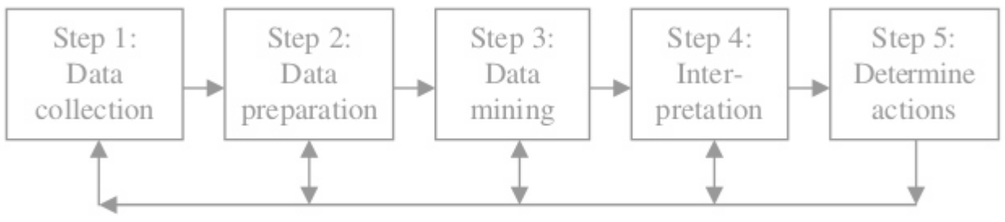


Fig. 1.1 Steps in the KDD process

[3]

STEP 0

KNOWLEDGE DISCOVERY in DATA

STEP 1

KNOWLEDGE DISCOVERY in DATA

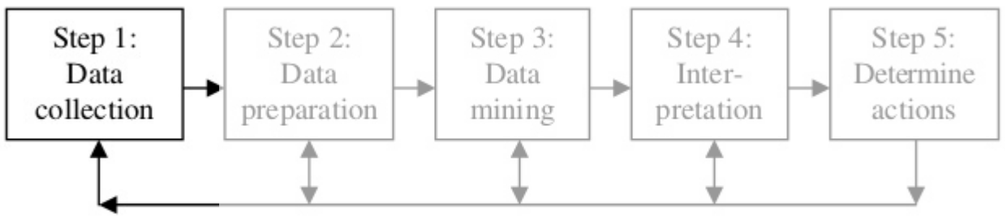


Fig. 1.1 Steps in the KDD process

[3]

PATTERN module:

pattern.web
- url downloads
- interval requests
- search engine requests
- use google translate
- crawl
  - wikipedia articles
  - fb comments + reactions
  - dbpedia
  - twitter
  - rss
- parse HTML elements, PDFs
- retrieve emails via imap
- retrieve local information (eg. tweets)

[1]

PATTERN module:

pattern.db
- built database
- work with time/date

[1]

sources

[1] : http://www.clips.ua.ac.be/pages/pattern

[2] : CLiPS - Guy de Pauw, Pattern workshop - Cqrrelations, January 2015

[3] : Data Mining and Profiling in Large Databases, Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (Eds.) (2013)

STEP 2

KNOWLEDGE DISCOVERY in DATA

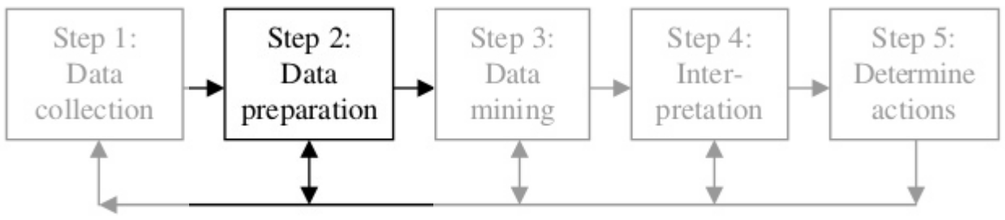


Fig. 1.1 Steps in the KDD process

[3]

Step 2: – text-preperation

syntax:
tokenization/normalization (98%)*
simplest thing/important thing
identifying the units in your text
to read the punctuation, e.g.:
- dr.
- This is a sentence.

lemmatization:
reduce wordforms to their dictionary item
is/been/was/be
--> belongs to 'to be'
+ plurals --> singulars

syntactical:
part-of-speech tagging
important elements for object text-mining
--> nouns
for subjective text-mining
--> adjectives
word sense disambiguation
bank / bank
--> river bank / money bank
semantic role labeling

pragmatics: (?)
named entity recognition
co-reference resolution (50%)*
<-- meaning output

*(% refers to accuracy)

[2]

PATTERN module:

pattern.en |es|de|fr|it|nl
- text preperation
- sentiment analysis tool
- WordNet interface
- wordlists interface

[1]

PATTERN module:

pattern.search
- a pattern matching system similar to regular expressions, that can be used to search a string by *syntax* (word function) or by *semantics* (word meaning).
- eg.:('{NP} be * than {NP}')

[1]

sources

[1] : http://www.clips.ua.ac.be/pages/pattern

[2] : CLiPS - Guy de Pauw, Pattern workshop - Cqrrelations, January 2015

[3] : Data Mining and Profiling in Large Databases, Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (Eds.) (2013)

