

Pre-processing Educational Data

Cristóbal Romero

Córdoba University, Campus Universitario de Rabanales, 14071, Córdoba, Spain
cromero@uco.es

Preprocessing EduData

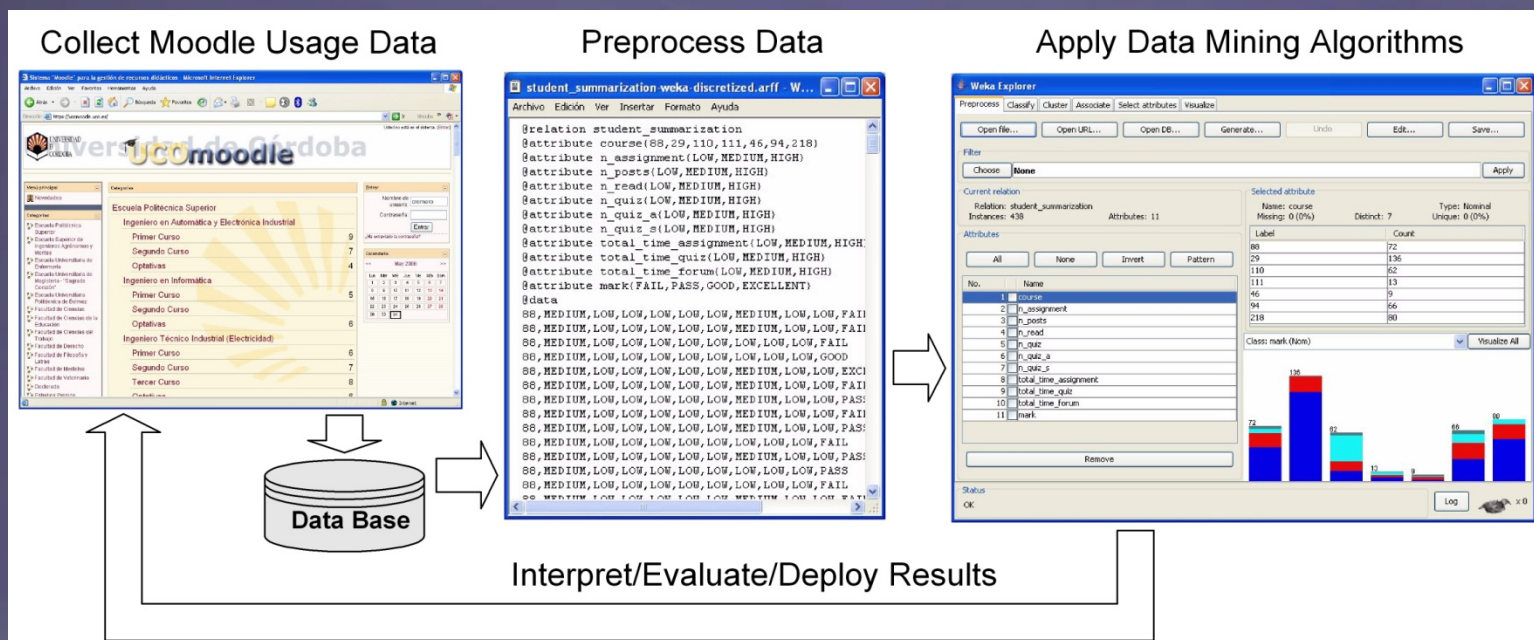
Introduction

- The first step in any KDD process is the transformation of data into an appropriate form for the mining process.
- Data pre-processing in educational context is considered the most crucial phase in the whole educational data mining process, and it can take more than half of the total time spent in solving the data mining problem.
- The data pre-processing phase typically consumes 60-80% of the time of the KDD process.

Preprocessing EduData

Introduction

- Data Mining Process with Moodle data:



Preprocessing EduData

MoodleDatabase

- Moodle provide a Data Manipulation API:
https://docs.moodle.org/dev/Data_manipulation_API
- **Tables** in Moodle database about student interaction:

Name	Description
mdl_user	Information about all the users.
mdl_user_students	Information about all students.
mdl_log	Logs every user's action.
mdl_assignment	Information about each assignment.
mdl_assignment_submissions	Information about assignments submitted.
mdl_forum	Information about all forums.
mdl_forum_posts	Stores all posts to the forums.
mdl_forum_discussions	Stores all forum discussions.
mdl_message	Stores all the current messages.
mdl_message_reads	Stores all the read messages.
mdl_quiz	Information about all quizzes.
mdl_quiz_attempts	Stores various attempts at a quiz.
mdl_quiz_grades	Stores the final quiz grade.

