Data Mining Governance for Service Oriented Architecture

Ali Beklen
Software Group
IBM Turkey
Istanbul, TURKEY
alibek@tr.ibm.com

Turgay Tugay Bilgin
Dept. of Computer Engineering
Maltepe University
Istanbul, TURKEY
ttbilgin@maltepe.edu.tr

Abstract— **The aim of this study is to propose a platform called Data Mining Registry, Repository and Statistics (DMRRS). The concept of this platform is to govern the data mining algorithm which needs to be integrated to service oriented architecture and to be used in the cloud analytics environment. The focus is on the notion of a reference architecture for DMRRS, XML schema-based algorithm definition data models and data mining algorithm life cycles.**

Keywords -service oriented architecture; soa governance; data mining.

I. INTRODUCTION

In recent years, data mining has attracted a great deal of attention in the information industry, as well as in society as a whole. This is due to the wide availability of huge amounts of data and the imminent need for converting such data into useful information and knowledge [1]. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration [1]. Data mining can be viewed as a result of the natural evolution of information technology [1].

Present day data can no longer be labelled as ‘simple’ [2]. As data in various domains becomes more heterogeneous, complex and peculiar, more intelligent techniques are required to mine it and to extract useful knowledge from it [2].

Data mining is a compilation of techniques, methods and algorithms utilized in order to extract knowledge hidden amongst huge amounts of data. It is, therefore, much more than a list of statistical formulas applied to a collection of data [2].

According to current trends, cloud computing and service oriented architecture are emerging as complex applications of the implementation and presentation type. Regarding data mining, there are many topics that need to be researched to adopt data mining algorithms with cloud computing and service oriented architecture.

In order to implement the data mining algorithm as a service, many interface options are available, for example web service. In order to administer the services, service registry and repository are necessary to define and manage the interfaces. Although the services could be organized in a service registry and repository, this does not allow developers or architects to manage and define the algorithm itself.

Management of the data mining algorithm necessitates a governance lifecycle. This lifecycle must allow information architects to test the maturity level of the algorithm in terms of performance and resource consumption. On the other hand, long running algorithms need to be monitored to predict the duration of further mining requests.

The aim of this study is to build a reference approach to adopt data mining algorithms to service oriented architecture (SOA) in a governed way to address the above issues. In order to implement the reference approach, data mining governance life cycle, the architecture of Data Mining Registry, Repository and Statistics (DMRRS) implementation and a sample algorithm definition schema have been developed.

II. RELATED WORK

There have been many studies published with the aim of adopting data mining to service oriented architecture. Chen et al. proposed service rating for data mining to improve information sharing [3]. Xu et al. proposed a service based architecture for data mining applications, including configuration service, service engine, monitor service, analysis service, visualization service, computing service and data and algorithm provision service [4]. Tsai et al. proposed a Dynamic Data Mining Process (DDMP) system based on Service-Oriented Architecture (SOA), which enables each activity in the data mining process to be promotable as a web service operated on the Internet, providing data preprocessing, data mining algorithm, and visualization analysis functions [5].

The Data Mining Registry, Repository and Statistics concept proposed in this paper however, is about data mining governance which is very important in terms of algorithm definition, life cycle management, runtime analysis reporting, and managing long and short running processes related to governed algorithms.

III. GOVERNANCE

Governance is the process of making correct and appropriate decisions on behalf of the stakeholders of those decisions or choices. In its corporate application, governance is the process of ensuring that the best interests of a company's or organization's stakeholders are met through all corporate decisions, from strategy through to execution. In

ISBN: 978-1-61208-105-2

an IT framework, governance focuses on appropriate oversight and stakeholder representation for IT spending and overall IT management [6].

SOA governance is the creation, communication, enforcement, and adaptation of policies used to direct and control the creation and implementation of the life cycle of services. It is a run-time and design-time administrative capability that no organization should be without [7].

In this research, we are proposing data mining governance as a new governance type to define the life cycle of data mining algorithms, which requires management and monitoring in a service oriented accessible repository.

IV. DATA MINING GOVERNANCE LIFECYCLE

In order to manage the life cycle of data mining algorithms and to integrate with service oriented architecture, defined policies and life cycle stages are necessary. In this study, a candidate reference life cycle is proposed. This life cycle consists of the following stages:

1. Define the purpose and development of algorithm.
2. Define the algorithm metadata and communication interface to make it discoverable
3. Test the maturity level with a pre-defined training data set
4. Deploy the algorithm
5. Monitor the algorithm runtime environment
6. Collect the feedback from the runtime environment

Figure 1 shows the life cycle.