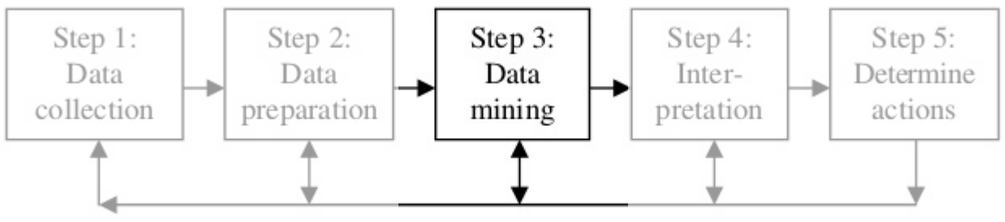


Fig. 1.1 Steps in the KDD process

[3]

STEP 3

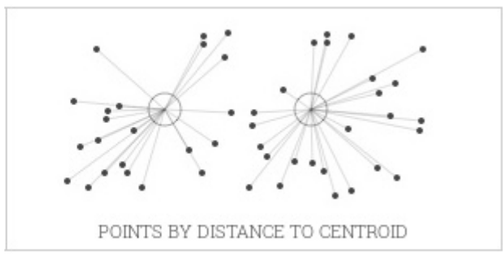
KNOWLEDGE DISCOVERY in DATA

3 a

PATTERN module:

pattern.vector
- machine learning tools:
  - word count functions
  - bag-of-word documents
  - a vector space model
  - latent semantic analysis (context analysis)
- algorithms for *clustering*
  - k-means (similar clusters)
  - hierarchical (nested clusters)
- and *classification*
  - NB (Naive Bayes)
  - KNN (k-nearest neighbor)
  - SLP (Single-layer perceptron)
  - SVM (Support vector machine)
- genetic algorithm

[1]



[1]

3 b

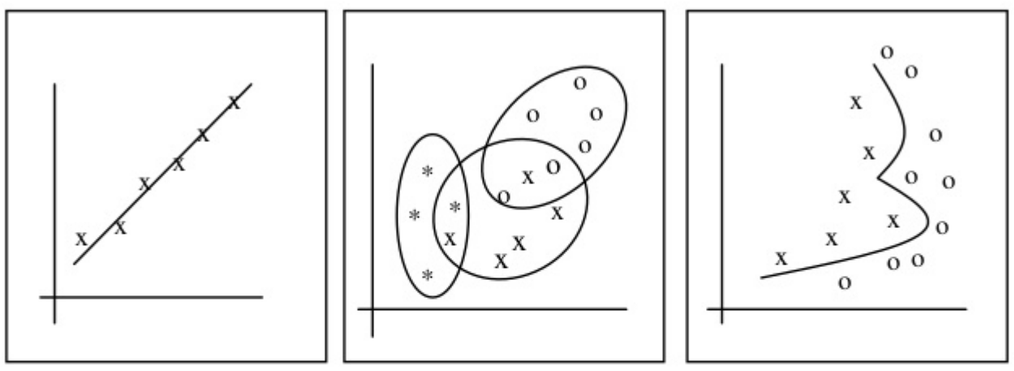
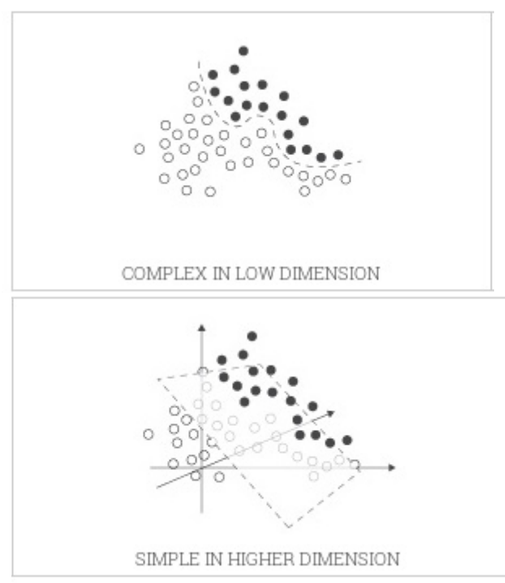


Fig. 2.2 Examples of different types of discovery algorithms: Pattern mining with a linear regression function (A), clustering (B), and classification (C)

[3]