Preprocessing EduData

Exporting student Moodle course data

- Back up course content
- Export gradebook
- Export reports (logs)

Preprocesing EduData

Back up course content

Ajustes -> (Mas -> Administracion) -> Copia de seguridad

Página Principal / Mis cursos / Másteres / TRANSVERSALES MÁSTERES UNIVERSITA

Administración del curso

Administración del curso Usuarios

- Editar ajustes
- Activar edición
- Finalización del curso
- Filtros
- Configuración Calificac
- Copia de seguridad
- Restaurar
- Importar
- Reiniciar
- Archivos de curso here

- ⚙ Editar ajustes
- ✎ Activar edición
- ⚙ Finalización del curso
- ⏴ Filtros
- ⚙ Configuración Calificaciones
- 📄 Copia de seguridad
- ⬆ Restaurar
- ⬆ Importar
- ⬅ Reiniciar
- 📁 Archivos de curso heredados
- ⚙ Más ...

1. Ajustes iniciales ▶ 2. Ajustes del esquema ▶ 3. Confirmación y revisión ▶ 4. Ejecutar copia de seguridad ▶ 5. Completar

Configuración de la copia de seguridad

IMS Common Cartridge 1.0

Incluir usuarios matriculados

Hacer anónima la información de usuario

Incluir asignaciones de rol de usuario

Incluir actividades y recursos

Incluir bloques

Incluir filtros

Preprocesing EduData

Export gradebook

Ajustes -> Mas -> Administración -> Configuración Calificaciones ->Exportar

Exportar a Hoja de cálculo Excel

Vista Configuración Escalas Letras Importar **Exportar**

Hoja de cálculo OpenOffice Exportar a SIGMA Archivo en texto plano **Hoja de cálculo Excel** Archivo XML

▼ **Ítems de calificación a incluir**

- Envío de prácticas (opcional)
- Examen Enero 2020
- Examen Febrero 2019
- Sala del Curso
- Total del curso

Seleccionar todos/ninguno

► **Opciones de los formatos de exportación**

Descargar

Preprocessing EduData

Export reports (logs)

Ajustes -> Administracion -> Informes -> Registros

Fundamentos de Informatica (GIELE) ▾ Todos los participantes ▾ Todos los días ▾

Todas las actividades ▾ Todas las acciones ▾ Todos los recursos ▾ Todos los eventos ▾ [Conseguir estos registros](#)

1 2 3 4 5 6 7 8 9 10 ... 54 »

Hora	Nombre completo del usuario	Usuario afectado	Contexto del evento	Componente	Nombre evento	Descripción	Origen	Dirección IP
10 de December de 2020, 11:04	Cristobal Romero Morales	-	Curso: Fundamentos de Informatica (GIELE)	Registros activos	Informe de registro en tiempo real visualizado	The user with id '3826' viewed the live log report for the course with id '679'.	web	150.214.117.251
10 de December de 2020, 11:00	Cristobal Romero Morales	-	Curso: Fundamentos de Informatica (GIELE)	Sistema	Curso visto	The user with id '3826' viewed the course with id '679'.	web	150.214.117.251
9 de December de 2020, 15:59	Cristobal Romero Morales	-	Archivo: Práctica 4	Recurso	Módulo de curso visto	The user with id '3826' viewed the 'resource' activity with course module id '3727'.	web	172.30.178.202

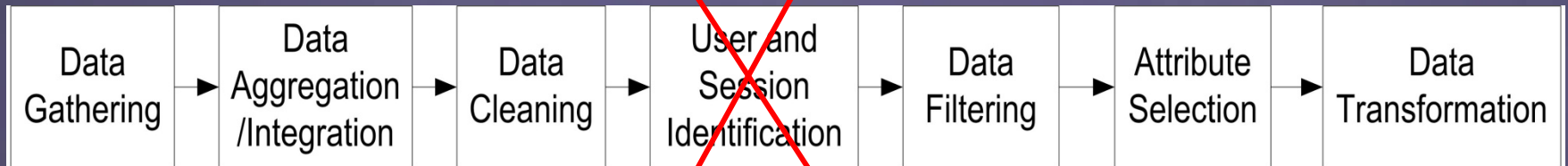
Descargar datos de tabla como ▾ [Descargar](#)

1 2 3 4 5 6 7 8 9 10 ... 54 »

