Advanced Topics in Data Mining

Instructor:
Dr. Hamid Turab Mirza

Department of Computer Science
CUI, Lahore

KNOWLEDGE DISCOVERY PROCESS

Outline

- Introduction
- What is the Knowledge Discovery Process?
  - Overview
- Knowledge Discovery Process Models
  - Academic
  - Industrial
  - Hybrid
  - Comparison of the models
- Research Issues
  - Metadata and Knowledge Discovery Process

Introduction

Before attempting to extract useful knowledge from data, it is important to focus on the process that leads to finding new knowledge:

- define a sequence of steps (with feedback loops) that should be followed to discover new knowledge (e.g. patterns)
- each step of the process is usually realized with the help of available commercial or open-source software tools

Introduction

Why do we need standardized knowledge discovery (KD) process (KDP) model?

- **KDP model is a logical, cohesive, well-thought-out structure and approach to help understand the need, value, and mechanics behind a KDP**
- **to ensure the end product is useful for the user/owner of the data**
- **KD projects require a significant project management effort that needs to be grounded in a solid framework**
- **KD follows other disciplines that have established models**
- **there is a widely recognized need for a standardization to stimulate growth of the data mining (DM) industry**

Introduction

KDP is defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data:

- **consists of many steps (one is DM), each attempting at the completion of a particular task**
- **KDP includes how the data is stored and accessed, how to use efficient and scalable algorithms, how to interpret and visualize the results, and how to model and support interaction between human and machine**
- **concerns support for learning and analyzing the application domain**

Overview of the KDP

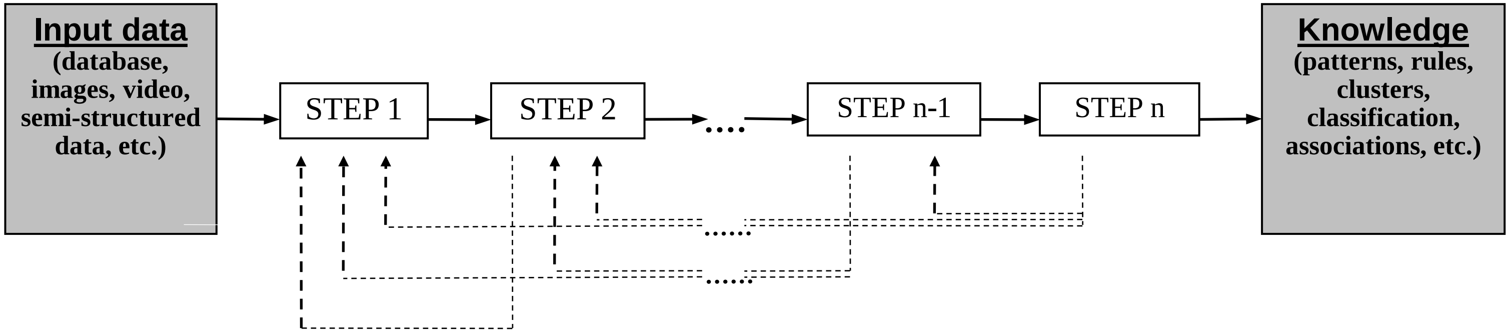
The KDP model

- its steps are executed in a sequence

- the next step is initiated upon successful completion of the previous step - the result generated by the previous step are its input

- it stretches between the task of understanding the project domain and data, through data preparation and analysis, to evaluation and application of the generated results

- it is iterative, i.e. includes feedback loops that are triggered by revisions


Input data
(database, images, video, semi-structured data, etc.)
STEP 1
STEP 2
....
STEP n-1
STEP n
Knowledge
(patterns, rules, clusters, classification, associations, etc.)
.....
.....

Overview of the KDP

KDP consists of a set of processing steps that are to be followed by practitioners when executing a Knowledge Discovery project

- model describes procedures that are performed at each step
- it is primarily used to plan, work through, and reduce the cost of any given project

Overview of the KDP

Since 1990s several different KDP models were developed

- the main differences are in the number and scope of specific steps

- a common feature of all models is definitions of inputs and outputs
  - inputs include data in various formats, such as numerical, nominal stored in databases or flat files, images, video, semi-structured data like XML or HTML, etc.
  - the output is the generated new knowledge –

    in terms of rules, patterns, classification models, associations, statistical analysis, etc.

