Information and Software Technology 51 (2009) 977–992

Contents lists available at ScienceDirect

Information and Software Technology

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

A UML profile for the conceptual modelling of data-mining with time-series in data warehouses

Jose Zubcoff [a,*], Jesús Pardillo [b], Juan Trujillo [b]

[a] Lucentia Research Group, Department of Sea Sciences and Applied Biology, University of Alicante, 03690 Alicante, Alicante, Spain
[b] Lucentia Research Group, Department of Software and Computing Systems, University of Alicante, 03690 Alicante, Alicante, Spain

ARTICLE INFO

Article history:
Received 6 March 2008
Received in revised form 25 August 2008
Accepted 7 September 2008
Available online 14 November 2008

Keywords:
Conceptual modelling
Multidimensional modelling
Data-mining
Time-series
Data warehouses
UML profiles

ABSTRACT

Time-series analysis is a powerful technique to discover patterns and trends in temporal data. However, the lack of a conceptual model for this data-mining technique forces analysts to deal with unstructured data. These data are represented at a low-level of abstraction and their management is expensive. Most analysts face up to two main problems: (i) the cleansing of the huge amount of potentially-analysable data and (ii) the correct definition of the data-mining algorithms to be employed. Owing to the fact that analysts' interests are also hidden in this scenario, it is not only difficult to prepare data, but also to discover which data is the most promising. Since their appearance, *data warehouses* have, therefore, proved to be a powerful repository of historical data for data-mining purposes. Moreover, their foundational modelling paradigm, such as, *multidimensional* modelling, is very similar to the problem domain. In this article, we propose a *unified modelling language* (UML) extension through UML profiles for data-mining. Specifically, the UML profile presented allows us to specify time-series analysis on top of the multidimensional models of data warehouses. Our extension provides analysts with an intuitive notation for time-series analysis which is independent of any specific data-mining tool or algorithm. In order to show its feasibility and ease of use, we apply it to the analysis of fish-captures in Alicante. We believe that a coherent conceptual modelling framework for data-mining assures a better and easier knowledge-discovery process on top of data warehouses.

1. Introduction

Computer systems currently collect a huge amount of time-stamped data within the context of scientific and business scenarios. Financial databases such as stock prices, web sites, point-of-sale systems, or scientific databases with sensor data, are just some examples of computer systems capturing huge amounts of data related to time. However, these data are stored in *operational databases* which are designed for transactional processing rather than knowledge discovery, i.e., they focus on the high performance of business transactions rather than on practical mechanisms with which to empower data analysis. Moreover, in complex projects in which knowledge discovery is required, data must be previously integrated from different operational data sources. This integration involves expensive pre-processing tasks if we are to obtain a cleansed and suitable data repository. As a result, the repository contains a huge amount of data which must be managed. This situation can be overcome by focusing solely on the problem domain, i.e., the data-mining process itself, rather than on the solution space in which the low-level details of a specific software technology makes any analysis unpractical. For instance, flat files are widely used as the data repository for time-series analysis.

The *knowledge discovery in databases* (KDD) process [1] is defined as the extraction of previously unknown and potentially useful information from large data sets. This process can be divided into three main phases (Fig. 1): (i) data integration (and pre-processing), (ii) data-mining, and finally (iii) knowledge discovery. The first stage corresponds to the most time-consuming phase in the KDD process. As a result, a unique data repository for analysis purposes is created. This repository, called the *data warehouse* was defined in [2] as "a subject-oriented, integrated, time-variant and non-volatile collection of data in support of a management's decision making process." There are three main advantages of using a data warehouse to perform data-mining with time-series analysis: (i) it contains data which are integrated and prepared for analysis; (ii) data are always associated with a particular time period; and (iii) the data are modelled under the *multidimensional modelling* paradigm which provides an intuitive view for analysis [3]. Thus, analysts can concentrate on data-mining rather than on cleansing and integrating data [4]. Nevertheless, time-series analysis is carried out more as an art than a science [5]. It is traditionally performed on top of flat files which do not explicitly represent the

* Corresponding author. Fax: +34 965 90 93 26.
E-mail addresses: Jose.Zubcoff@ua.es (J. Zubcoff), jesuspv@dlsi.ua.es (J. Pardillo), jtrujillo@dlsi.ua.es (J. Trujillo).

