Introduction- to An'alsis Intelligent Data (IDA)

Maria-Amparo Vila

vila@decsai.ugr.es

Investigaci'on group Databases and Intelligent Systems Informaci'on https://idbis.ugr.es/

Department of Sciences of the computation and artificial intelligence

University of Granada

Introduction to the topic

Presentaci' structure on

one. Introduction b'asicas ideas and current motivaci'on

two. KDD and An'alisis Intelligent data.

3. KDD process and the CRISP-DM process

Four. Data concept Miner'ıa

4.1 Miner'ıa data and Estad'ıstica

4.2 Miner'ıa data and learning

5. M'as important problems Miner'ıa Data.

5.1 EDA.Generalizaci'on and summary. DM and DW

5.2 descriptive models: Grouping

5.3 Descriptive models: Modelizaci'on dependency

5.4 predicitivos models: Classi fi caci'on

5.5 Predictive models: Predicting and An'alisis sequences.

6. Data types Miner'ıa

IDA, KDD and DM

Introduction- IDA: b'asicas ideas, motivaci'on hist' Orica

one. Since the beginning of the civilizaci'on man has compiled num'ericos data: Babylonians, Egyptians, Chinese, Greeks Romans, hac'ıan census counted crops and collected taxes etc.

two. Tambi'en since the beginning of the civilizaci'on man has tried to describe the world around him through "patterns" are understood as regularities: astronomia etc. Until the eighteenth century these were verbal descriptions.

IDA, KDD and DM

Introduction- IDA: b'asicas ideas, motivaci'on hist'   Orica

3 From the eighteenth century appears the "m'etodo scienti fi c" for study of natural and social fen'omenos:

IDA, KDD and DM

Introduction- IDA: b'asicas ideas, motivaci'on hist' Orica

3 From the eighteenth century appears the "m'etodo scienti fi c" for study of natural and social fen'omenos:

- The descripci'on of natural or social fen'omenos are hip'otesis that should be supported by experimentaci'on or Remark

IDA, KDD and DM

Introduction- IDA: b'asicas ideas, motivaci'on hist' Orica

3 From the eighteenth century appears the "m'etodo scienti fi c" for study of natural and social fen'omenos:

- The descripci'on of natural or social fen'omenos are hip'otesis that should be supported by experimentaci'on or Remark
- Data must be analyzed and studied by num'ericas t'ecnicas and / or estad'ısticas regularities and relationships between them.

IDA, KDD and DM

Introduction- IDA: b'asicas ideas, motivaci'on hist' Orica

3 From the eighteenth century appears the "m'etodo scienti fi c" for study of natural and social fen'omenos:

- The descripci'on of natural or social fen'omenos are hip'otesis that should be supported by experimentaci'on or Remark
- Data must be analyzed and studied by num'ericas t'ecnicas and / or estad'ısticas regularities and relationships between them.
- The lack of "m'aquinas to calculate and store data" drive the development of disciplines such as Teor'ıa of samples, Estad'ıstica Matem'atica, the An'alsis Num'erico throughout the nineteenth century and part of XX

IDA, KDD and DM

Introduction- IDA: b'asicas ideas, motivaci'on hist'   Orica

4 From the use of computers in the t'ecnicas of an'alsis

Data can process data in bulk and the possibilities of descripci'on of fen'omemos is ampl'ıan:

IDA, KDD and DM

Introduction- IDA: b'asicas ideas, motivaci'on hist' Orica

4 From the use of computers in the t'ecnicas of an'alsis

Data can process data in bulk and the possibilities of descripci'on of fen'omemos is ampl'ıan:

- The type and properties of the variables used is less rigorous
- EDA (exploratory data An'alisis), num'erica Taxonom'ıa, M'etodos of predicci'on not probabil'ısticos etc.: new t'ecnicas of an'alisis appear

IDA, KDD and DM

4 From the use of computers in the t'ecnicas of an'alsis

Data can process data in bulk and the possibilities of descripci'on of fen'omemos is ampl'ıan:

- The type and properties of the variables used is less rigorous
- EDA (exploratory data An'alisis), num'erica Taxonom'ıa, M'etodos of predicci'on not probabil'ısticos etc.: new t'ecnicas of an'alisis appear

5 Because of these t'ecnicas developed as m'etodos

learning within the IA were called to Intelligent Data An'alisis

IDA, KDD and DM

Introduction- to KDD: b'asicas ideas, and motivaci'on

- Data and databases have grown vertiginously
- A consultative approach to data bases cl'asico does not actually provide solutions for managers. Informaci'on is necessary to summarize and present it in an intelligible form.

