INTERNATIONAL STANDARD

ISO/IEC 13249-6

First edition
2002-12-01

Information technology — Database languages — SQL multimedia and application packages —

Part 6: Data mining

Technologies de l'information — Langages de bases de données — Multimédia SQL et paquetages d'application —
Partie 6: Exploration de données

Reference number
ISO/IEC 13249-6:2002(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Figures

Tables

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

ISO/IEC 13249-6 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 32, *Data management and interchange*.

ISO/IEC 13249 consists of the following parts, under the general title *Information technology — Database languages — SQL multimedia and application packages*:

— *Part 1: Framework*

— *Part 2: Full-Text*

— *Part 3: Spatial*

— *Part 5: Still Image*

— *Part 6: Data mining*

Annexes A, B and C of this part of ISO/IEC 13249 are for information only.

Introduction

The purpose of this International Standard is to define multimedia and application specific types and their associated routines using the user-defined features in ISO/IEC 9075.

SQL/MM is structured as a multi-part standard. At present it consists of the following parts:

Part 1: Framework

Part 2: Full-Text

Part 3: Spatial

Part 5: Still Image

Part 6 : Data mining

The organization of this part of ISO/IEC 13249 is as follows:

1) Clause 1, "Scope", specifies the scope of this part of ISO/IEC 13249.

2) Clause 2, "Normative references", identifies additional standards that, through reference in this part of ISO/IEC 13249, constitute provisions of this part of ISO/IEC 13249.

3) Clause 3, "Definitions, notations, and conventions", defines the notations and conventions used in this part of ISO/IEC 13249.

4) Clause 4, "Concepts", presents concepts used in the definition of this part of ISO/IEC 13249.

5) Clause 5, “Data Mining Model Types”, defines the user-defined types and associated routines for data mining models.

6) Clause 6, “Data Mining Settings Types”, defines the user-defined types and associated routines for data mining settings.

7) Clause 7, “Data Mining Application Result Types”, defines the user-defined types and associated routines for the results of the application of a data mining model.

8) Clause 8, “Data Mining Test Result Types”, defines the user-defined types and associated routines for the results of a test of a data mining model.

9) Clause 9, “Data Mining Data Types”, defines the user-defined types and associated routines for the data definitions of data mining.

10) Clause 10, “Status Codes “, defines the SQLSTATE codes used in this part of ISO/IEC 13249.

11) Clause 11, “Conformance “, defines the criteria for conformance to this part of ISO/IEC 13249.

A) Annex A, “Implementation-defined elements”, is an informative Annex. It lists those features for which the body of this part of ISO/IEC 13249 states that the syntax or meaning or effect on the database is partly or wholly implementation-defined, and describes the defining information that an implementor shall provide in each case.

B) Annex B, “Implementation-dependent elements”, is an informative Annex. It lists those features for which the body of this part of ISO/IEC 13249 states explicitly that the meaning or effect on the database is implementation-dependent.

C) Annex C, “A scenario using the classification technique”, is an informative Annex. It presents explanatory material on how to use this part of 13249.

In the text of this part of ISO/IEC 13249, Clauses begin a new odd-numbered page, and in Clause 5, "Data Model Types", through Clause 9, "Data Mining Data Types", subclauses begin a new page. Any resulting blank space is not significant.

Information technology — Database languages — SQL multimedia and application packages —

Part 6: Data mining

1 Scope

This part of ISO/IEC 13249:

a) introduces the Data mining part of ISO/IEC 13249,

b) gives the references necessary for this part of ISO/IEC 13249,

c) defines notations and conventions specific to this part of ISO/IEC 13249,

d) defines concepts specific to this part of ISO/IEC 13249,

e) defines data mining user-defined types and their associated routines.

The data mining user-defined types defined in this part adhere to the following:

— A data mining user-defined type is generic to data mining data handling. It addresses the need to store, manage and retrieve information based on elements such as data mining models, data mining settings, and data mining test results.

— A data mining user-defined type does not redefine the database language SQL directly or in combination with another data mining data type.

2 Normative references

The following normative documents contain provisions which, through reference in this text, constitute provisions of this part of ISO/IEC 13249. For dated references, subsequent amendments to, or revisions of, any of these publications do not apply. However, parties to agreements based on this part of ISO/IEC 13249 are encouraged to investigate the possibility of applying the most recent editions of the normative documents indicated below. For undated references, the latest edition of the normative document referred to applies. Members of ISO and IEC maintain registers of currently valid International Standards.

2.1 International standards

ISO/IEC 9075-1, *Information technology — Database languages — SQL — Part 1: Framework (SQL/Framework)*

ISO/IEC 9075-2:1999, *Information technology — Database languages — SQL — Part 2: Foundation (SQL/Foundation)*

ISO/IEC 9075-4:1999, *Information technology — Database languages — SQL — Part 4: Persistent Stored Modules (SQL/PSM)*

ISO/IEC 13249-1, *Information technology — Database languages — SQL multimedia and application packages — Part 1: Framework*

2.2 Publicly-available specifications

Extensible Markup Language (XML) 1.0 (Second Edition). 2 October 2000, http://www.w3.org/TR/REC-xml

Predictive Model Markup Language (PMML) 1.1, http://www.dmg.org/pmml-v1-1.htm

3 Definitions, notations, and conventions

3.1 Definitions

3.1.1 Definitions provided in Part 1

This part of ISO/IEC 13249 makes use of all terms defined in part 1 of ISO/IEC 13249.