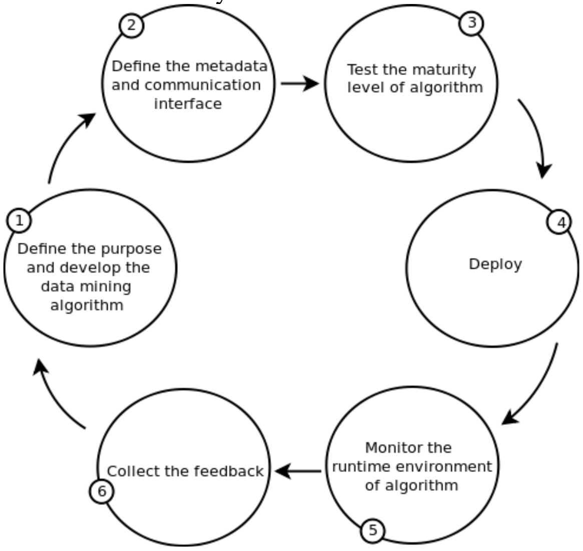
2
Define the metadata and communication interface
3
Test the maturity level of algorithm
1
Define the purpose and develop the data mining algorithm
4
Deploy
Collect the feedback
6
Monitor the runtime environment of algorithm
5

Figure 1. Data Mining Governance Life Cycle

The SOA governance addresses administration of the service life cycle. However, if the project needs to expose the data mining algorithms as a service, several problems may occur, such as:

- Atomic web services are not suitable for long running transactions, and data mining algorithms need to be managed and monitored as long running.
- SOA governance defines service and its life cycle but data mining algorithm does not fall within this scope.

The most important stage of the proposed life cycle is testing the maturity level of data mining algorithm. This stage guarantees the algorithm quality, in terms of algorithm duration, hardware resource consumption and mining result. For example, if the data mining algorithm owner would like to publish the algorithm, this step forces it to compare the algorithm with the best performing similar ones by using the same training data set. If it performs at least the mature ones this lifecycle stage allows him/her to move forward in the lifecycle.

On the other hand, provision of the data mining capabilities of the project as a service necessitates that it be discoverable by another system in the SOA. In order to mitigate with data mining governance requirements and to operate the life cycle, we propose the data mining registry, repository and statistics concept.

V. DATA MINING REGISTRY, REPOSITORY AND STATISTICS

The SOA service registry and repository is the service repository used for storing, accessing, and managing metadata used in the reuse, selection, invocation, management, and governance of SOA services in SOA architecture. It helps to define a central point for accessing predefined and custom service description artefacts acquired from a number of sources, including service and service application deployments, development tools, and other service metadata repositories. Interfaces are provided for finding, retrieving, and advertising services using classifications and properties during application and service design and development, service change and release processes, service invocation and execution, and operational management of services.

DMRRS is a concept which defines data mining algorithms and their metadata, manages their life cycle and monitors their performance. With the help of this concept, it becomes possible to govern native data mining algorithms, expose their interfaces in a transparent way to the SOA, help other services to discover algorithms, and link the Predictive Model Markup Language (PMML) to algorithms.

One of the critical requirement about proposing the data mining algorithm as a service is building and managing an association between PMML and algorithms. The DMRRS concept provides setting up an association between PMML and algorithms and allows the mining requestor to select and associate multiple algorithms to one mining model.

Another important factor in data mining is the runtime duration of the long running algorithms. This research proposes the statistics manager component. The proposed component in the tool's architecture is responsible for measuring the duration of running algorithm instances, and for providing this information to assist in prediction of the transaction duration for the next request to the other consumer systems. This also helps the architect to understand the behaviour of the algorithm, and to serve it as synchronous or asynchronous.

The component architecture diagram of the DMRRS represented in Figure 2 consists of six major components:

- Authentication: This component is responsible for authentication of users who have a definition in the user repository.
- Authorization: This component is responsible for authorization of users and generates user interfaces based on use rights. There are three types of users :
    - Mining Algorithm Developer: This role can create a candidate data mining algorithm definition and initiate a lifecycle. This role is also responsible for defining the metadata and communication interface and testing the maturity level of the algorithm.
    - SOA Architect: This role is responsible for auditing the compatibility of definitions with the SOA governance and approves the verified algorithm for deployment.
    - Data Architect: This role is responsible for linking the different algorithms and connecting them to Predictive Model Markup Language (PMML) documents. This approach proposes the building of a composite algorithm which consists of a unique flow of different types of algorithms, and links the algorithm runtime interface to the PMML definition.
- Document Manager: This component is responsible for managing the PMML documents and Data Mining Algorithm Definition Language (DMADL) documents. It proposes management of basic operations concerning these types of documents, and generation of DMADL type documents.
- Document life cycle modelling: This component is responsible for managing the data mining governance life cycle stages. It proposes assignation of responsibility roles to the stages.
- Algorithms runtime statistics: This component is responsible for providing the algorithms' runtime related statistical data as a web service. The last algorithm request uses this statistical output if it is available to predict the duration of last mining request. It does this by comparing the requested training data set size and algorithm type with the similar historical requests.
- Algorithm runtime manager: This component monitors the instances of running algorithms and feeds the runtime statistics database with data about transaction duration, central processing unit (CPU), random access memory (RAM) and hard disk drive input/output (HDD I/O) utilization.