0950-5849/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.infsof.2008.09.006

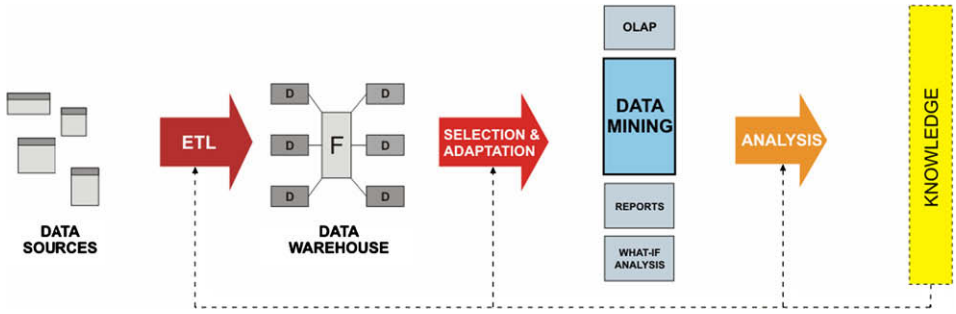

Fig. 1. Phases of the KDD process with multidimensional models of data warehouses.

underlying complex data relationships. This kind of unstructured data is very difficult to manage: flat files usually have hundreds of attributes and millions of observed values, which are unsuitable for large projects.

We believe that the coherent conceptual modelling framework, together with its corresponding tools, facilitate the success of KDD. Thus, taking the time-series analysis into account from the early stages of the data warehouse development assures data quality. Moreover, an isolated interpretation of the KDD sub-processes may lead analysts to duplicate time-consuming tasks. Unfortunately, data miners have not considered methods or models from a conceptual perspective, and have concentrating solely upon the implementation of the corresponding algorithm on top of unstructured raw data.

Two standards have been presented in relation to data-mining modelling: the *common warehouse model* (CWM) [6] and the *predictive model markup language* (PMML) [7]. The former proposes a data-mining model focused on metadata interchange. The latter uses the *extensible markup language* (XML) to describe statistical and data-mining schema. However, neither proposes a model with which to specify time-series analysis.

In our previous works, we have presented the conceptual modelling for the first phase of KDD: the *extraction transformation load* (ETL) processes [8], and the data-warehouse repository [9]. These works deal with the integration and multidimensional modelling of data warehouses (Fig. 1). By following the KDD process, we have proposed a model-based data-mining with association rules [10,11], classification [12], and clustering [13] on top of multidimensional models of data warehouses.

1.1. Contribution

In this article, we propose an extension of the well-known *unified modelling language* (UML) [14] through the use of UML profiles in order to allow users to specify data-mining time-series models on top of multidimensional models of data warehouses. This takes place in an integrated manner from the early stages of a data warehouse project. Our framework is independent of a particular software platform, thus providing intuitive modelling elements which are focused on the problem domain. In addition, time-series analysis models defined at the conceptual level can be easily mapped into low-level models that implement them in a given data-mining platform.

1.2. Outline

The remainder of this article is structured as follows: Section 2 introduces the fundamentals of time-series analysis. Section 3 summarises the multidimensional modelling based on UML. In Section 4, we present our new UML extension for the design of time-series mining models within the context of multidimensional modelling. Section 6 outlines a case study and applies our framework to the design of time-series analysis of multidimensional data. Section 7 discusses related works. Finally, Section 8 presents conclusions and outlines future work.

2. Fundamentals of times-series analysis

In a time-series analysis, time is the main variable used to describe the data under analysis. Nevertheless, a time period can be annotated using different scales. For instance, in a point-of-sales system, sales are recorded using the lowest level required timestamp. However, for analysis purposes, analysts can aggregate data by days, weeks, months, and so on, according to their needs (this issue can easily be addressed by the multidimensional model of the data warehouse). Hence, an inherent constraint of this data-mining technique is that it must contain at least one time attribute at any level of granularity.

Most time-series patterns can be described in terms of two basic concepts:

Trend: This concept represents a general systematic linear or (more often) nonlinear component that changes over time. It does not repeat, or at least it does not repeat within the time range captured in the data.
Seasonality: This can be identified by regularly spaced peaks and troughs. These have a consistent direction and approximately the same magnitude every year which is relative to the trend.

For instance, Fig. 2 represents the fish-capture time series in Alicante during the last three years. It also shows the trend (straight line) and seasonality (cyclical behaviour of the values in a calendar year, in this case). This is a real case study which will be explained in greater detail in Section 6.

Several algorithms can be used to reveal hidden temporal patterns, such as: *autoregressive integrated moving average* (ARIMA) [15], *autoregressive tree* (ART) [16], exponential smoothing [17], or *vector autoregression* (VAR) [18]. These algorithms require specific settings which must be tuned during the discovery of patterns.

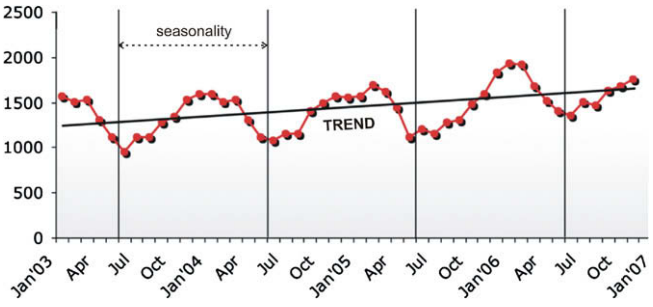

Fig. 2. Time-series representing the fish-captures in Alicante from 2002 to 2006.