3.1.2 Definitions provided in Part 6

For the purposes of this part of ISO/IEC 13249, the following definitions apply.

3.1.2.1
active field
An input field that is used during the data mining run to compute a data mining model, evaluate or predict a single input row.

Note: Active fields are defined in the logical data specification.

3.1.2.2
application phase
Phase, during which an input row is evaluated against a data mining model and one or more values are computed: the identification of the cluster for data mining clustering, the predicted value for data mining classification or data mining regression.

Note: The application phase is only used for data mining clustering, classification and regression now.

3.1.2.3
association rule
Combination of field values, which appears frequently (according to a specified threshold), i.e. there is a significant number of groups containing the combination of the values.

Example of application: store layout.

3.1.2.4
association rule support
The percentage of transactions that contain all items of an association rule.

Note: The association rule support of the association rule “X and Y implies Z” is 100 times the number of transactions that contain X, Y and Z divided by the total number of transactions.

3.1.2.5
categorical field type
A field being of categorical field type supports only the <equal operator> as a comparison operator.

Note: There is no defined order, and no arithmetic operations are supported.

3.1.2.6
class label
Target field used in data mining classification to create a classification model, whose application to data without a class label allows it to predict a value for this class label.

Note: The class label must have a categorical field type.

3.1.2.7
classification cost rate
Maximal classification error tolerated during the validation phase of data mining classification, above which the classification model is considered not good enough.

3.1.2.8
classification error
Percentage of wrongly predicted values among the total number of values predicted during the validation phase of the classification technique.

Note: This error is returned at the end of the test phase.

3.1.2.9
classification model
The result of data mining classification.

3.1.2.10
data mining application task
Abstraction for the complete data needed to apply a data mining model, containing values of the data mining model, the data mining data for input and output, and the result field used for the operation.

3.1.2.11
data mining classification
Data mining technique that computes a classification model such that the class label value can be predicted by using the model and input field values without the class label value.

Note: In this document, the term “classification” is sometimes used as a synonym for “data mining classification” where the meaning is unambiguous from the context.

Example of application: insurance risk prediction.

3.1.2.12
data mining cluster
A set of input rows with common characteristics (see data mining clustering).

Note: In this document, the term “cluster” is sometimes used as a synonym for “data mining cluster” where the meaning is unambiguous from the context.

3.1.2.13
data mining clustering
A data mining technique that discovers sets of input rows with common characteristics - the *clusters*, so that rows are possibly homogeneous inside a cluster and possibly heterogeneous between two clusters.

Note: In this document, the terms “cluster” or “segment” are sometimes used as a synonym for “data mining cluster” where the meaning is unambiguous from the context.

Example of application: customer mailings.

3.1.2.14
data mining data
Description of data contained in tables that represents the metadata required for access to the data for data mining training, test or application runs.

3.1.2.15
data mining data type
A data mining data type represents an abstraction of a table. A data mining data type maps the input field names (i.e. column names of the table) to alias names. These alias names will be compared, before the data mining run, to the field names used in the specified logical data specification.

3.1.2.16
data mining model
The result of a data mining run of a data mining technique over a given set of data, which contains the usable correlations discovered in the data.

Note 1: The data mining model can also be used alone as a substitute of the original large amount of data to associate, classify or predict the behavior of additional data (in application and testing phases) .

Note 2: In this document, the terms “mining model” or “model” are sometimes used as a synonym for “data mining model” where the meaning is unambiguous from the context.

3.1.2.17
data mining regression
Data mining technique similar to data mining classification except for the numerical type of the target field, which computes a regression model allowing to predict a numerical value.

Note 1: The predicted value might not be identical with any value contained in the data used to build the model.

Note 2: In this document, the term “regression” is sometimes used as a synonym for “data mining regression” where the meaning is unambiguous from the context.

Example of application: customer ranking.

3.1.2.18
data mining run
Computation of a data mining model using data mining settings.

Note: According to the data mining technique, a data mining run may contain a training phase, a validation phase and a testing phase. A data mining run is called a training run when it contains a training phase, and a test run when it contains only a test phase.

3.1.2.19
data mining segmentation
see data mining clustering.

3.1.2.20
data mining settings
Settings used in the data mining run that specify the logical data specification and parameters depending on the data mining technique.

Note: In this document, the terms “mining settings” or “settings” are sometimes used as a synonym for “data mining settings” where the meaning is unambiguous from the context.

3.1.2.21
data mining task
Abstraction for the complete data needed to compute, test or apply a data mining model, containing the data mining settings and the data mining data type.

Note: In this document, the terms “mining task” or “task” are sometimes used as a synonym for “data mining task” where the meaning is unambiguous from the context.

3.1.2.22
data mining technique
Four types of data mining techniques are distinguished: discovering association rules, clustering/segmentation, classification and regression.

3.1.2.23
data mining test result
Statistical result returned at the end of the testing phase that contains an evaluation of the correctly and incorrectly predicted values in the test data.

Note 1: This result can be considered a quality criterion of the data mining model.

Note 2: In this document, the term “test result” is sometimes used as a synonym for “data mining test result” where the meaning is unambiguous from the context.