IDA, KDD and DM

Introduction- to KDD: b'asicas ideas, and motivaci'on

- Data and databases have grown vertiginously
- A consultative approach to data bases cl'asico does not actually provide solutions for managers. Informaci'on is necessary to summarize and present it in an intelligible form.

- From 80 to a system it calls for:
  - Provide properties no data The Explicit.
  - Allow to know relationships between data.
  - Provide summary and / or classifying each informaci'on.
  - All these facilities must be integrated into a user-friendly interface and interactive.

IDA, KDD and DM

Introduction- to KDD: b'asicas ideas, and hist' motivaci'on Orica

Systems cl'asicos data bases offer little functionality to perform applications such as we have discussed.

IDA, KDD and DM

Introduction- to KDD: b'asicas ideas, and hist' motivaci'on Orica

Systems cl'asicos data bases offer little functionality to perform applications such as we have discussed.

A new form of approximation to databases

Miner'ıa data (Data Mining, DM)

Extracci'on knowledge (Knowledge Discovery, KDD)

IDA, KDD and DM

Introduction- to KDD: b'asicas ideas, and hist' motivaci'on Orica

Systems cl'asicos data bases offer little functionality to perform applications such as we have discussed.

A new form of approximation to databases

Miner'ıa data (Data Mining, DM)

Extracci'on knowledge (Knowledge Discovery, KDD)

De fi nition

Miner'ıa Data (DM or KDD) *is a nontrivial process caci'on identify patterns in v'alidos, novel, potentially data '*

useful and understandable (Frawley et al. 1991)

The t'ermimo patr'on must be taken in a broad sense (relationships, trends, groupings, classifications etc ..)

IDA, KDD and DM

Introduction- to KDD: b'asicas ideas, and hist' motivaci'on Orica

De fi nition

Extracci'on knowledge

The process of using a database for any query that is required; including:

- *Preprocessing, sampling and transformations,*
- *Application of t'ecnicas of data miner'ıa for employers*
- *The results evaluaci'on said miner'ıa to identify patterns are considered qu'e knowledge (Fayyad et al. 1996)*

IDA, KDD and DM

De fi nition

Extracci'on knowledge

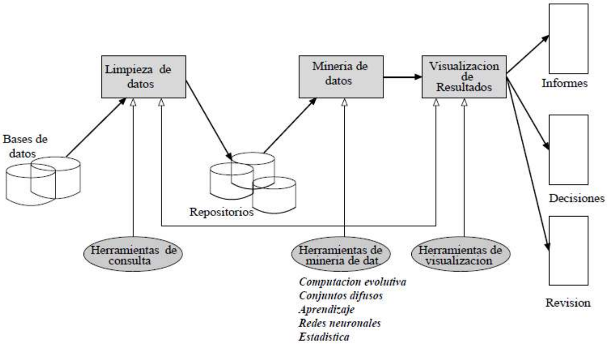
The process of using a database for any query that is required; including:

- *Preprocessing, sampling and transformations,*
- *Application of t'ecnicas of data miner'ıa for employers*
- *The results evaluaci'on said miner'ıa to identify patterns are considered qu'e knowledge (Fayyad et al. 1996)*

IDA and KDD are considered sin'onimos. KDD has a business M'as connotaci'on and scienti fi c IDA M'as DM is a step of both processes

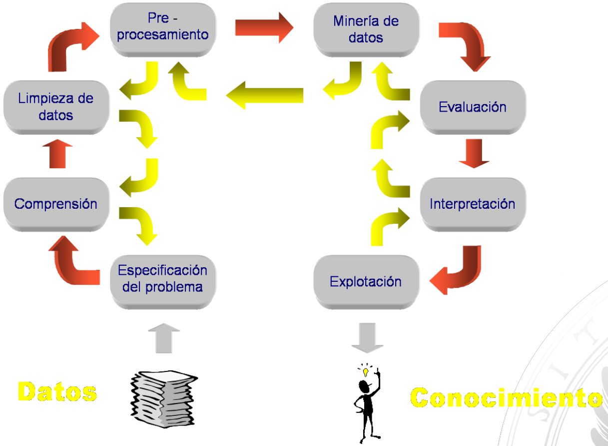
IDA, KDD and DM

Introduction- to KDD: stages in a process of KDD



IDA, KDD and DM

CRISP DM process (Cross Industry Standard Process for Data Mining)



IDA, KDD and DM

Stages: DM CRISP process

Understanding of the project

- What is exactly the problem?. What benefits are expected with soluci'on?
- ¿Qu'e type of soluci'on are looking for? ¿Qu'e answers ask?
- What we know about the project domain?
- What is the risk / cost of not solving it?

