architectures, software, hardware, and systems engineering challenges, without first thoroughly exploring how end-users will employ the new KDD technology. As a result of such "system-centered" design, there are many useless features that prolong development and significantly add to life cycle cost, making the system hard to operate and use[8].

In building of data mining ontology, it should be summarized that the historical data mining production and trends analysis, following the design conception of user-centered. The high usable data mining service should be offered to user.

Figure 2 shows the structure of data mining ontology, it shows the concept classes and their relations.

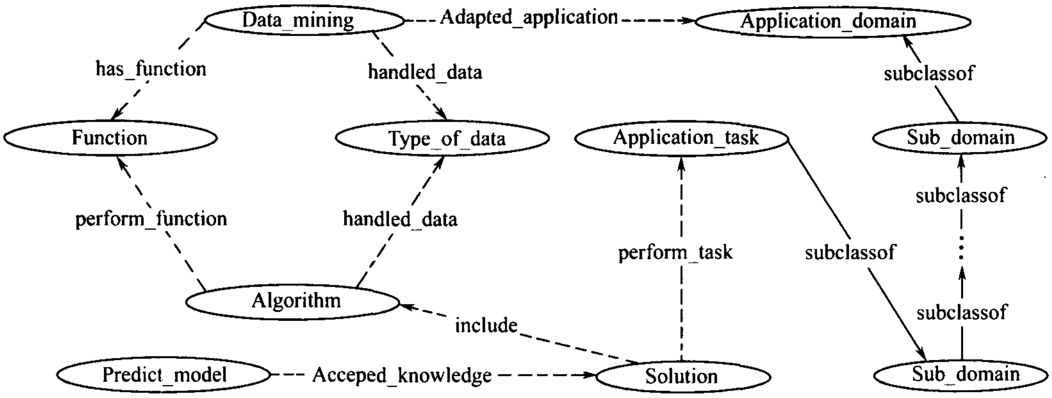

Fig. 2 The architecture of data mining ontology

According to function, the algorithms can be categorized as characterization, discrimination, clustering, classification, prediction, association, outlier analysis, link analysis, evolution analysis, visualization, etc. General algorithm is used to handle structured data. Essential data mining technique should be developed to cope with complex types of data, such as complex objects, spatial data, multimedia data, time-series data, text data and the World Wide Web.

Data mining is focusing on application, and is used in wide application domain. According to the data mining applications poll of Kdnuggets[9], the application domain is divided into banking, bioinformatics/Biotech , Direct marketing/fundraising , Fraud Detection, eCommerce/Web, entertainment/News, insurance, investment/stocks, manufacturing, medical/pharma, retail, scientific data, security, telecommunications, travel and other. Each domain can be divided into several sub-domains. Each sub-domain includes several application tasks. Several domains may have same sub-domains. Several sub-domains may have one same application tasks. Each task has one or more solutions, every solution includes one or multiple algorithms. One algorithms can be included by several solutions.

For example, banking, eCommerce, insurance, telecommunications, retail, travel , Direct marketing have Client analysis. Client analysis can be classified as client class, client behavior analysis, client credit analysis etc.

The predictive model accepted from the solution is expressed by means of the Predictive Model Markup Language (PMML)[10]. PMML is an XML-based language that provides a way for applications to define statistical and data mining models and to share models between PMML compliant applications.

2.2 Ontology Implementation

In ontology coding, OWL[11] is adopted that is recommended by W3C. OWL can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms. This representation of terms and their interrelationships is called an ontology. OWL has more facilities for expressing meaning and semantics than XML, RDF, and RDF-S, and thus OWL goes beyond these languages in its ability to represent machine interpretable content on the Web.

The ontology is coded with Stanford's protege 2000. It includes concept class Data_mining, Function, Type_of_data, Algorithm, Application_domain, Sub_domain, Application_task, Solution, Predictive_model etc. The relations of the concept, except the hierarchical concept such as subclassof relations of "Application_domain" and “ Sub_domain ”, “ Sub_domain ” and “ Application_ task ”, are expressed by concept properties.

Due to space limitation, only two of the concept class relations are introduced. The codes shown in Fig. 3 describe that “ Sub_domain ” is the subclass of “ Application_domain ” and “ Application_ Task ” is the subclass of “ Sub_domain ”.

```
〈owl :Class rdf : ID =“ Application_Domain ”/〉
   〈owl :Class rdf :about =“ # Sub_domain ”〉
      〈rdfs : subClassOf  rdf : resource = “ # Application _ Do-
main ”/〉
      〈rdfs :label〉Sub domain〈/ rdfs :label〉
〈/ owl :Class〉
 …
〈owl :Class rdf : ID =“ Application_ Task ”〉
      〈rdfs :subClassOf〉
         〈owl :Class rdf : ID =“ Sub_domain ”/〉
      〈/ rdfs :subClassOf〉
〈/ owl :Class〉
```