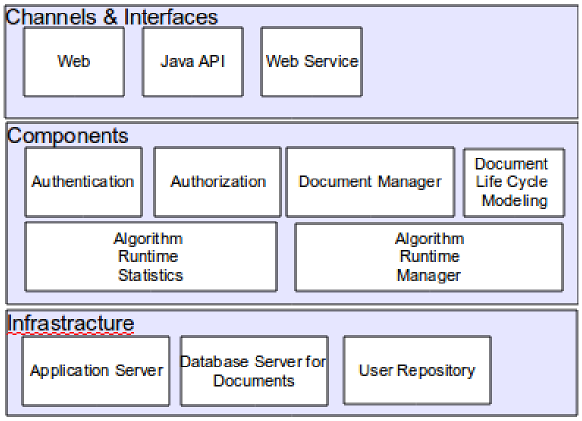

Figure 2. DMRRS Component Architecture Diagram

In many successful SOA implementations, all of the services base their communication on enterprise semantics. These semantics usually include a common vocabulary, a semantic information model, and common schemas. What is helpful is that such an approach does not require data transformation throughout the enterprise. Rather it is the responsibility of service consumers and providers to implement the abstractions from their internal data models to enterprise semantics [8].

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information [10]. DMRRS tool allows data mining algorithm owner to define its metadata to make it discoverable. Otherwise, algorithm requestor does not able to search and discover the required algorithm.

In order to govern the data mining algorithm, define its metadata and to adopt them to SOA governance, this research proposes the data mining algorithm definition data model (DMALD) which helps to define common data mining metadata fields and implementation details. Because of data mining algorithm characteristics, every data mining algorithm requires different types of metadata. In this research we developed a sample algorithm metadata model called association rule learner algorithm (ARMD). In order to define the DMALD and ARMD, Extensible Markup Language (XML) schema is used.

The XML schemas enable you to declare the type of textual data allowed within attributes and elements, using simple type declarations. For example, by utilizing these types you could specify that an element may contain only date values, only positive numbers, or numbers within a certain range. Many commonly used simple types are built into XML Schemas. This enables you to easily create documents that are intended to represent databases, programming languages, and objects within programming languages [9].

In Figure 3, the relationship between the ARMD and DMALD XML schemas is clearly shown at a high level. The boxes titled Governance and Implementation are inherited from DMALD as a complex type and help to avoid duplication of common fields among the mining algorithm definitions. These schemas are also used in the DMRRS tool to govern and integrate the algorithm to the SOA environment in a standard way.

A part of the schema designed for association rule mining definition can be seen as follows:

```
<xs:include schemaLocation="DMALD.xsd"></xs:include>
  <xs:annotation>
      <xs:documentation>
         Assocation rule mining definition schema.
      </xs:documentation>
  </xs:annotation>
  <xs:complexType name="ARMD">
      <xs:sequence>
        <xs:element name="GovernanceInfo" type="Governance"/>
        <xs:element name="Implementation" type="Implemantation">
         </xs:element>
        <xs:element  name="InputParams"  type="InputParameters"/>
        <xs:element name="OutputParams"
                    type="OutputParameters">
      </xs:sequence>
  </xs:complexType>
  <xs:complexType name="InputParameters">
      <xs:sequence>
        <xs:element name="MaxItemsetCount" type="xs:long"/>
        <xs:element name="MaxItemsetSize" type="xs:long"/>
        <xs:element name="MaxSupport"
                    type="xs:nonNegativeInteger"/>
        <xs:element     name="MinItemSetSize"     type="xs:long"/>
        <xs:element name="MinProbability"
                    type="xs:nonNegativeInteger"/>
        <xs:element name="MinSupport"
                    type="xs:nonNegativeInteger"/>
      </xs:sequence>
  </xs:complexType>
```

VI. CONCLUSION

In this study, the importance of adopting data mining governance to SOA governance has been examined in detail. The key requirements of data mining governance in enterprise level applications have been mentioned.

In order to implement the data mining governance concept in a SOA project, DMRRS tool, governance life cycle and data mining algorithm definition data model have been designed and proposed as a reference.

Implementation design has been kicked off and our subsequent studies will focus on making it live and increasing the number of implemented data mining algorithm definition data model types.

On the other hand, cooperating with a cloud analytics project will be a challenging continuation in terms of integrating the proposed approach and tools to a cloud environment.

REFERENCES

[1] J. Han and M. Kamber, "Introduction". Data Mining: Concepts and Techniques, Second Edition. Morgan Kaufmann Publishers, 2006, pp. 1-4.

