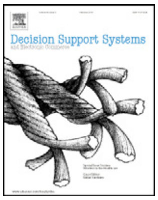
Decision Support Systems 55 (2013) 399–411

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss



A service oriented architecture to provide data mining services for non-expert data miners

Marta Zorrilla *, Diego García-Saiz

Department of Mathematics, Statistics and Computation, University of Cantabria, Avda. Los Castros s/n, Santander, 39005, Spain

ARTICLE INFO

Available online 29 May 2012

Keywords:
Analytics service
BI-as-a-Service
Knowledge discovery database
Data mining
Service-oriented architecture
Web Services

ABSTRACT

In today's competitive market, companies need to use discovery knowledge techniques to make better, more informed decisions. But these techniques are out of the reach of most users as the knowledge discovery process requires an incredible amount of expertise. Additionally, business intelligence vendors are moving their systems to the cloud in order to provide services which offer companies cost-savings, better performance and faster access to new applications. This work joins both facets. It describes a data mining service addressed to non-expert data miners which can be delivered as Software-as-a-Service. Its main advantage is that by simply indicating where the data file is, the service itself is able to perform all the process.

1. Introduction

In a market as competitive and global as today's, currently affected by a deep economic crisis, information is one of the main managerial assets since its analysis helps in effective steering, as De Leeuw [35] pointed out 28 years ago.

Regardless of the size of the company, the need for having an accurate and reliable knowledge of what is affecting its business and for discovering new useful information hidden in the data for correct decision making has meant that since the end of nineties, business intelligence (BI) tools have been used more and more although the sector growth has not been so high in the last few years as a consequence of the economic crisis [72].

Business intelligence tools, as is well-known, encompass a wide range of techniques and technologies which are used to gather, provide access to and analyze data from the operational systems of the organization and other external sources (for instance surveys, information from competitors or data from the web, among others) with the aim of offering decision makers a more comprehensive knowledge of the factors affecting their business and, in this way, help them to take more accurate and effective managerial actions.

Among the different elements which make up a BI environment [33], we consider four of them, the data warehouse (DW), the On-Line Analytical Processing (OLAP) technology, the reporting tools and the data mining techniques to be the most essential.

The DW is the integrated repository of the strategic information of the organization which generally includes measurements, metrics and facts from the different business processes of the company (known as key performance indicators — KPI). These measurements are defined according to the different users' perspectives. The OLAP technology meets managers' and business analysts' needs to quickly search and explore accurate, up-to-date, complete information from the DW, this information being detailed as well as aggregated. The reporting tools and, in particular, dashboards and scorecards aim to help analysts to monitor and analyze the status of their KPI and drill into detailed data to identify the root causes of problems and intervene while there is still time. Lastly, the data mining techniques facilitate the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and models which can be used directly in decision making (for instance, a model for preventing credit risk).

Nowadays the majority of large companies and corporations have to a greater or lesser extent a DW and they use reporting and OLAP tools to extract and analyze the information which allows them to position themselves strategically in the market. However, although there are areas where data mining techniques are being used more and more, such as business [48], marketing [61], education [16], banking [46], health systems [78], and so on [52], their use is still not generalized. This is mainly due to the fact that data mining projects need highly qualified professionals (expert data miners) to achieve, in reasonable time, useful results for business. According to Fayyad et al. [20], these results must be non-trivial, valid, novel, potentially useful, and ultimately understandable patterns to be able to be used in decision making. One of the reasons for which expert data miners are required is that the knowledge discovery in databases (KDD) process involves multiple stages [20], and regretfully, in each one, there is a large number of decisions that have to be taken with little or no formal guidance. The lack of a theoretical framework that unifies different data mining tasks [77] explains why the KDD process is said to be as much an “art” as it is “science” [45,71].

* Corresponding author. Tel.: +34 942 202063; fax: +34 942 201402.
E-mail addresses: marta.zorrilla@unican.es (M. Zorrilla), diego.garcia@unican.es (D. García-Saiz).

0167-9236/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2012.05.045

Except for some specific cases, business intelligence needs can be grouped in domain specific solutions as for example retail banking [27], insurance risk assessment [63], discovering web access patterns [34,81], selective marketing campaigns [8,71], acquiring and retaining customers [26,32], and so on. Since the information which companies have in their transactional systems as well as the questions they want answered have a lot in common, generic data mining models can be designed in order to satisfy the needs of all of them. One easy way to define these models is by means of templates, which specify the data set to be processed, the kind of result which is required (for instance a segmentation, a rule set or a predictive model), the preprocessing tasks to be carried out and the mining algorithms to be used. These templates would be defined by a data miner, expert in the business domain, and exploited by all the users who access the service proposed in this work.

As far as we know, there is no service in the cloud which allows an end-user to extract patterns and models by simply sending his data file without having to carry out the tedious job of selecting attributes, pre-processing and setting data mining algorithms. A service like this does not only offer non-expert data miners a tool for analysis but also facilitates the work of the expert data miners who can use it to obtain initial patterns easily and quickly.