Knowledge Discovery Process Models

Popular KDP models include:

- **nine-step model by Fayyad et al.**
  - **academic**
- **CRISP-DM (CRoss-Industry Standard Process for Data Mining) model**
  - **industrial**
- **six-step KDP model by Cios et al.**
  - **hybrid (academic/industrial)**

Knowledge Discovery Process Models

Nine-step model by Fayyad et al.

1. **Developing and Understanding of the Application Domain**
   It includes learning the relevant prior knowledge and the goals specified by the end-user.
2. **Creating a Target Data Set**
   It selects a subset of attributes and data points (examples), which will be used to perform discovery tasks. It includes querying the existing data to select a desired subset.
3. **Data Cleaning and Preprocessing**
   It consists of removing outliers, dealing with noise and missing values, and accounting for time sequence information.
4. **Data Reduction and Projection**
   It consists of finding useful attributes by applying dimension reduction and transformation methods, and finding invariant representation of the data.

Knowledge Discovery Process Models

Nine-step model by Fayyad et al.

5. Choosing the Data Mining Task
   It matches the goals defined in step 1 with a particular DM method, such as classification, regression, clustering, etc.

6. Choosing the Data Mining Algorithm
   It selects methods for searching patterns in the data, and decides which models and parameters may be appropriate.

7. Data Mining
   It generates patterns in a particular representational form, such as classification rules, decision trees, regression models, trends, etc.

Knowledge Discovery Process Models

Nine-step model by Fayyad et al.

8. **Interpreting Mined Patterns**

   It usually involves visualization of the extracted patterns and models, and visualization of the data.

9. **Consolidating Discovered Knowledge**

   It consists of incorporating the discovered knowledge into the performance system, and documenting and reporting it to the end user. It may include checking and resolving potential conflicts with previously believed knowledge.

Knowledge Discovery Process Models

Nine-step model by Fayyad et al.

- the process is iterative
  - a number of loops between any two steps but the authors provide no specific details
- the model details technical description with respect to data analysis but lacks description of business aspects
- major applications
  - a commercial Knowledge Discovery system called MineSet™ (see Purple Insight Ltd. at http://www.purpleinsight.com).
  - was used to facilitate projects in a number of domains including engineering, medicine, production, e-business, and software development

Knowledge Discovery Process Models

CRISP-DM (CRoss-Industry Standard Process for Data Mining) model

- **designed by Integral Solutions Ltd. (provider of commercial Data Mining solutions), NCR (database provider), Daimler Chrysler (automobile manufacturer), and OHRA (insurance company); the latter two provided data and case studies**

- **Secial Interest Group was created to support the developed process model (over 300 users and tool/service providers)**

- **the model consists of six steps**

Knowledge Discovery Process Models

CRISP-DM model

1. **Business Understanding**
   Focus is on understanding objectives and requirements from a business perspective. It converts them into a DM problem definition, and designs a preliminary project plan to achieve the objectives. It is broken into several sub-steps:
   - determination of business objectives
   - assessment of situation
   - determination of DM goals, and
   - generation of project plan.
2. **Data Understanding**
   Starts with an initial data collection and familiarization with the data. Includes identification of data quality problems, discovery of initial insights into the data, and detection of interesting data subsets. It is broken down into:
   - collection of initial data
   - description of data
   - exploration of data, and
   - verification of data quality

Knowledge Discovery Process Models

CRISP-DM model

3. Data Preparation

   Covers all activities to construct the final dataset, which constitutes the data to be fed into DM tool(s) in the next step. It includes table, record, and attribute selection, data cleaning, construction of new attributes, and data transformation. This step is divided into:

   - selection of data
   - cleansing of data
   - construction of data
   - integration of data, and
   - formatting of data sub-steps.

Knowledge Discovery Process Models

CRISP-DM model

4. Modeling

   Selects and applies various modeling tools. It involves using several methods for the same DM problem and calibration of their parameters to optimal values. Since some methods require a specific format for input data, often reiteration into the previous step is necessary. This step is subdivided into:

   - selection of modeling technique(s)
   - generation of test design
   - creation of models, and
   - assessment of generated models.

