

# SoPhIA: A Unified Architecture for Knowledge Discovery

Dimitrios K. Korentzelos, Huaglory Tianfield, and Tom Buggy

School of Computing and Mathematical Sciences, Glasgow Caledonian University, 70  
Cowcaddens Road, Glasgow G4 0BA, UK

{D.Korentzelos, H.Tianfield, T.Buggy}@gcal.ac.uk

**Abstract.** This paper presents a novel architecture Soph.I.A (Sophisticated Intelligent Architecture), which integrates Knowledge Management and Data Mining into a unified Knowledge Discovery Process. Within SophIA Data Mining is driven by knowledge captured from domain experts. Knowledge Grid is briefly reviewed to envision the implementation of the proposed framework.

**Keywords:** Data Mining, Knowledge Discovery Process, Knowledge Grid, Knowledge Management, Ontology.

## 1 Introduction

Data Mining (DM) is basically about knowledge discovery ([1] [5] [17]). Another field on knowledge discovery is Knowledge Management [15]. The adherence of DM with KM could be seen in Knowledge Mining. For instance, Knowledge-Based Systems can be used to discover knowledge on Knowledge Repositories [manipulating experts' knowledge], Data Warehouses and OLAP systems.

"Knowledge Discovery is the Process based on an intelligent and sophisticated mechanism, which works for the collection and process of data, information, and knowledge. This leads towards the discovery of expedient facts or relationships between them" [12]. The main characteristic of the Knowledge Discovery Process is that it has no compulsory tasks.

The use of *Domain Knowledge* could significantly improve the efficiency of Knowledge Discovery Process (KDP), if it is integrated within the process ([10] [18]). The domain expert can prevent the data miner from researching or being misguided within the database. It has to be noticed that during KDP, many potentially interesting patterns can be found but few of them contain nuggets interesting for the creation of new knowledge on the domain. Piatetsky-Shapiro [16] did a further analysis on objective rule interestingness. Obviously, there is a link between interestingness measures and domain knowledge. One of the challenges is how to capture the domain knowledge from the domain expert so to integrate it within the KDP mechanism.

Knowledge Management can have an important role during the KDP, in supporting the assimilation and capture of essential knowledge from the domain

expert. For successful Knowledge Management, it is essential to have an effective collection, documentation, refinement, dissemination and utilisation of knowledge. Frank and Hampe [4] stress that Knowledge Management Systems (KMS) provide a collection of views, which are illustrated from user's perspective and conclude by sustaining that KMS can contribute towards overcoming the gap between business and technology. This gap is one of the factors that can block the effective use of computers and communications [11].

Another interesting perspective for processing human knowledge, so as to be generally accessible (e.g. in an organizational environment), is to implement Knowledge Management by the use of Ontologies. Gruber [7], defined 'Ontology', as the formal explicit specification of shared conceptualisation. Guarino [8] denoted that the problem of the above definition is the vagueness of the term 'conceptualisation'. After analysing several definitions, Guarino defined 'Ontology' as: "... a logical theory that constraints the indented models of a logical language" [8]. The construction of a knowledge base can be based on ontology and ontological theory. Guarino and Giarretta [9] gave three possible technical senses to the word 'Ontology': a)'ontological theory'; b)'Ontology' is a synonym of 'specification of an ontological commitment'; and c)'Ontology' is a synonym of 'conceptualisation'.

Ontologies can help Knowledge Management to form knowledge in a way so to be easily reusable. "Next generation knowledge management systems will likely rely on conceptual models in the form of ontologies to precisely define the meaning of various symbols" [14].

## 2 SophIA: An Architecture for Knowledge Discovery Process

This paper presents SophIA. SophIA's Knowledge Discovery Process is divided into four steps i.e. Domain Knowledge Assimilation, Data Preparation, Data Mining, and Knowledge Dispersion. Defining the boundaries, between different fields such as Data Mining and Knowledge Management, is helpful for the identification of malfunctions in KDP. Usually, several malfunctions happen during the transition from one step to another, especially when these steps belong to different disciplines. For example, the transition from 'Assimilation of Domain Knowledge' step (which involves how to manage knowledge) to 'Data Preparation' step (that has to do with managing data). For a smooth transition from one step to the other, the use of linkage steps (such as Data Understanding) is considered more than essential. SophIA's design supports the use of an object oriented design, a common sense language/rules (either for business people or for data miners) so to be able to translate/understand the messages from one discipline to another, an integrated user interface system and finally an easily comprehensible functionality.

SophIA's Design Rationale is characterised by the following attributes: a higher level of automation in KDP, higher level of abstraction, an effective User Interface, utilisation of Domain Knowledge (in order to reduce both the time that

is needed for Knowledge Discovery, the amount of data that has to be processed during the KDP), and finally a unique repository of expert’s knowledge, which can be fortified as intellectual capital.