In short, our objective in this paper is to describe a software architecture which meets the necessity of non-expert data miners to extract useful and novel knowledge using data mining techniques in order to obtain patterns which can be used in their business decision making process. Our proposal follows a service-oriented architecture with the aim of being easily configured and hosted in the web and can be deployed as an Analytic Software-as-a-Service. Furthermore, a service-oriented architecture implemented by means of Web Services facilitates its extension with new functionalities (services), developed by ourselves or by third-parties (through an orchestration of services). Another additional advantage that SOA offers is its design, based on layers, which allows the improvement of certain parts of the system without affecting the rest.

This paper is organized as follows. First, we write a preliminary section in which our interpretation of some concepts and terms used in the paper are explained. Next, we review the context of BI-as-a-Service and enumerate some currently available on-demand tools. Likewise, we relate other works published with a specific focus on the knowledge discovery process and discuss these in relation to our proposal. After that, we describe the architecture of our service and discuss some details about its implementation. In Section 4, we present an application which uses the proposed data mining service, called E-learning Web Miner, which allows virtual course instructors to extract knowledge from the clickstream stored in the e-learning platform logs. And, finally, we close by summarizing the contents of this chapter and discussing our future work.

2. Preliminaries

In the last few years, a set of terms and concepts have appeared which are not clearly and accurately defined in the software world and all of them are used profusely. We refer to terms such as business intelligence as-a-Service (BIaaS), Analytics as-a-Service, Software On-demand, business intelligence in the Cloud, service-oriented architecture (SOA), Web Service (WS) or service-oriented computing (SOC) among others. In this section, we do not attempt to define these terms but indicate the sense in which we understand and use them in this work.

When we talk about a data mining service we understand "service" as a software product which offers a solution or gives an answer to the needs of a customer, this being either a person or another software application. So, there are at least two parties involved, the service provider and the service consumer, although a third party could exist, a service broker, which would act as the intermediary. Here we link with the On-demand term which means, in our view, the ability for customers to have instant access to the service and pay for it based on usage, only if this service is not free. In general, these services are offered across the Internet and therefore, On-demand Software and Software as-a-Service (SaaS) are used as synonyms. According to the Software & Information Industry Association [65], SaaS applications are based on a recurring subscription fee and typically follow a pay-as-you-go model. However, according to Srinivasa [66], currently most SaaS are free, as for example, web applications for communication and collaboration offered by Google or recently Office Web Apps offered by Microsoft.

SaaS applications are characterized by being: easy to use, feature-rich, easy to access and they promise good consumer adaptation. Generally, SaaS is used to refer to business software rather than consumer software since this delivery model avoids the need to install and run the applications on the computer of the user and to carry out the maintenance and support tasks. So, the adaptation of the SaaS concept to provide business intelligence services is known as business intelligence-as-a-Service (BIaaS).

Another relevant characteristic of SaaS applications is that they run entirely in Cloud Computing which, according to NIST [41], is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. That means, Cloud Computing provides environments to enable resource sharing in terms of scalable infrastructures (storage, computing power, etc.), middleware (databases, operation systems, application servers), application development platforms, and value-added business applications which can be delivered through a SaaS model or as utilities, namely loosely coupled sub-processes inside customers' business processes.

Much like other software, SaaS can also take advantage of service oriented architecture (SOA) to allow software applications to communicate with each other. Each software service can act as a service provider, exposing its functionality to other applications via public brokers, and can also act as a service requester, incorporating data and functionality from other services.

Channabasavaiah et al. [7] defines SOA as follows: "SOA is an application architecture within which all application logic is defined as services, which can be called in defined sequences to form business processes". Erikson et al. [18] gather seven different definitions of SOA which come from organizations such as W3C, IBM, or OMG among others and conclude that SOA is commonly seen as a way of assembling, building or composing the information infrastructure of a business or organization. Additionally, SOA Manifesto [64] states that SOA is a type of architecture that results from applying service orientation, a new way of conceiving and designing the software, focused mainly on the business processes of an organization, known as Service-Oriented Computation (SOC). Unlike previous architectures, SOA focuses on business processes, rather than technical components.

Although an official set of service-orientation principles does not exist [19,44], there is a common set that is mostly associated with service orientation, namely loose coupling, autonomy, reusability, statelessness, abstraction, composability, and discoverability. In spite of the fact that there is a variety of technologies and standards to implement an SOA including RPC, DCOM, CORBA or WCF, Web Services are the most widely used [50]. One of the main reasons is that Web Services are based on open standards that are independent from any implementation platform. Some of these standards are: XML (eXtensible Markup Language) for writing data exchange files and messages which are used for the communication between the service requester and the service provider; XSD (XML Schema Definition) for providing a means of defining the structure, content, and semantics of XML documents; SOAP (Simple Object Access Protocol) for transporting these messages in an envelope across the Internet [13]; WSDL (Web Services Description Language) for describing the details of the service such as the