Promoting Data Mining Methodologies by Architecture-Level Optimizations

Ge Xin\(^1\), Ding Enjie\(^2\), Xie Hongxia\(^1\)

1) School of Computer Science & Technology, China University of Mining and Technology, Xuzhou Jiangsu, 221116 (E-mail: \{gexin, xiehx\}@cumt.edu.cn)
2) School of Information & Electronic Engineering, China University of Mining and Technology, Xuzhou Jiangsu, 221116 (E-mail: enjied@vip.sina.com)

Abstract—This paper presents a new theoretical data mining framework that adapts the existing data mining systems with the architecture of the Knowledge Grid, the mechanism of the ontologies, and the factor of the human-driven knowledge. Aiming at much of the research to date focusing on the technique and algorithms, the new framework describes the essential factors from systemic and technical viewpoints respectively in order to balance the effect between the two aspects.

Ⅰ. INTRODUCTION

Data mining, which is the process of extracting meaningful knowledge from large quantity of data, has been applied to a wide range of applications in its relatively short existence [1]. But its researches have primarily been technical and algorithmic rather than systemic; indeed, the enormous research effort expended in developing industrially suitable data mining technologies has rendered further incremental developments unnecessary [2].

Additionally, without explanation data are just a heap of meaningless numbers or symbols. Combining with human knowledge, they will become information, from which data mining tools can extract valid rules. However, these rules may describe dependencies between attributes and classes in a quantitative way. It would be possible to obtain experience (knowledge about knowledge [3]) about an application area by fusing this kind of knowledge with the combined qualitative knowledge of several experts (see Fig. 1).

The aim of this study is thus to present a new way of thinking about data mining projects which is organizationally inclusive and identifies the importance of experts factor during data processing. This new theoretical framework, which includes key organizational issues that are absent from existing approaches, may be used to reduce the likelihood of data mining projects producing high quality knowledge with little or no business benefit. Moreover, it combines expert knowledge with automatically extracted knowledge of data mining approaches to obtain more comprehensive and more valuable knowledge about an application area. It serves also to strengthen the theoretical foundations of the field.

The remainder of this paper is structured as follows. Section II discusses the current state of data mining and some related work about existing data mining methodologies. Section III introduces some important terms and explains how they work in the new framework. Section IV presents the design of the theoretical framework. Finally, Section V concludes this paper and gives an outlook to additional ideas.

Ⅱ. RELATED WORK

Data mining has been applied to an enormous variety of projects across a wide range of application areas. In a recent survey, 65% of data mining projects are found to be successful [4].
The key components of any data mining project are the source data, the project personnel and the systems that will apply data mining algorithms to the data to discover knowledge. These systems have been through several generations since their introduction in the 1980s: first generation, typically restricted to a single discovery algorithm and requiring expert operators; second generation, supporting multiple discovery algorithms and other phases in the data mining process such as data transformation and third generation, which undertake the entire data mining process for a specific business problem, with an interface designed for business users [5]. And much of the systems to date have concentrated upon their more technical aspects, such as algorithms, data structures, data access; existing methodologies are primarily guidelines for data mining specialists.

Although ad hoc approaches to perform data mining are popular, standardized models have evolved [6, 7]; they distill the lessons learned from many projects into methodologies that compose the systems development lifecycle. For example, CRoss-Industry Standard Process (CRISP) is one of the established approaches used in data mining [6].

![The CRISP methodology](image)

The CRISP methodology [6], presented in Fig. 2, is used as a starting point for the theoretical data mining framework. The diagram shows the six phases in order of execution from left to right; it should be noted that this ordering is not fixed, and scope exists for flexible movement in a similar manner to the systems development lifecycle [6]. Although CRISP incorporates many organizational issues, its focus is predominantly upon technical items, and analysis of the balance between these at each stage is not performed.

Recently the advent of the Grid has introduced substantial changes in the way that data and computations are conceived and developed within industrial and scientific applications. Cannataro et al. [8] developed the Knowledge Grid to help researchers and professionals analyze the very large amount of data that distributed over corporate or worldwide Grids. Currently, the Knowledge Grid mechanisms are being designed and implemented following the Service Oriented Architecture (SOA) paradigm [10] and the emerging Web Services Resource Framework (WSRF) family of standards [11] are being adopted for re-implementing the Knowledge Grid services.

This study proposes a new theoretical data mining framework that encompasses organizational factors of the Knowledge Grid and human-driven knowledge base to optimize the performance of data mining system from architecture viewpoint.