Figure 1 presents the SophIA’s Framework for the Knowledge Discovery Process. SophIA’s kernel is the *Core Repository* (CR). CR holds information generated by both DM and KM, which has the potential to be transformed into new knowledge. The *Sophisticated Mechanism*, (see Figure 1) is responsible for managing knowledge that comes from a domain expert. Its basic parts are as follows: The *Domain Knowledge Worker Repository* (D-KnoWR ), where the knowledge provided by the domain expert (with the help of a user interface) is stored, an *Ontology* responsible to translate the above knowledge, and the *Knowledge Repository* (KnoWR) where the translated domain knowledge from the Ontology is transformed into information. The KnoWR is also responsible to find new information, which could be considered as potential new knowledge (and so send them to the CR). The *Intelligent Mechanism*, is responsible for managing the data stored on a *Data Warehouse*. The results of mining, the *Data Warehouse* are also stored on CR. The information stored on CR, is processed so to give the new domain knowledge. With the help of a user interface, SophIA disseminates the new knowledge. The user interface is also used to answer queries related to the new information.

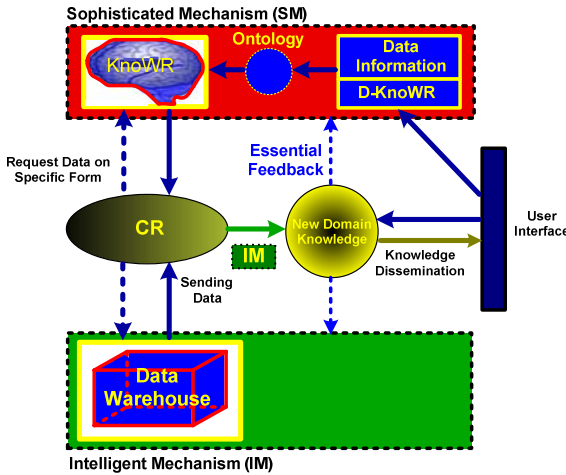


Fig. 1. Framework of SophIA

The Functionality steps of SophIA include: domain knowledge acquisition, data preparation, data mining, and knowledge dispersion.

*Assimilation of Domain Knowledge.* In general, the contribution of the business analyst is crucial, in guiding the data miner, for designing the DM project. Further assistance is essential for the data miner so to identify basic relationships or patterns, which the business analyst already knows. It would be easier

for the data miner to extract information from a database, instead of trying to take advantage of the expert's knowledge instantly. Another significant reason for codifying business analyst's knowledge is that the human brain has capacity limitations and aging, comparing to computer databases, which are able to store, transfer, and manipulate petabytes. This step is an effort in the direction of the best possible way to codify domain knowledge. The main purpose is to enable the codified knowledge to be easily and comprehensibly accessible for the data miner. Figure 1 depicts SophIA's sophisticated mechanism for codifying domain knowledge: The business analyst, with the help of a user interface, passes his/her knowledge on to D-KnoWR. By the use of text mining techniques, several keywords are mined from D-KnoWR so to formulate an ontology. An object oriented schema, can be used to identify data, which has the potential to be interesting for the data miner. Finally, the data described from the ontology, is passed on to KnowR. Consecutively, the data miner searches KnowR for finding knowledge relative with the project's needs.

*Data Preparation*, in SophIA, involves the following steps: Transform the data from KnowR to the CR, enrich the CR with data from the DW and finally, prepare the data stored in CR for mining. This is feasible by organizing the data from knowR (used as metadata) and the data from the DW, in a structured model.

*Data Mining*. After the data is ready for mining, a specific type of algorithm (e.g. a back-propagation Neural Network algorithm) can be implemented. The type of algorithm depends on the type and the volume of the data. For instance, a clustering algorithm can be used to describe data that resides on the CR database. The purpose is to discover patterns with the potential to be characterised as new knowledge.

*Knowledge Dispersion* is the step where the results are expressed by the help of a user interface in a comprehensible way for the business analyst. The success of the whole process is measured primarily by checking if criteria as in manager's initial questionnaire have been answered, and secondly by investigating the optimisation of the new patterns that came up from SophIA.

### 3 The Advent of Knowledge Grid

The astonishing evolution of the World Wide Web has proved the power of global networking. The global use of the Web has created massive volume of data, information, and knowledge, which resides on servers in the form of HTML, XML documents and their derivatives. The superinduction of Ontologies, RDF and OWL languages, have transformed the Web into Semantic Web. In chorus, Grid technologies make an effort to take advantage of the data [data, information, and knowledge] that resides on Semantic Web.

Knowledge Grid is a relatively new concept regarding Grid terminology and definitions [2]. It can be said that Knowledge Grid is the evolution of merging Semantic Web with Grid Technology. Its effort is to collect and distribute Knowledge using the power of a Grid. Zhuge et al defined 'Knowledge Grid' as "...a mechanism that shares and manages the distributed heterogeneous re-