IDA, KDD and DM

Stages: DM CRISP process

Understanding of the project

- What is exactly the problem?. What benefits are expected with soluci'on?
- ¿Qu'e type of soluci'on are looking for? ¿Qu'e answers ask?
- What we know about the project domain?
- What is the risk / cost of not solving it?

The Numbers

- What data do we have?
- Are they relevant to the problem? .¿Son fi ables, v'alidos?
- Are the data in your Centes fi t'erminos of: quality, quantity and timing?

IDA, KDD and DM

Stages: DM CRISP process

Preparaci'on data

- ¿Qu'e data we focus on?
- How I can improve their quality?
- Do they need to be processed (preprocessed)?

IDA, KDD and DM

Stages: DM CRISP process

Preparaci'on data

- ¿Qu'e data we focus on?
- How I can improve their quality?
- Do they need to be processed (preprocessed)?

Modeling

- What type (s) model (s) / problem (s) correponde my project? (Model selection)
- What is the right M'as t'ecnica to build the model? (Construcci'on model)
- Is it correct model from the point of view t'ecnico ?. (Validaci'on model)

IDA, KDD and DM

Stages: DM CRISP process

Evaluation

- Does it meet the model requirements of our project?
- Have we learned about our Qu'e problem trav'es model?

IDA, KDD and DM

Stages: DM CRISP process

Evaluation

- Does it meet the model requirements of our project?
- Have we learned about our Qu'e problem trav'es model?

Implantation

- How can it be acquired useful knowledge for decision decisions?
- ¿C'omo I know if the model is still v'alido ?.

Miner'ıa Data (Data Mining)

Data concept Miner'ıa

De fi nition

we mean by Miner'ıa Data (DM) t'ecnicas assembly allowing prepare data, to build the model and validating a process or IDA KDD

Miner'ıa Data (Data Mining)

Data concept Miner'ıa

De fi nition

we mean by Miner'ıa Data (DM) t'ecnicas assembly allowing prepare data, to build the model and validating a process or IDA KDD

- The DM t'ecnicas as'ı are varied as the problems they deal with.
- The mayor'ıa of miner'ıa t'ecnicas of data must be scalable
- Some authors believe that the m'etodos of preparaci'on and study pre-model data selecci'on not part of the DM.

Miner'ıa Data (Data Mining)

Miner'ıa levels of data

Miner'ıa Data (Data Mining)

Miner'ıa levels of data

No-directed or pure There are no restrictions on the system, nor indications about what the user expects. Answers : *Tell me something interesting about the data*

It corresponds to An'alisis Exploratory Data.

Miner'ıa Data (Data Mining)

Miner'ıa levels of data

No-directed or pure There are no restrictions on the system, nor indications about what the user expects. Answers : *Tell me something interesting about the data*

It corresponds to An'alisis Exploratory Data.

directed The objective speci fi c M'as something speci fi c but It has an attitude descriptive . Example: *¿C'omo are grouped customers of a bank '?*

It corresponds to the pattern extracci'on

Miner'ıa Data (Data Mining)

Miner'ıa levels of data

No-directed or pure There are no restrictions on the system, nor indications about what the user expects. Answers : *Tell me something interesting about the data*

It corresponds to An'alisis Exploratory Data.

directed The objective speci fi c M'as something speci fi c but It has an attitude descriptive . Example: *¿C'omo are grouped customers of a bank '?*

It corresponds to the pattern extracci'on

Prediction It is about predict the value of a variable. Example: *Can I characterize a customer defaulting on age, type of work etc ..?*

Miner'ıa Data (Data Mining)

Miner'ıa levels of data

No-directed or pure There are no restrictions on the system, nor indications about what the user expects. Answers : *Tell me something interesting about the data*

It corresponds to An'alisis Exploratory Data.

directed The objective speci fi c M'as something speci fi c but It has an attitude descriptive . Example: *¿C'omo are grouped customers of a bank '?*

It corresponds to the pattern extracci'on

Prediction It is about predict the value of a variable. Example: *Can I characterize a customer defaulting on age, type of work etc ..?*

These three levels correspond to the models discussed in DM:

Exploratory, descriptive and predictive

Miner'ıa Data (Data Mining)

Data Miner'ıa and Estad'ıstica

Miner'ıa Data (Data Mining)

Data Miner'ıa and Estad'ıstica

- Initially the Estad'ıstica is desarroll'o to solve tratmiento data without the help of computers. Enriqueci'o subsequently be using computers.