Knowledge Discovery Process Models

CRISP-DM model

5. Evaluation

   After building one or more high quality (from a data analysis perspective) models, they are valuated from business objective perspective and review of the steps executed to construct the models is performed. A key objective is to determine if there are important business issues that have not been considered. At the end, a decision on the use of the DM results is reached. The key sub-steps include:

   – evaluation of the results

   – process review

   – determination of the next step.

Knowledge Discovery Process Models

CRISP-DM model

6. Deployment

   Involves organization and presentation of the discovered knowledge in a user-friendly way. Depending on the requirements, this can be as simple as generating a report or as complex as implementing a repeatable KDP.
   This step is subdivided into:

   – planning of the deployment
   – planning of the monitoring and maintenance
   – generation of final report, and
   – review of the process sub-steps.

Knowledge Discovery Process Models

CRISP-DM model

- **uses easy to understand vocabulary and is well documented**
- **acknowledges the iterative nature of the process with loops between the steps**
- **extensively used model, mainly because of its grounding in industrial real-world experience**
- **major applications**
  - **medicine, engineering, marketing, sales**
  - **turned into a commercial KD system called Clementine® (see SPSS Inc. at http://www.spss.com/clementine)**

Knowledge Discovery Process Models

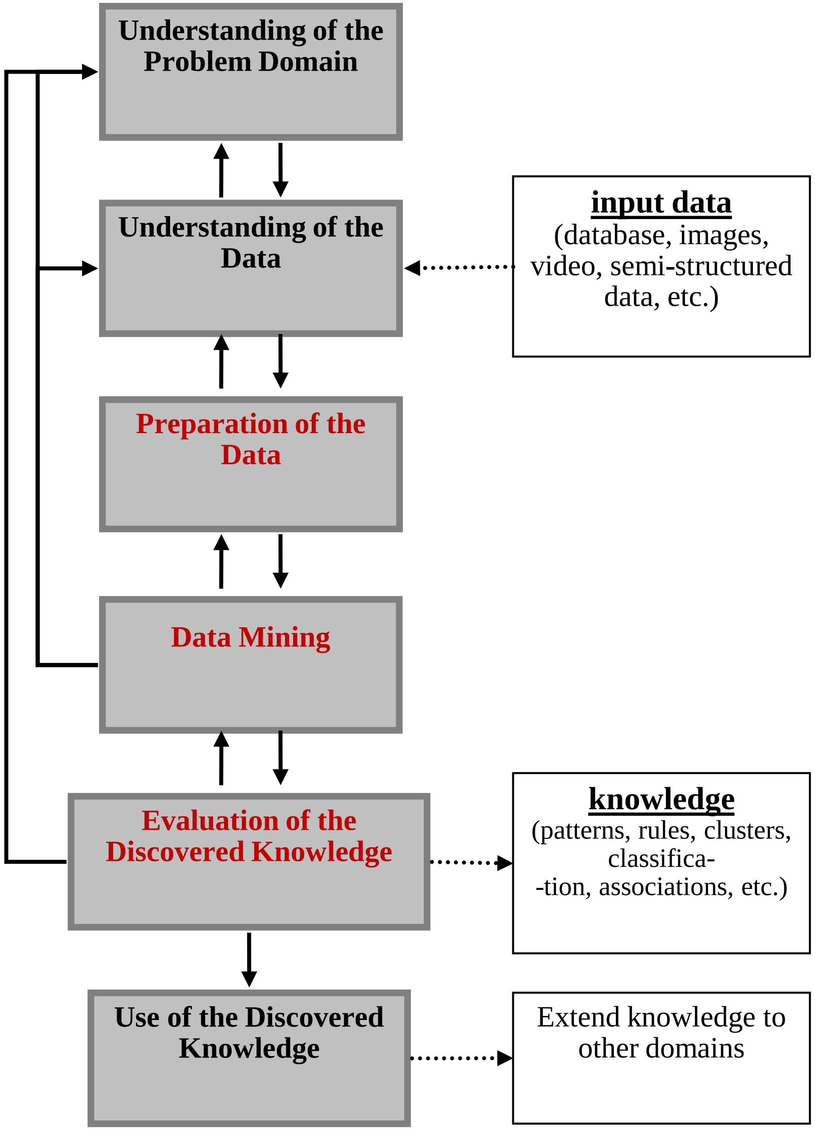
Six-step model by Cios et al.

