STUDIA INFORMATICA 2003
Volume 24 Number 2A (53)

Janusz ŚWIERZOWICZ
Politechnika Rzeszowska, Zakład Informatyki

IMPACT OF DATA MINING STANDARDIZATION ON INFORMATION TECHNOLOGY DEVELOPMENT

Summary. This paper examines the objective assumptions for Data Mining Process standardization, which simplifies integration of Information Systems with Data Mining models. In doing so it provides an overview of the most important characteristics of Cross Industry Standard Process Model for Data Mining (CRISP-DM), Application Programming Interface OLE DB for Data Mining (API OLE DB DM), and Predictive Model Markup Language (PMML). These standards have essential impact on Information Technology development.

Keywords: data mining, standardization.

WPŁYW STANDARDÓW EKSPLORACJI DANYCH NA ROZWÓJ TECHNOLOGII INFORMATYCZNEJ

Streszczenie. Artykuł przedstawia obiektywne założenia standaryzacji procesu eksploracji danych. Standaryzacja upraszcza integrację systemów informacyjnych z modelami eksploracji danych. Przedstawiono charakterystyki CRISP – DM, API OLE DB DM oraz PMML. Standardy te mają istotny wpływ na rozwój technologii informatycznej.

Słowa kluczowe: eksploracja danych, standaryzacja.

1. Introduction

Information Technology development has strong influence on data resources. In this environment of fast rising volumens of data, human abilities in memory capacities and low data complexity or dimensionality analysis cause data overload problem. It is impossible to solve this issue in a human manner – it takes strong effort to use intelligent and automatic software

tools for turning rough data into valuable information [2-7, 9-10]. One of the central activities associated with understanding, navigating and exploring the world of digital data is Data Mining. It is an intelligent and automatic process of identifying and discovering useful structures in data such as patterns, models and relations. We can consider Data Mining as a part of the overall Knowledge Discovery in Data process, which is defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [4], it should support us as we struggle to solve data overload and complexity issues.

Data mining applications have to process data of diverse nature, drawn from different storage architectures, use multiple data-specific exploration algorithms, and present results in a variety of forms. Data mining processes and models are used as a part of commercial Information Systems including those in Enterprise Resource Planning, Customer Relationship Management and in processing engineering and scientific data as well. With the fastest acceleration of online data resources in the Internet, the World Wide Web is a natural domain for using data mining techniques to automatically discover and extract actionable information from Web documents and services, especially in e-business. We have named those techniques as Web Mining. We also consider text mining as a data-mining task that helps us summarize, cluster, classify and find similar text documents.

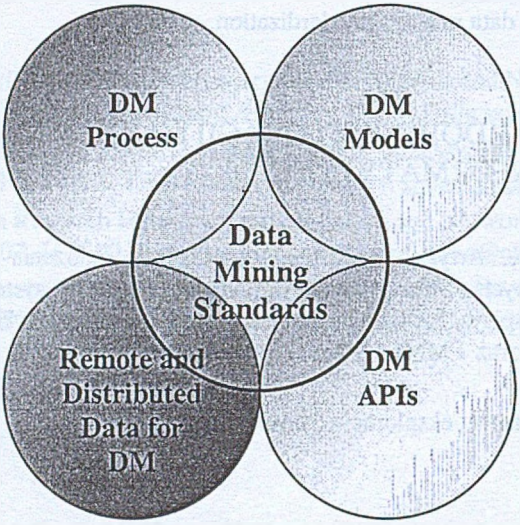

Fig. 1. Data mining standards in various aspects
Rys. 1. Różne aspekty standardów eksploracji danych

Technological standards play an important role in Information Technology development [7]. Now, many organizations are developing technological standards for various aspects of data mining. Several standardization efforts [6] are undertaken on models,

attributes, application programming interfaces, processing of remote and distributed data as depicted in Fig. 1. This issue is discussed in following chapters.

2. Cross Industry Standard Process Model for Data Mining - CRISP DM

CRISP DM was developed in the year 2000 by the consortium of data mining vendors and advanced users (e.g. SPSS, NCR Daimler-Benz, Mercedes-Benz and OHRA) [2]. The CRISP-DM applies across different industry sectors (e.g. automotive, aerospace, insurance) was designed to make data mining projects easily adopted as a key part of business processes. The main assumption in this model preparation was its neutrality with respect to industry, method, tool and application. It consists of tasks described at four levels of abstraction as

- phases,
- generic tasks,
- specialized tasks
- process instances.