Fig. 3 The subclass relation description of the concept classes

The OWL codes of Fig. 4 show part of the Solution class. It uses three properties “ include ”, “ perform_ task ”, “ accepted _knowledge ” to build its relations with class “ Algorithm ”, “ Application_ Task ”, “ Predictive_ Model ”.

```
〈owl :Class rdf : ID =“ Solution ”/〉
 …
〈owl : FunctionalProperty rdf : ID =“ include ”〉
      〈rdfs :range rdf :resource =“ # Algorithm ”/〉
      〈rdfs :domain rdf :resource =“ # Solution ”/〉
      〈rdf :type
rdf : resource = “ http :/ / www. w3. org/ 2002/ 07/ owl # Ob-
ject Property ”/〉
〈/ owl : FunctionalProperty〉
〈owl : FunctionalProperty rdf : ID =“ perform_task ”〉
      〈rdfs :domain rdf :resource =“ # Solution ”/〉
      〈rdf :type rdf : resource = “ http :/ / www. w3. org/ 2002/
07/ owl # Object Property ”/〉
      〈rdfs :range rdf :resource =“ # Application_ Task ”/〉
〈/ owl : FunctionalProperty〉
〈owl : FunctionalProperty rdf : ID =“ accepted _knowledge ”〉
      〈rdfs :domain rdf :resource =“ # Predictive_Model ”/〉
      〈rdfs :range rdf :resource =“ # Solution ”/〉
      〈rdf :type
rdf : resource = “ http :/ / www. w3. org/ 2002/ 07/ owl # Ob-
ject Property ”/〉
   〈/ owl : FunctionalProperty〉
```

Fig. 4 Part of description of the solution class

2.3 One Example of Money Laundering Data Mining Application Solution

Money laundering[12] is a sub-domain of Security, Fraud detection and Banking. The mapping of its application solutions and their used algorithms are showed in Table 1.

Table 1 Money laundering application solutions and their used algorithms

Application	Function of Algorithm	Algorithm
Trade network analysis	Link analysis	Coplink Concept Space[13], CLIQUE[14], two-tree PFS[15]
Trade attributes related analysis	Classification	SLIQ[16]
Group different cases	Clustering	CLIQUE[14]
To detect unusual amounts of fund transfers	Outlier analysis	TOP-n LOF [17]
To characterize unusual access sequences	Sequence model analysis	Multi-objective GA[18]
Trade trends analysis	Prediction	Multiple regression[19]

2.4 The Application and Characteristics of the System

The data mining ontology offers a semantic model of user 's application needs, solutions and data mining algorithm resources. With the help of ontology the users run through the following steps to accomplish one special data mining task.

1) Through the intelligent composer, the user selects data sources through data integration ontology, enters its application task such as Trade network analysis;

2) The system selects suitable data mining application solution through searching data mining ontology, it includes algorithms of Coplink Concept Space, CLIQUE and two-tree PFS;

3) Transfer the information of data source and solution to the Metadata directory server, execute the solution, access the acquired predict model with knowledge integration ontology and test it;

4) Store the tested predict model into knowledge base and feed back the result to user through the intelligent composer;

5) The user uses the tested predict model into his transaction system for decision-making.

6) Other users direct use the predict model in knowledge base for decision-making or after modifying it.

The system has high usability and Extensibility.

a) High usability

The user needn't understand the data mining algorithms, only selects the application solution according to his application task and then implement it.

The user can use the predict model expediently into their business system supporting decision-making. The new application solution can be defined using system data mining algorithms by the domain expert and help them to find the appropriate solution.

b) Extensibility

UKG can share predict model with other PMML application and make predict model, source data and data mining tools self-existent.

2.5 Discussion

Many of the recently appeared knowledge discovery-oriented systems, such as TeraGrid, InfoGrid, DataCutter, AdAM, Discovery Net, Terabyte Challenge Testbed[20], have been designed for specific domains, and later they have been extended to support more general application construction. Some of such systems are essentially advanced interfaces for integrating, accessing and processing large datasets. Furthermore, they provide specific functionalities for the support of typical knowledge discovery processes.

The Knowledge Grid[21] builds a domain-independent knowledge discovery environment on the Grid. It provides specifically designed services for the integration of parallel and sequential data mining algorithms, and the management of base datasets and extracted knowledge models. The system is only used expertly by experts who are familiar with data mining algorithms. It is hard to get good knowledge model without knowing data mining algorithm very well.

The UKG is an ontology-based grid architecture model for building large-scale distributed knowledge system on the grid. UKG emphasizes geographically distributed high-performance knowledge discovery applications and integration services. It can help various hierarchy users from different application domains for data mining and knowledge integration service, offering high extensibility and usability.