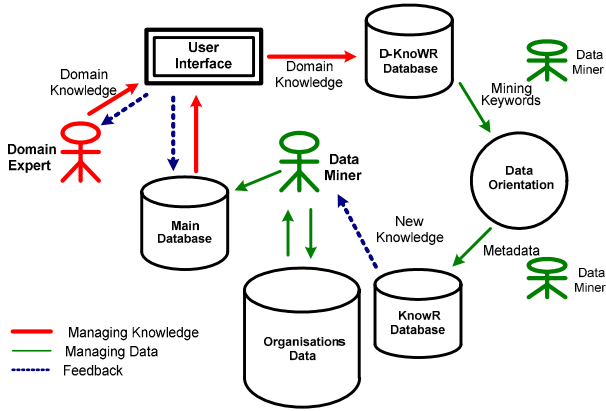


Fig. 2. Acquisition of Domain Knowledge

sources spread across the Internet in a uniform way [19]. It has been advocated that Grid Architectures should be compliant with Open Grid Services Architecture (OGSA) [13] and Web Services Conceptual Architecture (WSCA) [3]. In an extended report for Semantic Web, Grid Computing and Cognitive/Knowledge/Semantic Grid, Geldof [6] describes the benefits, the applications, the current status, critical issues and the challenges of Knowledge Grid.

## 4 Outlook

SophIA has been designed to take into consideration the use of domain expertise for the best possible implementation of a Knowledge Discovery Process. SophIA, is intended to manipulate knowledge that resides on the Knowledge Grid so to utilise it for more efficient guidance of Data Mining algorithms. Therefore, SophIA integrates an Unstructured Knowledge-Base and a Meta-Knowledge-Base, using Ontology as an interface. Currently, SophIA is in the stage of evaluating possible tools and platforms particularly from Knowledge Grid research for its implementation.

## Acknowledgements

We are grateful to Professor Julian Newman for comments on an earlier version of this paper.

## References

1. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*. **39**(11) (1996) 27–34

2. Foster, I.: What is Grid? A Three Point Checklist. Argonne National Laboratory and University of Chicago. (2002)
3. Foster, I., Kesselman, C., Nick, J.M., Tuecke, S.: An Open Grid Services Architecture. A Unified Framework for Distributed Systems Integration. Technical Report. Globus Project Technical Report. [Online: <http://www.globus.org/research/papers/ogsa.pdf>]. (2002)
4. Frank, U., Hampe, J. F.: An Object-Oriented Architecture for Knowledge Management Systems. Report Nr.16. Arbeitsberichte des Institut fur Wirtschaftsinformatik IWI. (1999)
5. Frawley, W.J., Piatetsky-Shapiro, G., Matheus, C.J.: Knowledge Discovery in Databases: An Overview. *AI Magazine*. **13**(3). (1992) 57–70
6. Geldof, M.: The Semantic Grid: Will the Semantic Web and Grid go Hand in Hand?. European commission DG Information Society Unit "Grid Technologies". (2004)
7. Gruber, T. R.: A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*. **5**(2). (1995) 199–220
8. Guarino, N.: Understanding, Building, and Using Ontologies. LADSEB-CNR. National Research Council. Padova. Italy. [<http://ksi.cpsc.ucalgary.ca/KAW/KAW96/guarino/guarino.html>]. (1996)
9. Guarino, N., Giaretta, P.: Ontologies and Knowledge Bases: Towards a Terminological Clarification. In Mars, N. (eds) (1995). *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*. IOS Press. Amsterdam. (1995) 25–32
10. Hsu, W. H., Welge, M., Redman, T., Clutter, D.: High Performance Commercial Data Mining: A Multi-strategy Machine Learning Application. *Proceedings of the Data Mining and Knowledge Discovery*. Kluwer Academic Publishers. **6**. (2001) 361–391
11. Keen, P.W.: *Shaping the Future. Business Design through Information Technology*. Harvard Business School. Press Boston. Massachusetts. (1991)
12. Korentzelos, D.: *SophIA: A Mechanism for Knowledge Discovery*. PhD Transfer Report CMS/COM/2004/9. Glasgow Caledonian University. Glasgow. UK. (2004)
13. Kreger, H.: *Shaping the Future. Web Services Conceptual Architecture*. Technical Report WSCA 1.0. IBM Software Group. (2001)
14. Maedche, A., Motik, B., Stojanovic, L., Studer, R., Volz, R.: *Ontologies for Enterprise Knowledge Management*. IEEE Computer Society. (2003) 26–33
15. Newman, B.: *An Open Discussion of Knowledge Management*. [Online: <http://www.km-forum.org/>]. (1991)
16. Piatetsky-Shapiro, G.: Discovery, analysis, and presentation of strong rules. In Piatetsky-Shapiro, G., Frawley, W. (eds). In *Knowledge Discovery in Databases*. AAAI/MIT Press. Menlo Park.CA. (1991) 229–248
17. Simoudis, E., Cabena, P., Haddjimian, P., Standler, R., Varhees, J., Zanasi, A.: *Discovering Data Mining, From Concept to Implementation*. Prentice Hall PTR. (1997)
18. Yoon, S. C., Henschen, L. J., Park, E. K., Makki, S.: Using Domain Knowledge in Knowledge Discovery. *Proceedings of the 4th International Conference on Information and Knowledge Management (CIKM'99)*. ACM Press. (1999) 243–250
19. Zhuge, H., Liu, J.: A Fuzzy Collaborative Assessment Approach for Knowledge Grid. *Future Computer Systems* **20**. (2004) 101–111