At the top level, the data mining process is organized into the following phases (see Fig. 2.):

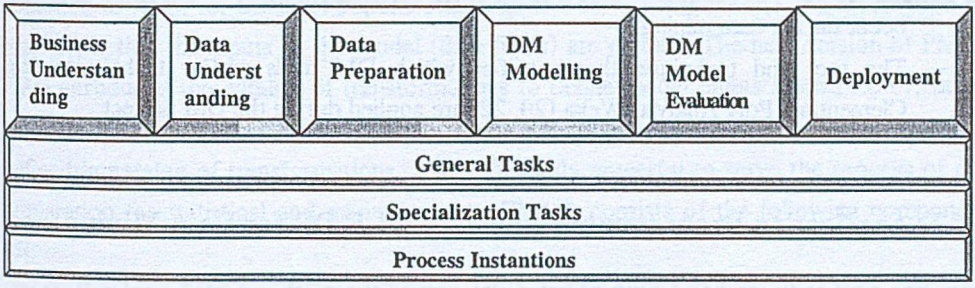

Fig. 2. CRISP DM – phases and levels
Rys. 2. Fazy i poziomy modelu CRISP DM

- Business understanding that focuses on understanding the project objectives and requirements from business perspective.
- Data understanding that includes initial data collection, identification of data quality problems and detection interesting data subset to form hypotheses for hidden valuable information.
- Data preparation that covers construction of the data set for modeling tools. This phase focuses on tables, records and attributes selection as well as transformation and cleaning of data.

- Modeling that focuses on selection of various modeling techniques and on tuning for values of optimal parameters.
- Evaluation of the model quality with respect to achieving the business objectives.
- Deployment that involves applying models within a decision-making process in organization. It takes simple forms like report generation as well as repeatable mining process.

The second level is the level of generic tasks. It was introduced to cover whole data mining process, all possible data mining applications and new modeling techniques.

The third level is the specialized tasks. It describes how the general task differs in various situations.

The last but not least is the process instance level. It is a record of the actions, decision and results of actual data mining engagement.

CRISP-DM distinguishes between following dimensions of data mining context:

- The application domain (e.g., banking, education, Customer Relationship Management [1-2, 6, 18-22]) is the area in which project takes place.
- The data mining problem type (e.g., data description and summarization, segmentation, concept descriptions, classification, prediction, dependency analysis, etc.) describes specific classes of objectives that the mining process deals with.
- The technical aspect (e.g., missing values) describes technical challenges that usually occur during data mining.
- The tool and technique that specifies which DM tools and/or techniques (e.g., Clementine, Poly Analyst, Weka [20, 22] are applied during the DM project.

3. Application Programming Interface OLE DB for Data Mining

The API OLE DB DM is an example of a new protocol that simplifies communication and provides better integration of data mining tools with data based management applications. It was released in 2000 by Microsoft [13, 16, 17]. It defines a data mining API for SQL Server 2000 data and supports the basic data mining operations such as defining a mining model, populating a mining model from training data set, predicting attributes for new data and browsing a mining model for reporting and visualization applications. A virtual object that is similar to a table (Data Mining Model, DMM) can be created with CREATE statement, browsed with SELECT, populated with INSERT INTO, reset via the DELETE statement, refined or used to derive prediction with PREDICTION JOIN statement. A fundamental

operation is the training of DMM, follow by use of the model to derive prediction. The operation is executed in the following steps:

- Create an OLEDB data source and obtain an OLE DB session object
- CREATE MINING MODEL ...
- INSERT INTO //training data into the model
- SELECT ...

FROM .....

PREDICTION JOIN

4. Predictive Model Markup Language (PMML)

Predictive Model Markup Language (PMML), presented by the Data Mining Group [15], is the most workable standard of the data mining. It is based on the XML language for description of models, inputs, and transformations used for data preparation and parameters for defining of data mining models. The aim of the PMML is to provide a suitable infrastructure for an application that will be able to generate a model - as the PMML producer - as well as different application that will be able to consume it (the PMML consumer) by reading the PMML data file. The first version of PMML (v.1 1999) was concentrated on the definition of XML for the most popular statistical and data mining models. An assumption has been made that the inputs to the model (data fields) are defined. The next version of PMML v.2.0 introduced the catalog of transformations to define model inputs more flexibly. In this case inputs to the model can be either data fields, defined in a catalog of data, or derived fields defined in catalog of transformations. It is sufficiently powerful to serve the process of data preparation for statistical and mining models. PMML consists of the following components [6]:

- Catalog of data – defines the input attributes to model and specifies type and range value. It contains data definitions that do not vary by the model.
- Mining schema – precisely one for each model, listing the attributes of the schema and their role in the model. There are a subset of attributes in the data catalog. The schema contains information specific to a certain model and also specifies an attribute's usage type. The usage type can be active (as an input of the model), predictive (as an output of the model), or supplementary (as a descriptive information, ignored by the model).
- Catalog of transformations – that contain one of the following transformations:
  - standardization (mapping continuous or discrete values to numbers),
  - discretization (mapping continuous values to discrete values),

- value mapping (mapping discrete values to discrete values),
- aggregation (summarizing or collecting group of values).

o Statistics of the model – univariate statistics about the attributes of the model.

o Models – parameters of the model specified by tags. We can use: decision tree, neural networks, regression models, cluster models, Bayesian models, association rules, and sequence models.