[2] D. Taniar, "Preface". Strategic Advancements in Utilizing Data Mining and Warehousing Technologies: New Concepts and Developments. IGI Global, 2010.

[3] Y. Chen, Brad Cohen and B. A. Hamilton, Data Mining and Service Rating in Service-Oriented Architectures to Improve Information Sharing. Aerospace Conference, 2005, pp. 1-10.

[4] L. Xu, Y. Wang, G. Geng, X. Zhao and Nan Du, SDMA: A Service-based Architecture for Data Mining Applications. IEEE International Conference on Services Computing, 2008, pp. 1-2.

[5] C. Tsai and M. Tsai, A Dynamic Web Service based Data Mining Process System, Computer and Information Technology, 2005, pp. 1-7.

[6] E. A. Marks, "Chapter 1 - The SOA Governance Imperative". Service-Oriented Architecture Governance for the Services Driven Enterprise. John Wiley & Sons. 2008.

[7] M. Rosen, B. Lublinsky, K. T. Smith, and M. J. Balcer, "Chapter 12 - SOA Governance". Applied SOA: Service-Oriented Architecture and Design Strategies. John Wiley & Sons. 2008.

[8] M. Rosen, B. Lublinsky, K. T. Smith, and M. J. Balcer, "Chapter 5 - Service Context and Common Semantics". Applied SOA: Service-Oriented Architecture and Design Strategies. John Wiley & Sons. 2008.

[9] D. Hunter, "Chapter 5 - XML Schemas". Beginning XML, 4th Edition. Wrox Press. 2007.

[10] National Information Standards Organization, “What is metadata”. Understanding metadata, NISO Press. 2010.

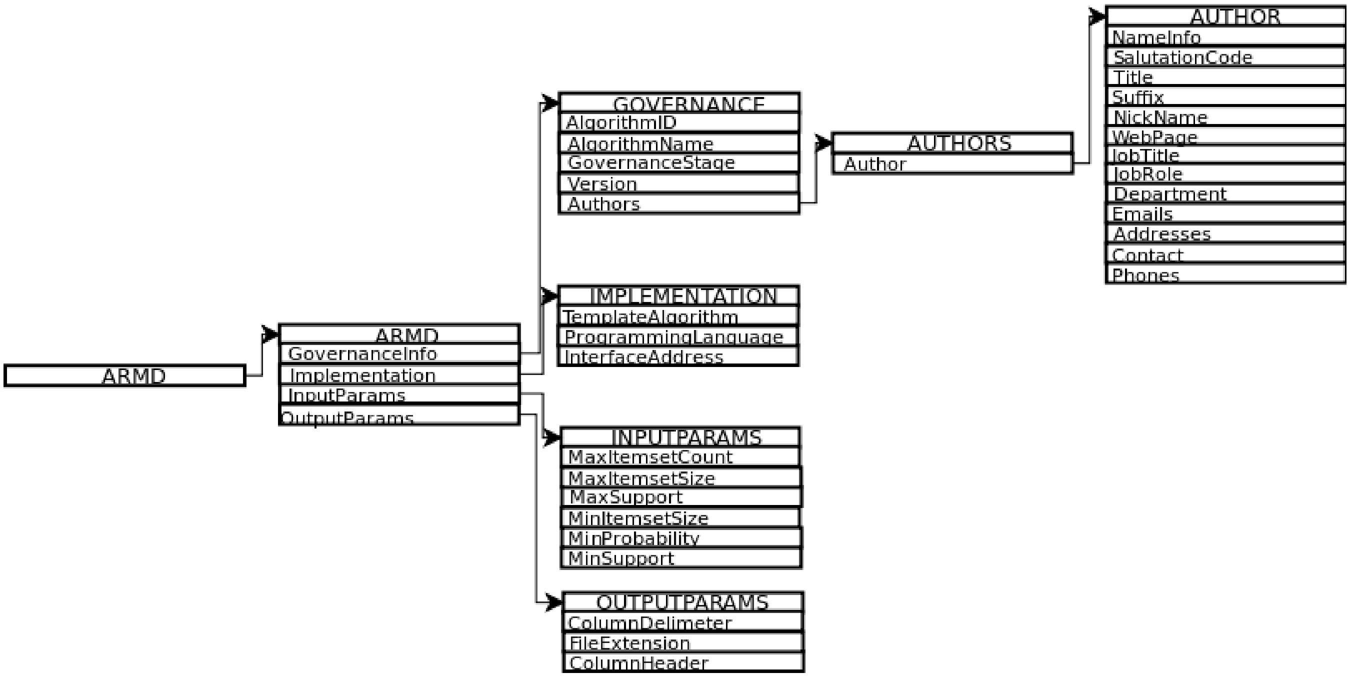

Figure 3. ARMD Schema Architecture