- inspired by the CRISP-DM model and adopted for academic research; main differences and extensions include:
  - providing more general, research-oriented description of the steps
  - has a Data Mining step instead of the Modeling step
  - introducing several new explicit feedback mechanisms. The CRISP-DM model has only 3 major feedbacks, while this model has 6 detailed feedback mechanisms
  - modification of the last step; the knowledge discovered for a particular domain may be used in other domains
- has six steps

Knowledge Discovery Process Models

Cios et al.

six-step model


Understanding of the Problem Domain
Understanding of the Data
input data
(database, images, video, semi-structured data, etc.)
Preparation of the Data
Data Mining
Evaluation of the Discovered Knowledge
knowledge
(patterns, rules, clusters, classifica-
-tion, associations, etc.)
Use of the Discovered Knowledge
Extend knowledge to other domains

Knowledge Discovery Process Models

Six-step model by Cios et al.

1. **Understanding the Problem Domain**

   Involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It involves learning domain-specific terminology. A description of the problem and its restrictions is prepared. Project goals are translated into the DM goals and initial selection of DM tools to be used is performed.

2. **Understanding of the Data**

   Includes collection of sample data and deciding which data, including its format and size, will be needed. Background knowledge is used to guide these efforts. Data is checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Includes verification of the usefulness of the data in respect to the DM goals.

Knowledge Discovery Process Models

Six-step model by Cios et al.

3. Preparation of the Data
   Concerns deciding what data will be used as input to DM tools in the next step. Involves sampling, running correlation and significance tests, data cleaning and checking completeness of data records, removing or correcting for noise and for missing values, etc. The cleaned data is further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say by discretization), and by summarization of data (data granularization). The results are data meeting specific input requirements of DM tools.
4. Data Mining
   It involves using various DM methods to derive new knowledge/information from the preprocessed data.

Knowledge Discovery Process Models

Six-step model by Cios et al.

5. **Evaluation of the Discovered Knowledge**
   Includes understanding results, checking whether the discovered knowledge is novel and interesting, interpreting results by domain experts, and checking possible impact of the discovered knowledge. Only the approved models are retained and the entire process is revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared.

6. **Use of the Discovered Knowledge**
   It consists of planning where and how the discovered knowledge will be used. The application in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project is documented. Finally, the discovered knowledge is deployed.

Knowledge Discovery Process Models

Six-step model by Cios et al. identifies and describes explicit feedback loops

- from the *Understanding of the Data* to the *Understanding of the Problem* step; the loop is caused by need of additional domain knowledge to better understand the data

- from the *Preparation of the Data* to *Understanding of the Data* step; the loop is caused by the need for additional/more specific information about the data to guide the choice of data preprocessing algorithms

- from the *Data Mining* to the *Understanding of the Problem Domain* step; the reason could be unsatisfactory results generated by used DM methods, which may requires modification of the DM goals

- from the *Data Mining* to the *Understanding of the Data* step; the most common reason is poor understanding of the data, which results in incorrect selection of DM method and thus its subsequent failure

Knowledge Discovery Process Models

Six-step model by Cios et al. identifies and describes explicit feedback loops

- **from the *Data Mining* to the *Preparation of the Data* step; the loop is caused by need to improve data preparation. This is often caused by the specific requirements of the used DM method, which may have not been known during the Data Preparation step**

- **from the *Evaluation of the Discovered Knowledge* to the *Understanding of the Problem Domain* step; the most common cause is invalidity of the discovered knowledge. Reasons include incorrect understanding/interpretation of the domain, incorrect design/understanding of problem restrictions, requirements or goals**

- **from the *Evaluation of the Discovered Knowledge* to the *Data Mining*; this loop is executed when the discovered knowledge is not novel, interesting, or useful. The least expensive solution is to choose a different DM tool and repeat the DM step.**