Preprocessing EduData

Introduction

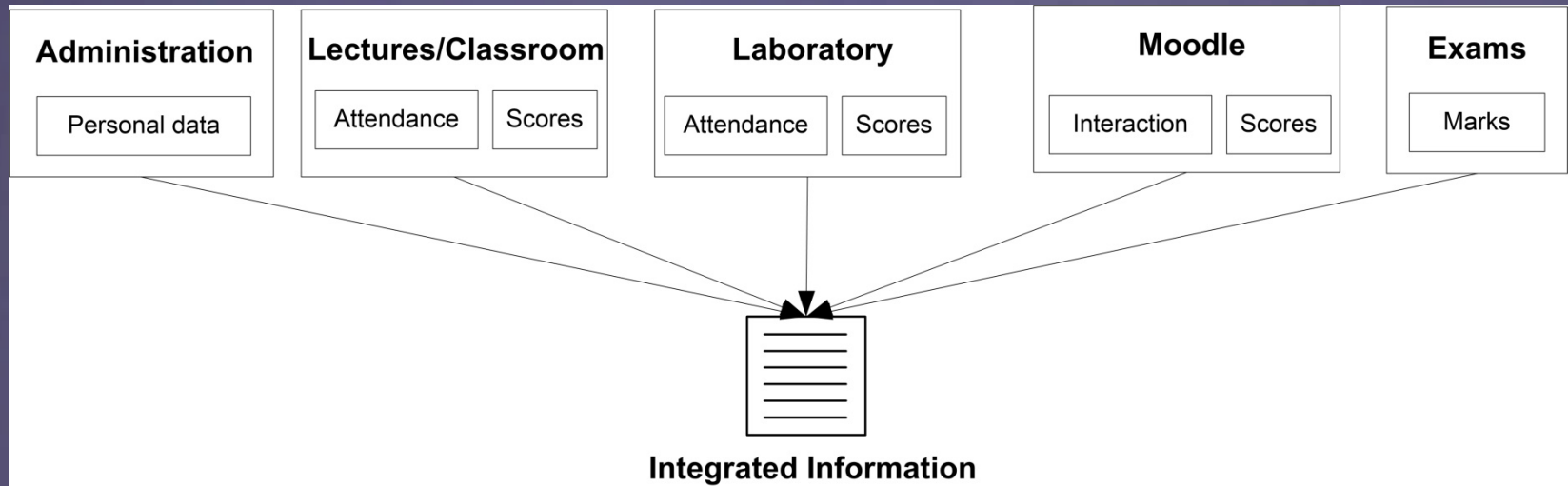
- The main steps/tasks of the overall process of preprocessing educational data are:



Preprocessing EduData

Data Gathering/Aggregation/Integration

- Example of gathering, data aggregation and integration:



EDM Data

Data Cleaning

Missing data is a common issue in education (usually appear when students have not completed or done all the activities in the course) and some possible solutions are:

- Students who have missing values can be removed.
- Whenever possible, these specific students may be contacted and asked (by the instructor) to complete the course.
- To codify missing/unspecified values by mapping incomplete values using for example the labels “?” (missing) and “null” (unspecified).
- To use a global constant to fill in the missing value or to use a substitute value, like the attribute mean or the mode.