PMML is used for a wide variety of software, including applications in e-business, direct marketing, finance, manufacturing and defense, in products released by such vendors as Angoss, IBM, Magnify, Microsoft, MINEIT, NCDM, NCR, Oracle, Salford Systems, SPSS, SAS and Xchange [15]. The current standard supports several predictive model types that cover the most popular data mining methods used in contemporary data mining tools [12-18].

5. Other Data Mining Standards

To assure the integration of data mining process with multimedia information systems Application Programming Interface SQL/MM [6] has been developed.

Java Specification Request –73 (JSR-73) [8] determines programming interface in Java language to build and compute of data mining model. The usage of the JSR-73 standard makes possible a software service as well as data and metadata access connected with data mining results. Common Warehouse Model for Data Mining has been defined by OMG Group [3]. The CWM DM defines metadata for model representations, its settings and results of data mining operations. Models are described through the Universal Modeling Language (UML) [23]. They use XML Document Type Definition for formal description of XML documents.

Standards mentioned above find a practical solution in Oracle 9i Data Mining Application Programming Interface (ODM API). It enables Java programs to access Data Mining Server [14].

Standardization has been also extended to program objects for the data mining definition; data mining used in business processes as well as the internet services for the mining of remote and distributed data [6].

6. Conclusion

The main reason why there are so many representations and communication standards is that data mining is used on many different ways in various domains. In combination with

variety of systems and services, it often results in incompatible solutions. It is possible to observe efforts of databases and software vendors to unificate a nomenclature and integration of data mining standards.

PMML, basing on XML, is a common platform for several emerging standards. SQL / MM, JSR -73 CWM, SQL Server 2000 Analytic Services and Oracle Data Mining use the PMML in their specifications, assuring the basic level of compatibility.

The main data mining research is connected with standardization of data cleaning, transforming and preparation as well as the coordination for running the common internet services that use remote and distributed data.

User participation in the standardization process is becoming more important. This issue should be also considered in the process of selection of data mining methods and tools.

REFERENCES

1. Berry M., Linoff G.: Data Mining Technique. John Wiley & Sons, Inc, New York 1997.
2. Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shaerer C., Wirth R.: CRISP-DM 1.0. Step -by - step data mining guide. CRISP-DM Consortium, 2000.
3. Common Warehouse Metamodel: Data Mining. Object Management Group; Cgi.omg.org/cgi-bin/doclist.pl
4. Fayyad U., Piatetsky-Shapiro G., Smyth P., Uthurusamy R.: From data mining to knowledge discovery: An overview. Fayyad U. et all (ed): Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, Menlo Park, CA 1996, pp. 1-34.
5. Fayyad U.: The Digital Physics of Data Mining. CACM, March, 2001, Vol. 44, No. 3 pp. 62-65.
6. Grossman R. L., Hornick M.F., Meyer G.: Data Mining Standards Initiatives, CACM, August, 2002 Vol. 45, No. 8 pp. 59-61.
7. Jacobs K.: Global Aspect of Information Technology Standards and Standardization. Information Management, Vol. 15. No. 1/2, 2002, pp. 8-35.
8. Java Specification Request 73, jcp.org/jsr/detail/073.jsp
9. Landauer T. K.: How much do people remember? Some estimates of the quantity of learned information in long-term memory. Cognitive Science, 10 (4) pp. 477-493 (Oct-Dec 1986).
10. Leavitt N: Data Mining for the Corporate Masses? Computer, May, 2002 Vol. 35, No. 5 pp. 22-24.
11. Liautaud B.: e-Business Intelligence: Turning Information into Knowledge into Profit. McGraw-Hill, New York 2001.