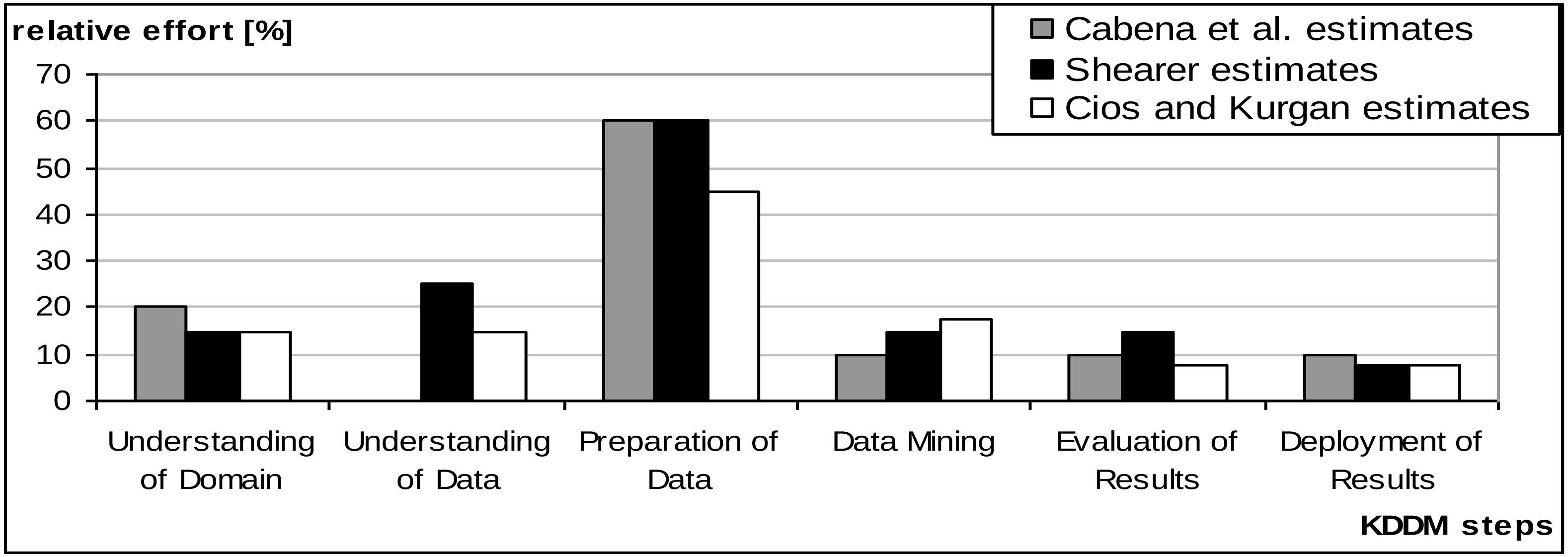
Comparison of the KDP Models

Model	Fayyad et al.	Cios et al.	CRISP-DM
domain of origin	academic	hybrid (academic/industry)	industry
# steps	9	6	6
Steps	1. Developing and Understanding of the Application Domain	1. Understanding of the Problem Domain	1. Business Understanding
	2. Creating a Target Data Set	2. Understanding of the Data	2. Data Understanding
	3. Data Cleaning and Preprocessing	3. Preparation of the Data	3. Data Preparation
	4. Data Reduction and Projection		
	5. Choosing the Data Mining Task		
	6. Choosing the Data Mining Algorithm		
	7. Data Mining	4. Data Mining	4. Modeling
	8. Interpreting Mined Patterns	5. Evaluation of the Discovered Knowledge	5. Evaluation
	9. Consolidating Discovered Knowledge	6. Use of the Discovered Knowledge	6. Deployment
Notes	provides detailed technical description with respect to data analysis, but lacks business aspects	draws from both academic and industrial models; emphasizes iterative aspects; identifies and describes explicit feedback loops	uses easy to understand well-defined vocabulary; has good documentation
supporting software	commercial system MineSet™	N/A	commercial system Clementine®
reported application domains	medicine, engineering, production, e-business, software	medicine, software	medicine, engineering, marketing, sales