Miner'ıa Data (Data Mining)

Data Miner'ıa and Estad'ıstica

- Initially the Estad'ıstica is desarroll'o to solve tratmiento data without the help of computers. Enriqueci'o subsequently be using computers.

- The mayor'ıa of estad'ısticas t'ecnicas impose very strong hip'otesis data: num'ericas and continuous variables, independence between variables, equal variances etc. *but they can be used within a DM problem if these are met hip'otesis*

Miner'ıa Data (Data Mining)

Data Miner'ıa and Estad'ıstica

- Initially the Estad'ıstica is desarroll'o to solve tratmiento data without the help of computers. Enriqueci'o subsequently be using computers.

- The mayor'ıa of estad'ısticas t'ecnicas impose very strong hip'otesis data: num'ericas and continuous variables, independence between variables, equal variances etc. *but they can be used within a DM problem if these are met hip'otesis*

- Many of the models t'ecnicas of validaci'on come from *Dise~not Experiments* .
- Some scalability issues are resolved by *Sample Teor'ıa*

Miner'ıa Data (Data Mining)

Miner'ıa data or Learning

- It is the independent Data Miner'ıa Learning?
- *The difference in est'a* quality *data.*

Miner'ıa Data (Data Mining)

Miner'ıa data or Learning

- It is the independent Data Miner'ıa Learning?
- *The difference in est'a* quality *data.*
  - *There is a large volume of data* B'space is reduced earch both rows and columns domains
  - *Var'ıan data over time* The knowledge acquired should be able to be updated as the databases change.
  - *The data are incomplete and / or inaccurate* At this point the teor'ıa fuzzy subsets is very applicable in both REPRESENTATION data and the extracted knowledge.

Miner'ıa Data (Data Mining)

Miner'ıa data or Learning

- It is the independent Data Miner'ıa Learning?
- *The difference in est'a* quality *data.*

  - *There is a large volume of data* B'space is reduced earch both rows and columns domains
  - *Var'ıan data over time* The knowledge acquired should be able to be updated as the databases change.
  - *The data are incomplete and / or inaccurate* At this point the teor'ıa fuzzy subsets is very applicable in both REPRESENTATION data and the extracted knowledge.

  - *Data can be "noisy" with no errors sistem'aticos*
  - *There may be lots of "lost data"*
  - *Some data may be redundant or not signi fi cant*

M'as important issues DM

Generalizaci' on and summary

- Data and objects in the databases contain informaci'on very detailed and very primitive levels
- The b'asica idea of generalization (summary) is to provide compact descriptions for subsets of data to a higher conceptual level.

- The summary data can be analyzed visually and exploratory. Suggesting M'as an'alisis sticados sophist.

M'as important issues DM

Generalizaci' on and summary

- Data and objects in the databases contain informaci'on very detailed and very primitive levels
- The b'asica idea of generalization (summary) is to provide compact descriptions for subsets of data to a higher conceptual level.

- The summary data can be analyzed visually and exploratory. Suggesting M'as an'alisis sticados sophist.

The data cube approach "data cube"

- The b'asica idea is to use multidimensional tables with aggregated data

- The structure obtained is called *multidimensional cube data* and it is supposed stored.

M'as important issues DM

Miner'ıa data and "Data Warehousing"

- Data Warehousing is part of the Data Miner'ıa?

M'as important issues DM

Miner'ıa data and "Data Warehousing"

- Data Warehousing is part of the Data Miner'ıa?
  - Although the problems of generalization and summary are considered within the miner'ıa data actually they correspond to the first phase of the KDD process.

M'as important issues DM

Miner'ıa data and "Data Warehousing"

- Data Warehousing is part of the Data Miner'ıa?
  - Although the problems of generalization and summary are considered within the miner'ıa data actually they correspond to the first phase of the KDD process.
  - The structure of "data cube" is t'ıpica of REPRESENTATION of informaci'on in DW but fu'e presented in "KDD & Data Mining"

M'as important issues DM

Miner'ıa data and "Data Warehousing"

- Data Warehousing is part of the Data Miner'ıa?
  - Although the problems of generalization and summary are considered within the miner'ıa data actually they correspond to the first phase of the KDD process.
  - The structure of "data cube" is t'ıpica of REPRESENTATION of informaci'on in DW but fu'e presented in "KDD & Data Mining"
  - The tecnolog'ıa DW is a commercial approach and focuses on implementaci'on, desempe~ and no interactive visualizaci'on.