III. PREREQUISITES

A. Human-driven knowledge

Human-driven knowledge is proposed in [3] by contrast with data-driven knowledge which is extracted from data by data mining systems. It is application-specific knowledge, but this kind of knowledge originates from human experts which have a certain expertise concerning an application area. Thus human-driven knowledge often describes dependencies without using information about the (numeric) characteristics of attributes.

Due to the objectivity of data and processing techniques, data-driven knowledge is often provided in an implicit way. It typically has a quantitative nature and it is too general and theoretic to be applied in a specific area. Thus in practice human-driven knowledge can help algorithms reduce search area and prune decision branches to expedite the digging rate (see Fig. 3).

![Data mining based on the human-driven knowledge](image)

The fuzzy rule (data-driven knowledge) extracted from data sets can be explained automatically by inferring from the human-driven knowledge base. This understanding of data-driven knowledge can now be used to select expert knowledge which is related to this kind of knowledge as they are able to initiate appropriate actions that solve a problem. Furthermore, it is possible to detect novel knowledge; for instance, if the fuzzy rule indicates some
interesting information which is not inferred from the human-driven knowledge base, it is knowledge about knowledge, that is, experience.

B. The Knowledge Grid

The Knowledge Grid [12] is designed to provide a higher level of abstraction and a set of services by which it is possible to integrate Grid resources to support all those phases of the knowledge discovery process.

The Knowledge Grid Services (K-Grid Services) are organized in two hierarchical levels [13] (see Fig. 4). One is the High-level K-Grid layer, which contains:

- The Data Access Service (DAS), responsible for the publication and searching of data to be mined (data sources), as well as the search of inferred models (mining results).
- The Tools and Algorithms Access Service (TAAS), responsible for publishing and searching extraction tools, data mining tools, and visualization tools.
- The Execution Plan Management Service (EPMS) defines the abstract execution plan which is represented by a graph describing interactions and data flows between data sources, extraction tools, data mining tools, and visualization tools. The execution plan is referred to as an abstract execution plan, because it may include both well identified resources and abstract resources, i.e., resources that are defined through constraints about their features, but are not known a priori.
- The Results Presentation Service (RPS) offers facilities for presenting and visualizing the extracted knowledge models.

The other is the Core K-Grid layer which offers basic services for the management of metadata describing features of hosts, data sources, data mining tools, and visualization tools. This layer also coordinates the application execution by attempting to fulfill the application requirements with respect to available Grid resources. The Core K-Grid layer comprises two main services:

- The Knowledge Directory Service (KDS), responsible for handling metadata describing Knowledge Grid resources. Such resources include hosts, data repositories, tools and algorithms used to extract, analyze, and manipulate data, distributed knowledge discovery execution plans, and knowledge models obtained as a result of mining processes.
- The Resource Allocation and Execution Management Service (RAEMS), used to find a suitable mapping between an abstract execution plan and available resources so as to achieve an instantiated execution plan, which defines the resource requests for each data mining process.

These services can be implemented in different ways using available Grid environments such as the Globus Toolkit, UNICORE, and Legion.

IV. DESIGN OF ORGANIZATIONAL INCLUSIVE DATA MINING SYSTEM

In this section we will discuss the new theoretical data mining framework (Fig.5) that adapts CRISP with the architecture of the Knowledge Grid, and meanwhile imports the factor of the Human-driven Knowledge. In addition we present the technical and systemic factors in each phase respectively, so that their influence may be compared (see Table 1).
guiding users to build a clear question description that can be understood by the computer. There are many ways to achieve it, such as heuristic questions, presenting preliminary findings as a catalyst for question identification [14]. Once the goals for a data mining project have been identified, its scope can be determined, which helps to reduce a number of project risk factors, such as the time frame, the person month size, personnel requirements.

The success of a data mining project will be achieved not only in technical factors but also systemic factors, such as top management support, staff skills and training, since they can affect the success of the project [15].

B. Data Understanding & Preparation

The source data is clearly a critical component of the framework. This aspect may be hampered by problems in following factors: data availability (e.g. disparate and distributed sources, un-access of private data), data quality (e.g. uncertain, missing and volatile data, poor organization and maintenance), data characters (e.g. semi-structured and unstructured data).

Data warehouses (DW) and ontologies are crucial in solving such problems. Through data preparation that includes data cleaning, data integration, data transformation, data filtering etc., data warehouses can improve the quality of data [16], and offer a platform on which all operations of data mining are performed. Since data warehouses mainly aim at centralized, structured data, ontologies can use the Metadata Repository (MR) [17] to deal with distributed, semi-structured or unstructured data. These repositories can store in different Grid nodes as long as they are available. And they can be accessed just by URIs [18, 19]. The sketch of the process is shown in Fig. 6 [20].