Comparison of the KDP Models

An important aspect of the KDP is the relative time spent to complete each of its steps

- it enables precise scheduling
- estimates by researchers and practitioners are shown below
  - specific estimated values depend on existing knowledge about the domain, skill level of the humans, complexity of the problem, etc.
  - data preparation step is by far the most time consuming step



Research Issues

The goal of the KDP model is to achieve integration of the entire KD process through the use of industrial standards

Another important issue is to provide interoperability and compatibility between different software systems and platforms used in the process

- integrated and interoperable models enable semi-automation of Knowledge Discovery systems

Research Issues

Metadata and Knowledge Discovery Process

- the goal is to perform a KDP without possessing extensive background knowledge and without manual procedures to exchange data/knowledge between different DM tools

    - technology used is the XML (eXtensible Markup Language)
        - it allows to describe and store structured or semi-structured data, and exchange data in a platform- and tool-independent way
        - XML standardizes communication between diverse database systems, build standard data repositories for sharing data between systems using different software platforms, and provides a framework for integration of the KDP

Research Issues

Metadata and Knowledge Discovery Process

- metadata standards based on the XML may provide a complete solution
  - PMML (Predictive Model Markup Language) is an XML-based standard that allows for interoperability among different DM tools and for achieving integration with database systems, spreadsheets, and decision support systems
    - It describes data models and shares them between compliant applications
    - XML and PMML can be stored in most database management systems
    - XML was designed by the Data Mining Group (http://www.dmg.org/), a vendor-led group that develops DM standards

Research Issues

Metadata and Knowledge Discovery Process

– **PMML snippet**

polynomial regression model for *iris* data generated by the *DB2 Intelligent Miner for Data* V8.1

```
<?xml version="1.0" encoding="windows-1252" ?>
<PMML version="2.0">
<DataDictionary numberOfFields="4">
   <DataField name="PETALLEN" optype="continuous" x-significance="0.89" />
   <DataField name="PETALWID" optype="continuous" x-significance="0.39" />
   <DataField name="SEPALWID" optype="continuous" x-significance="0.92" />
   <DataField name="SPECIES" optype="categorical" x-significance="0.94" />
   <DataField name="SEPALLEN" optype="continuous" />
</DataDictionary>
<RegressionModel modelName="…" functionName="regression"
algorithmName="polynomialRegression" modelType="stepwisePolynomialRegression"
targetFieldName="SEPALLEN">
<MiningSchema>
    <MiningField name="PETALLEN" usageType="active" />
    <MiningField name="PETALWID" usageType="active" />
                              …
</MiningSchema>
<RegressionTable intercept="-45534.5912666858">
   <NumericPredictor name="PETALLEN" exponent="1" coefficient="8.87" mean="37.58" />
   <NumericPredictor name="PETALLEN" exponent="2" coefficient="-0.42" mean="1722" />
                              …
</RegressionTable>
</RegressionModel>
<Extension>
   <X-modelQuality x-rSquared="0.8878700000000001" />
                              …
</Extension>
</PMML>
```