12. Kumar N., Gangopadhyay A.: Discovering Valuable Patterns through Internet Web-Log Access Analysis. Information Technology and Organizations: Trends, Issues, Challenges and Solutions. Khosrow-Pour M. (ed). Idea Group Publishing, Hershey, London, Melbourne, Singapore, Beijing 2003.
13. OLE DB for Data Mining Specification 1.0. Microsoft; www.microsoft.com/data/ole-db/default.htm
14. Oracle9i Data Mining Concepts Release 2 (9.2) Part Number A95961-01.
15. Predictive Model Markup Language (PMML). Data Mining Group; www.dmg.org
16. Seidman C.: Data Mining with Microsoft® SQL Server 2000 Technical Reference. Microsoft Press, 2000.
17. SQL Server 2000 Analyzis Services; www.microsoft.com/SQL/techinfo/bi/analysis.asp
18. Świerzowicz J.: Analysis of Current Data Mining Standards. Information Technology and Organizations: Trends, Issues, Challenges and Solutions, Khosrow-Pour M. (ed). Idea Group Publishing, Hershey, London, Melbourne, Singapore, Beijing 2003.
19. Świerzowicz J.: A Management Information System for Classification of Scientific Achievements, Evolution and Challenges in System Development, Zupancic et all (ed). Kluwer Academic/Plenum Publishers, New York 1999, pp. 735-740.
20. Świerzowicz J.: Decision Support System for Data and Web Mining Tools Selection, Issues and Trends of Information Technology Management in Contemporary Organizations, Khosrow-Pour M. (ed). Idea Group Publishing, Hershey, London, Melbourne, Singapore, Beijing 2002, pp. 1118-1120.
21. Świerzowicz J: Current Standards for Data Mining Process. Soft Computing and Distributed Processing, Six International Conference, SCDP 2002, UITM, Rzeszów 2002, pp. 120-122.
22. Świerzowicz J: Decision Support System for Data Webmining Tools. International Conference on E-Business and Web Technologies, UITM, Rzeszów 2001.

Recenzent: Dr inż. Marcin Gorawski

Wpłynęło do Redakcji 31 marca 2003 r.

Streszczenie

Standardy technologiczne odgrywają istotną rolę w rozwoju technologii informacyjnej. Wiele producentów oraz użytkowników baz danych rozwija technologiczne standardy eksploracji danych. Różnorodne aspekty standaryzacji dotyczące modeli, atrybutów,

interfejsów programowania aplikacji, przetwarzania danych zdalnych i rozproszonych przedstawiono schematycznie na rys. 1.

Standardowy model procesu dla drążenia danych - CRISP - DM jest neutralny w stosunku do dziedziny zastosowań, metody, narzędzi i aplikacji. Na rys. 2 wyróżniono zadania na poziomie faz, zadań ogólnych i wyspecjalizowanych oraz instancji procesów. Model ten zaprojektowano w taki sposób, aby można było łatwo wdrożyć proces eksploracji danych jako kluczowy czynnik procesu gospodarczego.

Interfejs API OLE DB DM jest przykładem protokołu, który upraszcza komunikację i dostarcza lepszej integracji narzędzi eksploracji danych z aplikacjami zarządzania bazą danych. Zaprojektowano go dla usług analitycznych SQL Servera 2000 w celu transformacji danych, drążenia i bieżącego przetwarzania analitycznego OLAP.

PMML jest jednym z najczęściej wdrażanych standardów dla opisu modeli eksploracji danych, wejść, transformacji oraz dla przygotowania danych i parametrów dla definiowania modeli. Jego celem jest zapewnienie odpowiedniej infrastruktury dla aplikacji, która będzie zdolna wygenerować model oraz innej aplikacji zdolnej do skonsumowania modelu. PMML zawiera katalog danych, schemat eksploracji, katalog transformacji, statystyki modelu i modele. Jest stosowany w wielu aplikacjach, a modele obsługują najpopularniejsze metody eksploracji danych, wykorzystywane we współczesnych narzędziach analizy i obsługi baz danych.

Główną przyczyną występowania rozmaitych reprezentacji danych i standardów komunikacyjnych jest fakt, że eksploracja jest stosowana w różnorodnych dziedzinach dla danych strukturalnych, semistrukturalnych i tekstowych. Daje to w kombinacji z wieloma systemami i usługami często niekompatybilne rozwiązania. Można zaobserwować wysiłki czołowych producentów baz danych zmierzające do ujednolicenia terminologii i integracji standardów. Główne kierunki badań dotyczą standardów czyszczenia, transformowania i przygotowania dla danych oraz uzgodnienie wspólnego zestawu usług internetowych dla pracy z danymi zdalnymi i rozproszonymi. Opisane standardy mają istotny wpływ na dalszy rozwój technologii informatycznej.

Adres

Janusz ŚWIERZOWICZ: Politechnika Rzeszowska, Wydział Budowy Maszyn i Lotnictwa, Zakład Informatyki, ul. W.Pola 2, 35-959 Rzeszów, Polska, jswierz@prz.rzeszow.pl .