EDM Data

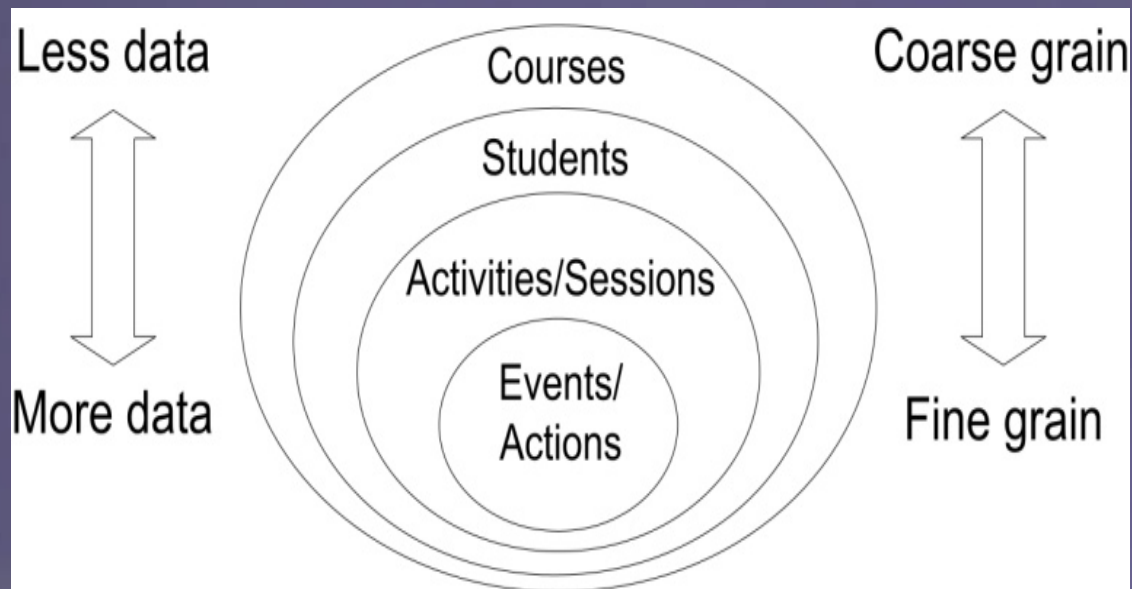
User and Session Identification

- Although **user and session identification** is not specific to education, it is especially relevant due to the longitudinal nature of student usage data.
- Computer-based educational systems provide user authentication (identification by login and password). So it is not necessary to do the typical user and session identification.
- It is also necessary to preserve student data anonymity/privacy but enabling that different pieces of information are linked to the same person. A common solution for it consists in using a number randomly or incrementally generated, like a user ID.

EDM Data

Data Filtering

- Example of **filtering** at different levels of granularity and their relationship to the amount of data:



EDM Data

Attribute Selection

- Example of **Summary Table** with a set of **attributes selected** per student in Moodle courses:

Name	Description
id_student	Identification number of the student.
id_course	Identification number of the course.
num_sessions	Number of sessions.
num_assignment	Number of assignments done.
num_quiz	Number of quizzes taken.
a_scr_quiz	Average score on quizzes
num_posts	Number of messages sent to the forum.
num_read	Number of messages read on the forum.
t_time	Total time used on Moodle.
t_assignment	Total time used on assignments.
t_quiz	Total time used on quizzes.
t_forum	Total time used on forum.
f_scr_course	Final score of the student obtained in the course.

EDM Data

Data Transformation

- Example of **transformation** is Discretization:
 - **Manual discretization** has the user himself directly specifying the cut-off points. Example (Marks/Scores depend on the country):

FAIL: if value is < 5

PASS: if value is ≥ 5 and < 7

GOOD: if value is ≥ 7 and < 9

EXCELLENT: if value is ≥ 9

EDM Data

Data transformation

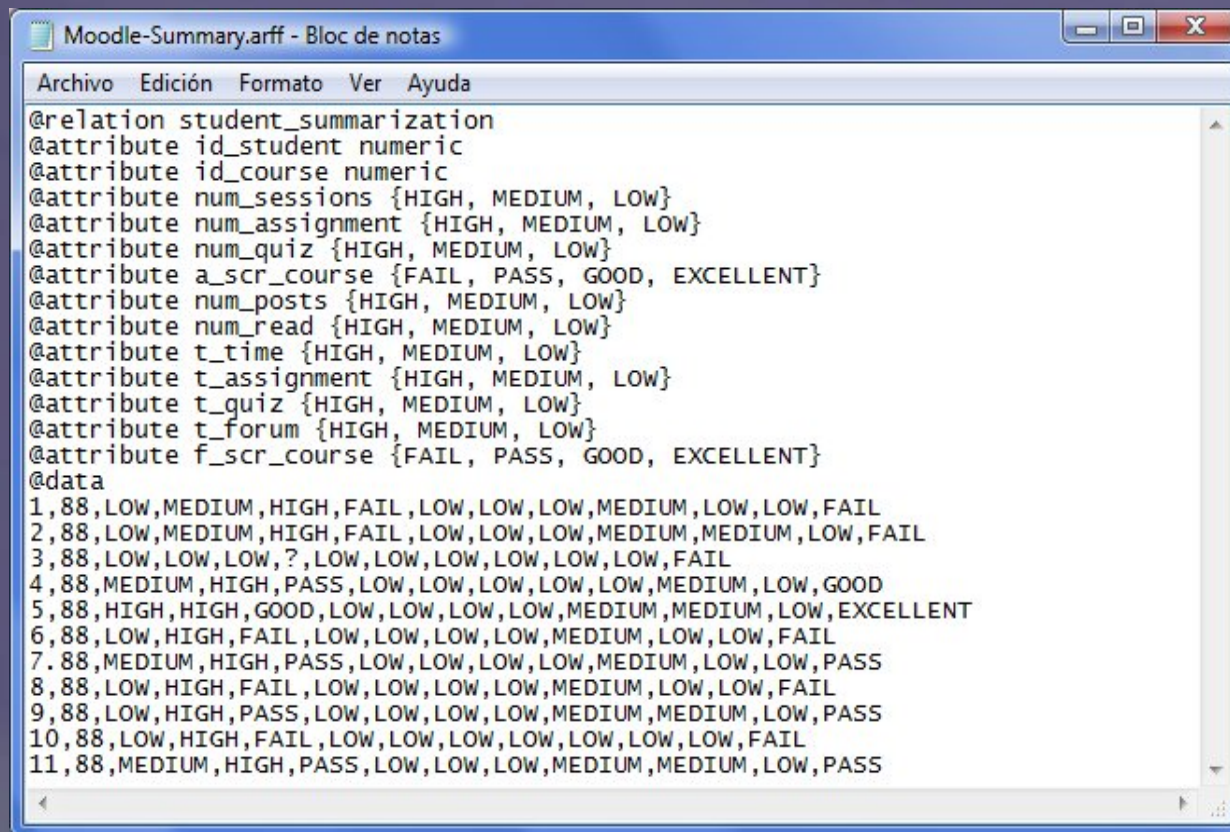
- Example of derived attributes, which enables to create new attributes starting from the current ones:

Name	Description
UserId	A unique identifier per user.
Performance	Percentage of correctly answered tests calculated as the number of correct tests divided by the total number of tests performed).
TimeReading	Time spent on pages (calculated as the total time spent on each page accessed) in a session.
NoPages	The number of accessed pages.
TimeTests	The time spent performing tests (calculated as the total time spent on each test).
Motivation	Engaged / Disengaged.

EDM Data

Data transformation

- Example of Moodle Summary ARFF file:



```
Moodle-Summary.arff - Bloc de notas
Archivo  Edición  Formato  Ver  Ayuda
@relation student_summarization
@attribute id_student numeric
@attribute id_course numeric
@attribute num_sessions {HIGH, MEDIUM, LOW}
@attribute num_assignment {HIGH, MEDIUM, LOW}
@attribute num_quiz {HIGH, MEDIUM, LOW}
@attribute a_scr_course {FAIL, PASS, GOOD, EXCELLENT}
@attribute num_posts {HIGH, MEDIUM, LOW}
@attribute num_read {HIGH, MEDIUM, LOW}
@attribute t_time {HIGH, MEDIUM, LOW}
@attribute t_assignment {HIGH, MEDIUM, LOW}
@attribute t_quiz {HIGH, MEDIUM, LOW}
@attribute t_forum {HIGH, MEDIUM, LOW}
@attribute f_scr_course {FAIL, PASS, GOOD, EXCELLENT}
@data
1,88,LOW,MEDIUM,HIGH,FAIL,LOW,LOW,LOW,MEDIUM,LOW,LOW,FAIL
2,88,LOW,MEDIUM,HIGH,FAIL,LOW,LOW,LOW,MEDIUM,MEDIUM,LOW,FAIL
3,88,LOW,LOW,LOW,?,LOW,LOW,LOW,LOW,LOW,LOW,FAIL
4,88,MEDIUM,HIGH,PASS,LOW,LOW,LOW,LOW,LOW,MEDIUM,LOW,GOOD
5,88,HIGH,HIGH,GOOD,LOW,LOW,LOW,LOW,MEDIUM,MEDIUM,LOW,EXCELLENT
6,88,LOW,HIGH,FAIL,LOW,LOW,LOW,LOW,MEDIUM,LOW,LOW,FAIL
7,88,MEDIUM,HIGH,PASS,LOW,LOW,LOW,LOW,MEDIUM,LOW,LOW,PASS
8,88,LOW,HIGH,FAIL,LOW,LOW,LOW,LOW,MEDIUM,LOW,LOW,FAIL
9,88,LOW,HIGH,PASS,LOW,LOW,LOW,LOW,MEDIUM,MEDIUM,LOW,PASS
10,88,LOW,HIGH,FAIL,LOW,LOW,LOW,LOW,LOW,LOW,LOW,FAIL
11,88,MEDIUM,HIGH,PASS,LOW,LOW,LOW,MEDIUM,MEDIUM,LOW,PASS
```

Thanks.

Questions?