M'as important issues DM

Miner'ıa data and "Data Warehousing"

- Data Warehousing is part of the Data Miner'ıa?
- Although the problems of generalization and summary are considered within the miner'ıa data actually they correspond to the first phase of the KDD process.
- The structure of "data cube" is t'ıpica of REPRESENTATION of informaci'on in DW but fu'e presented in "KDD & Data Mining"
- The tecnolog'ıa DW is a commercial approach and focuses on implementaci'on, desempe~ and no interactive visualizaci'on.
- The Miner'ıa Data is M'as acad'emica and develops processes M'as sticados sophist.

M'as important issues DM

Miner'ıa data and "Data Warehousing"

- Data Warehousing is part of the Data Miner'ıa?
- Although the problems of generalization and summary are considered within the miner'ıa data actually they correspond to the first phase of the KDD process.
- The structure of "data cube" is t'ıpica of REPRESENTATION of informaci'on in DW but fu'e presented in "KDD & Data Mining"
- The tecnolog'ıa DW is a commercial approach and focuses on implementaci'on, desempe~ and no interactive visualizaci'on.
- The Miner'ıa Data is M'as acad'emica and develops processes M'as sticados sophist.
- Actually two stages of a process that can be fed back.

M'as important issues DM

Miner'ıa data and "Data Warehousing"

Some "commercial" issues

- Large companies database tools offered DW.
- Additionally DM tools are offered in many cases not own.
- The best tools are not DM own homes databases
- There is little support for the user

M'as important issues DM

Input data

M'as data structure common to work with DM is the

Dataset

items variables	V_{one}	V_{two}	$\cdots\cdots$	V_n
i_{one}	d_{eleven}	d_{12}	$\cdots\cdots$	$d_{\text{one}\ n}$
$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
i_m	$d_{m\,\text{one}}$	$d_{m\,\text{two}}$	$\cdots\cdots$	d_{mn}

M'as important issues DM

Input data

M'as data structure common to work with DM is the

Dataset

items variables V_{one}		V_{two}	......	V_n
i_{one}	d_{eleven}	d_{12}	......	$d_{one\ n}$
⋮	⋮	⋮	......	⋮
⋮	⋮	⋮	......	⋮
i_m	$d_{m\,one}$	$d_{m\,two}$	......	d_{mn}

- items represent cases
- Variables can be of many types
- There may be missing data

M'as important issues DM

descriptive models: Grouping (Clustering)

- It is a process that groups the items of a "dataset" obtaining a set of "clusters" or classes.
- The mayor'ıa of t'ecnicas arise from the Taxonom'ıa Matem'atica and are based on the sameness between items.

M'as important issues DM

descriptive models: Grouping (Clustering)

- It is a process that groups the items of a "dataset" obtaining a set of "clusters" or classes.
- The mayor'ıa of t'ecnicas arise from the Taxonom'ıa Matem'atica and are based on the sameness between items.
- Seldom use additional knowledge about how the groups.
- There are many different t'ecnicas adapted to the types of data.
- Advanced M'as t'ecnicas solve scalability issues.

M'as important issues DM

Modelizaci'on dependency

- *Objective:* Describe dependencies signi fi cant among the variables included in the database

M'as important issues DM

Modelizaci'on dependency

- *Objective:* Describe dependencies signi fi cant among the variables included in the database models dependencies can be:
  - Qualitative or quantitative (functional units and an'alisis of regresi'on)
  - partial or complete units

M'as important issues DM

Modelizaci'on dependency

- *Objective:* Describe dependencies signi fi cant among the variables included in the database models dependencies can be:

  - Qualitative or quantitative (functional units and an'alisis of regresi'on)

  - partial or complete units

  When it comes to quantitative variables, and expected the existence of a relaci'on $y = f(x_{\text{one}}, .. x_n)$ *we have a* predictive model *usually an'alisis of regresi'on*

M'as important issues DM

Modelizaci'on dependency

- *Objective:* Describe dependencies signi fi cant among the variables included in the database models dependencies can be:

  - Qualitative or quantitative (functional units and an'alisis of regresi'on)