3 Conclusion

This paper presents the user-centered data mining service ontology on Universal Knowledge Grid, which not only offers abundant data mining algorithms for different function and different type of handling data, but also provides multiple data mining application solutions for different application domains. It can meet the user requirements of knowledge discovery in different domains and different hierarchies and make the system exoteric, extensible and high usable.

An example solution for money laundering is introduced, including the hierarchy of the domain and the mapping of application solutions and their used algorithms.

Future work is to integrate more application solutions into the system, develop more effective knowledge integration services and offer more usable knowledge discovery and knowledge integration services.

References

[1] Gruber T R. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 1993, **5** (2): 199-220.

[2] Li Yu-hua, Lu Zheng-ding. Ontology-Based Universal Knowledge Grid: Enabling Knowledge Discovery and Integration on the Grid. *Proceedings —2004 IEEE International Conference on Services Computing*. Shanghai, China, September 2004. 557-560.

[3] Klett F. The Impact of User-Centered Design Concepts in Virtual Learning Environments. *Proceedings of the Fifth International Conference on Information Technology Based Higher Education and Training*. Instanbut, Turkey, June 2004. 222-226.

[4] Grossman R. Supporting the Data Mining Process with Next Generation Data Mining Systems. *http://www.lac.uic.edu/~grossman/paper/esj-98.htm.* 98. December 1998.

[5] Piatetsky-Shapiro G. Knowledge Discovery in Database: 10 Years after. *SIGKDD Explorations*, 2000, **1**(2): 59-61.

[6] Lu H. Seamless Integration of DM with DBMS and Application. *http://www.cs.ust.hk/luhj/ps/pakdd01.pdf*, June 2001.

[7] Han Jia-wei. Data Mining-Current Status and Research Directions. *http://db.cs.sfu.ca/sections/publications/slides/slides.htm*, August 2001.

[8] Haseltine E. Invited Talks: User Centered Design for KDD. *http://www.acm.org/sigs/sigkdd/kdd2004/program/invited_talks.html*, August 2004.

[9] KDnuggets. KDnuggets Polls: Data Mining Applications. *www.kdnuggets.com/polls/2004/data_mining_applications_industries.htm*, August 2004.

[10] Bohanec M, Moyle S, Wettschereck D, *et al*. A Software Architecture for Data Preprocessing Using Data Mining and Decision Support Models. *Workshop Integration Aspects of Data Mining, Decision Support and Meta Learning*. Freiburg, Germany, September 2001. 13-24.

[11] McGuinness D L, Harmelen F V. OWL Web Ontology Language Overview. *http://www.w3.org/2004/OWL/*, July 2004.

[12] Chen Hsin-chun, Chung Wing-yan, Xu J, *et al*. Crime Data Mining: A General Framework and Some Examples. *IEEE Computer*, 2004, **37**(4):50-56.

[13] Hauck R V, Chen H. Coplink: A Case of Intelligent Analysis and Knowledge Management. *Proceedings of the International Conference on Information Systems*, Charlotte. North Carolina, USA, December 1999. 15-28.

[14] Agrawal R, Gehrke J, Gunopulos D, *et al*. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. *International Conference on Management of Data Proceedings of the* 1998 *ACM SIGMOD International Conference on Management of Data*. New York: ACM Press, 1998. 95-105.

[15] Xu J, Chen H. Fighting Organized Crime: Using Shortest-Path Algorithms to Identify Associations in Criminal Networks. *Decision Support Systems*, 2004, **38**(3):473-487.

[16] Mehta M, Agawal R, Rissanen J. SLIQ: A Fast Scalable Classifier for Data Mining. *In Proc of the* 5*th International Conference on Extending Database Technology (EDBT)*. Avignon, France, March 1996. 18-32.

[17] Jin W, Tung A K H, Han Jia-wei. Mining Top-n Local Outliers in Large Databases. *In Proceedings of* 7*th. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, USA, August 2001. 293-298.

[18] Kaya M, Alhajj R. Multi-Objective Genetic Algorithm Based Approach for Optimizing Fuzzy Sequential Patterns. *Proceedings of the* 16*th IEEE International Conference on Tools with Artificial Intelligence (ICTAI* 2004*)*. Boca Raton, FL, USA, November 2004. 396-400.

[19] Han Jia-wei, Kamber M. *Data Mining Concepts and Techniques*. Beijing: *Higher Education Press*, 2001. 319-321.

[20] Congiusta A, Pugliese A, Talia D, *et al*. Designing Grid Services for Distributed Knowledge Discovery. *Web Intelligence and Agent Systems*, 2003, **1**(2):91-104.

[21] Cannataro M, Comito C. A Data Mining Ontology for Grid Programming. 1*st International Workshop on Semantics in Peer-to-Peer and Grid Computing, in Conjunction with WWW*2003. Budapest, Hungary, May 2003. 113-134.

□