[1]

clustering *unsupervised learning*

group similarities

classification *supervised learning*

map into pre-defined classes

[3]

sources

[1] : http://www.clips.ua.ac.be/pages/pattern

[2] : CLiPS - Guy de Pauw, Pattern workshop - Cqrrelations, January 2015

[3] : Data Mining and Profiling in Large Databases, Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (Eds.) (2013)

STEP 4

KNOWLEDGE DISCOVERY in DATA

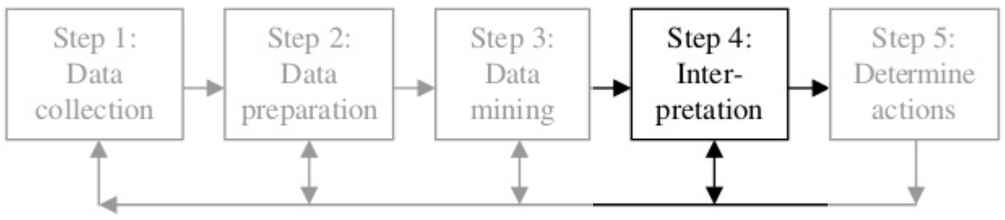


Fig. 1.1 Steps in the KDD process

[3]

Step 4: – interpretation

gold standard
--> annotated test set

10-fold cross validation
taking 1000 tweets
training 800 tweets
test 100 tweets
val 100 tweets

compare to baseline scores
- frequency-baseline
you expect 80% of the tweets to be neutral(?)

- informative baseline
i have a 60% chance that it will rain tomorrow
--> your result need to be higher
otherwise --> why do all the work?

[2]

sources

[1] : http://www.clips.ua.ac.be/pages/pattern

[2] : CLiPS - Guy de Pauw, Pattern workshop - Cqrrelations, January 2015

[3] : Data Mining and Profiling in Large Databases, Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (Eds.) (2013)

STEP 5

KNOWLEDGE DISCOVERY in DATA

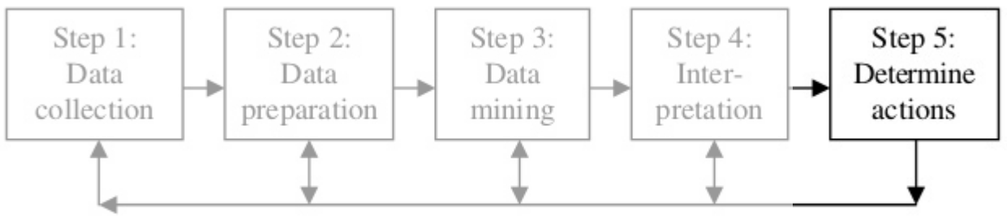


Fig. 1.1 Steps in the KDD process

[3]

Step 5: – Pattern in practise (actions)

> ... as a maker of such tools, it has a site of application that has not been intented.

How do you see this, and where will there be talked about certain issues?

< **three levels**:
- fundamental research
  without any concrete applications

- IWT* research center (applied)
  develop research to improve problems in society (AMiCA)

- industry sponsored efforts

* IWT, Agentschap voor Innovatie door Wetenschap en Technologie

[2]

PATTERN module:

pattern.graph

graph analysis (shortest path, centrality) and graph visualization in the browser. A graph is a network of nodes connected by edges. It can be used for example to study social networks or to model semantic relationships between concepts.

[1]

1.2.2 From Data to Knowledge

The KDD-process may be very helpful in finding pattern and relations in large databases that are not immediately visible to the human eye. Generally, deriving patterns and relations are considered creating added value out of databases, as the patterns and relations provide insight and overview and may be used for decision-making. The plain database may not (or at least not immediately) provide such insight. For that reason, usually a distinction is made between the terms data and knowledge. Data is a set of facts, the raw material in databases usable for data mining, whereas knowledge is a pattern that is interesting and certain enough for a user.[14] It may be obvious that knowledge is therefore a subjective term, as it depends on the user. For instance, a relation between vegetable consumption and health may be interesting to an insurance company, whereas it may not be interesting to an employment agency. Since a pattern in data must fulfill two conditions (*interestingness* and *certainty)* in order to become knowledge, we will discuss these conditions in more detail.

[3]

sources

[1] : http://www.clips.ua.ac.be/pages/pattern

[2] : CLiPS - Guy de Pauw, Pattern workshop - Cqrrelations, January 2015

[3] : Data Mining and Profiling in Large Databases, Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (Eds.) (2013)