  - partial or complete units

  When it comes to quantitative variables, and expected the existence of a relaci'on $y = f(x_{\text{one}}, .. x_n)$ we have a predictive model *usually an'alisis of regresi'on*

  When you have no prior knowledge, the variables are M'as general values and seek partnerships we have a
  descriptive model

M'as important issues DM

descriptive problems: rules asociaci' on

Discover significant associations between sets of attribute values

- A cl'asico example:

  Find connections between different types of products in a sales database. For example whether customers who buy milk buy bread

M'as important issues DM

descriptive models: Classi fi caci' on

- It has a set of data where one of the variables represents the class to which the item pertenence.
- We seek a *classifying procedure caci'on* that does not allow to include each new item in a class.
- The classi fi t'ecnicas caci'on are adapted from existing learning with a special'enfasis in *scalability issues*

M'as important issues DM

descriptive models: Classi fi caci' on

- It has a set of data where one of the variables represents the class to which the item pertenence.
- We seek a *classifying procedure caci'on* that does not allow to include each new item in a class.
- The classi fi t'ecnicas caci'on are adapted from existing learning with a special'enfasis in *scalability issues*
- Have adapted'exito ID-3, t'ecnicas estad'ısticas (An'alisis Discriminant, Bayesian), t'ecnicas yt'ecnicas based neural networks based on "rough sets" based on fuzzy yt'ecnicas l'ogica .

The classi fi caci'on is one of the most studied problems in DM

M'as important issues DM

An'alisis time series and sequences

- This type of an'alisis applies to time-dependent data and for which wants to find a temporary patr'on.
- Usually these problems have been addressed by the *Temporal series* but these statistician t'ecnicas of origin impose many limitations to the data.

- sequences are usually not stationary, no peri'odicas, irregular and even ca'oticas.

M'as important issues DM

An'alisis time series and sequences

- This type of an'alisis applies to time-dependent data and for which wants to find a temporary patr'on.
- Usually these problems have been addressed by the *Temporal series* but these statistician t'ecnicas of origin impose many limitations to the data.
- sequences are usually not stationary, no peri'odicas, irregular and even ca'oticas.
- T'ecnicas applied between processing is est'an ~ *n* ales adaptive algorithms gen'eticos, and teor'ıa of chaos and nonlinear din'amicos systems.

M'as important issues DM

An'alisis time series and sequences

- This type of an'alisis applies to time-dependent data and for which wants to find a temporary patr'on.
- Usually these problems have been addressed by the *Temporal series* but these statistician t'ecnicas of origin impose many limitations to the data.
- sequences are usually not stationary, no peri'odicas, irregular and even ca'oticas.
- T'ecnicas applied between processing is est'an ~ *n* ales adaptive algorithms gen'eticos, and teor'ıa of chaos and nonlinear din'amicos systems.

• It est'a devoting much effort to an'alisis pattern of discrete sequences, time dependent or not (Stream Mining)

Miner'ıa Data (Data Mining)

Data types Miner'ıa

Seg'an EL Description: FIELD where applicable appear different types of DM

Miner'ıa Data (Data Mining)

Data types Miner'ıa

Seg'an EL Description: FIELD where applicable appear different types of DM

Text Mining When word knowledge is extracted.

- Estructuraci'on no problems. intermediately
- Problems of semantics.
- T'ıpico a problem. Detecci'on identity

Miner'ıa Data (Data Mining)

Data types Miner'ıa

Seg'an EL Description: FIELD where applicable appear different types of DM

Text Mining When word knowledge is extracted.

- Estructuraci'on no problems. intermediately
- Problems of semantics.
- T'ıpico a problem. Detecci'on identity

web Mining When knowledge of the Web is extracted

- Web content mining. B'asicamente Text Mining
- Miner'ıa Web use.
- CONNECT Web Mineria (Mining Graph)
- Miner'ıa on social networks.

Miner'ıa Data (Data Mining)

Data types Miner'ıa

Seg'an EL Description: FIELD where applicable appear different types of DM

Text Mining When word knowledge is extracted.

- Estructuraci'on no problems. intermediately
- Problems of semantics.
- T'ıpico a problem. Detecci'on identity

web Mining When knowledge of the Web is extracted

- Web content mining. B'asicamente Text Mining
- Miner'ıa Web use.
- CONNECT Web Mineria (Mining Graph)
- Miner'ıa on social networks.

Bioinform'atica B'asicamente Stream Mining and Clustering