In addition, there are many ontology mapping and merging tools which can be employed to transform data, such as GLUE [21], RiMOM [22], HCOME-merge [23], MOA [24] and OntoDNA [25].

This phase mainly transforms the conceptual model into an abstract execution plan, and stores the plan in the Execution Plan Repository (EPR) for subsequent processing.

![Fig. 7 Distributed execution plan generating interactions](image)

Fig. 7 describes the interactions that occur when a conceptual model of the application to be executed (step 1) based on the Knowledge Grid Services [13, 26], and the RAEMS receives the abstract execution plan (step 2). First of all, the RAEMS queries the local KDS to obtain information about the resources needed to instantiate the abstract execution plan (step 3). Note that the KDS performs the searching both accessing the local MR and querying remote KDSs (step 4).

After the instantiated execution plan is obtained, the RAEMS coordinates the actual execution of the overall computation. To this purpose, the RAEMS invokes the appropriate data mining services (DM Services) and basic Grid services (e.g., file transfer services), as specified by the instantiated execution plan (step 5). The results of the computation are stored by the RAEMS into the DW (step 6), while the execution plan is stored in the EPR (step 7). To make available the results stored in the DW, it is necessary to publish results metadata into the MR. To this end, the RAEMS invokes the local KDS (steps 7 and 8) for publishing data or tools.

D. Resource Allocation & Execution Management

This procession is used to find a suitable mapping between an abstract execution plan and available resources, with the goal of satisfying the constraints (e.g., CPU, storage, memory, database, and network bandwidth requirements) imposed by the execution plan. The output of this process is an instantiated execution plan, which defines the resource requests for each data mining process. Generated execution plans are stored in the EPR. After the
execution plan activation, this service manages the application execution and the storing of results in the Data-driven Knowledge Base Repository (DKBR).

At each active Grid node invoked by services, the operations performed are just like the traditional data mining process. It is important in the sub-application to make sure the quality of results, which is dependent on the algorithms used, and can be affected by the production of a greater quantity of patterns than can feasibly be examined, or failure to check whether discoveries were caused by chance [27, 28].

Misinterpretation of the discovered knowledge, such as confusing correlation with causality, can be a danger [14]. As mentioned before, human-driven knowledge is useful at explanation data and description questions. Human-driven knowledge can be obtained by acquiring knowledge from single users, and the knowledge of different users will be combined into a more comprehensive knowledge base with an hybrid (symbolic/numeric) knowledge representation scheme. Most work related to these tasks is based on symbolic approaches such as propositional logic, description logic, first-order logic, modal logic, production systems, or ontologies.

**E. Knowledge Fusion**

This fusion component can be seen as an interface that enables the system to use human-driven knowledge in order to analyze, verify, and validate data-driven knowledge and vice versa. In a first step, we will focus on analyzing data-driven knowledge (fuzzy interesting rules) by using human-driven knowledge (causal relationships between concepts in the application domain). To solve this problem, the attributes and terms of the fuzzy rules are mapped onto the corresponding concepts entered by human experts by using the MR built by ontologies.

Moreover, the bottleneck of any knowledge based system is knowledge acquisition, so adapting the human-driven knowledge by the results of knowledge fusion is an approach to expand the bottleneck. The procession of distinguishing and absorbing new knowledge should be a self-awareness mechanism and needs a suitable learning approach.

**V. CONCLUSION**

This study has extended the primarily technical focus of existing data mining approaches to produce a new theoretical framework incorporating the architecture of the knowledge grid, the mechanism of the ontologies, and the factor of the human-driven knowledge, so as to optimize the organization of a data mining system.

These younger and still evolving technologies promote us taking the important steps on the way towards an open, distributed, and service-based data mining system - that can interact with components and data sources in various concrete Grid infrastructures, that is open in architecture and design, and that builds upon a powerful expert knowledge base.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>The factors of the new theoretical data mining framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Understanding &amp; Data Understanding</td>
<td>Distributed Execution Plan Generating</td>
</tr>
<tr>
<td>Systemic factors</td>
<td>• Application area</td>
</tr>
<tr>
<td></td>
<td>• Task goals</td>
</tr>
<tr>
<td></td>
<td>• Management support</td>
</tr>
<tr>
<td></td>
<td>• Resource</td>
</tr>
<tr>
<td></td>
<td>• Staff skills</td>
</tr>
<tr>
<td></td>
<td>• Risk factors</td>
</tr>
<tr>
<td>Technical factors</td>
<td>• Graphical user interface</td>
</tr>
<tr>
<td></td>
<td>• Interaction</td>
</tr>
<tr>
<td></td>
<td>• Describing questions</td>
</tr>
</tbody>
</table>

**REFERENCES**