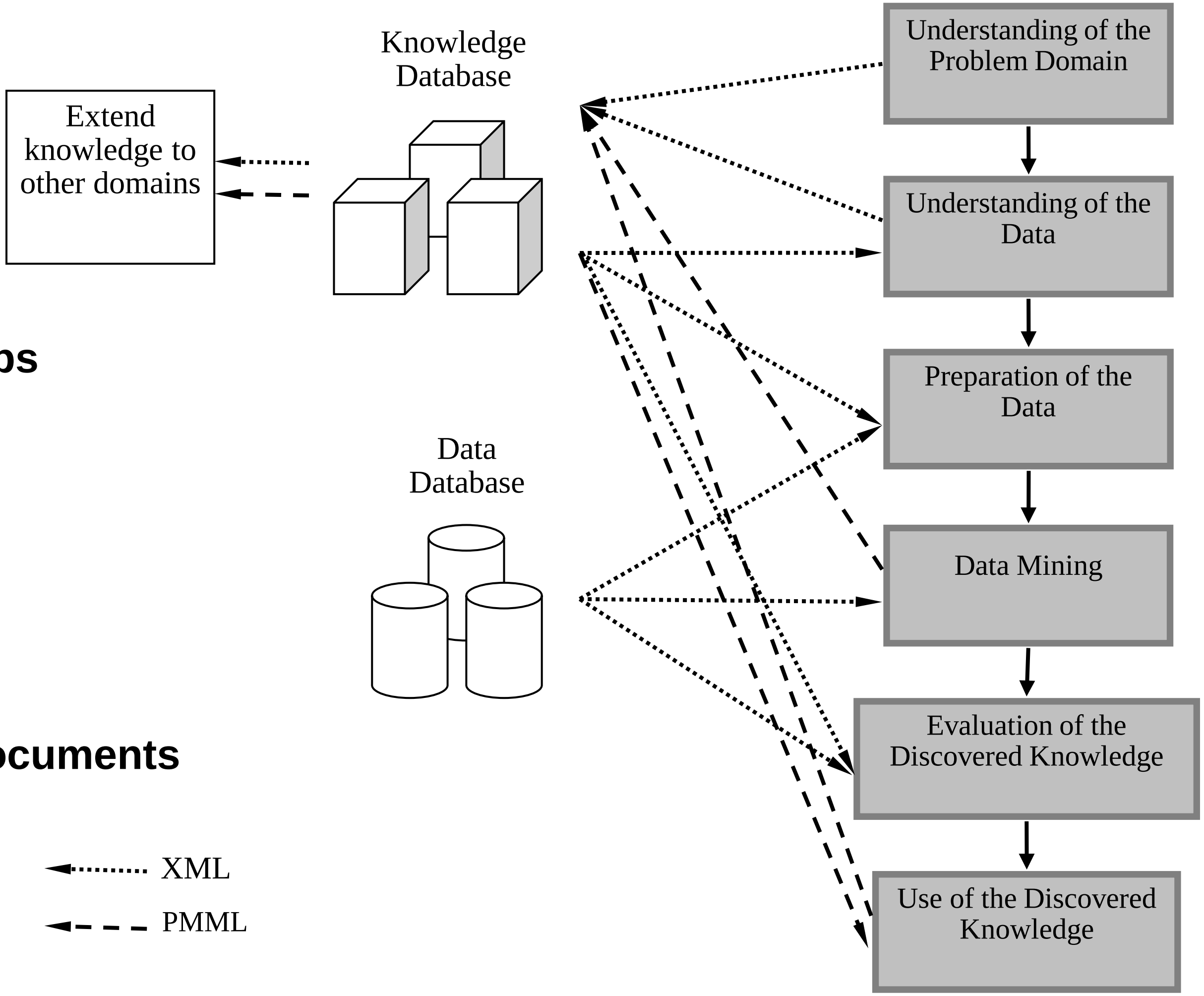
Research Issues

XML, PMML, and the KDP

- information collected during the domain and data understanding steps can be stored as XML
  - and later used in data preparation and knowledge evaluation steps

- knowledge extracted in the DM step is verified in the evaluation step, and domain knowledge gathered in the domain understanding

  step is stored using PMML documents


Knowledge Database
Extend knowledge to other domains
Data Database
Understanding of the Problem Domain
Understanding of the Data
Preparation of the Data
Data Mining
Evaluation of the Discovered Knowledge
Use of the Discovered Knowledge
XML
PMML

© 2007 Cios / Pedrycz / Swiniarski / Kurgan

References

Cios, K. and Kurgan, L. 2005. Trends in Data Mining and Knowledge Discovery, In: Pal, N., Jain, L., and Teoderesku N. (Eds.), *Knowledge Discovery in Advanced Information Systems*, Springer

Fayyad, U., Piatesky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds.) 1996. *Advances in Knowledge Discovery and Data Mining*, AAAI Press

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. 1996. The KDD Process for Extracting Useful Knowledge from Volumes of Data, *Communications of the ACM*, 39(11):27-34

Kurgan, L. and Musilek, P. 2006. A Survey of Knowledge Discovery and Data Mining Process Models, *Knowledge Engineering Review*, 21(1):1-24

Shearer, C. 2000. The CRISP-DM Model: The New Blueprint for Data Mining, *Journal of Data Warehousing*, 5(4):13-19