Data Mining - Han and Kamber, Elsevier

Chapter-1

Data mining → Extraction of knowledge from large amount of data.

→ Also called as KDD (Knowledge Discovery / mining from database)

Knowledge extraction

pattern analysis.

Data archaeology

Knowledge → Understanding of information by experience, Learning

History

→ Since 1960s, we were basically doing data collection and creation of database for primitive file processing. or day to day transaction.

→ The research and development in database systems ~~started~~ since 1970s. During this period relational database systems (where data are stored in relational table structures, data modeling tools (E-R models), indexing and accessing methods (B-trees, hashing etc.), ~~Query languages → SQL etc~~. were developed.

→ Users had flexible data access through query languages (SQL), user interfaces, ~~forms and reports~~ optimized query processing and transaction management.

→ Efficient methods for on-line transaction processing (OLTP), where a query is viewed

viewed as a read-only transaction, have developed, which helped in efficient storage, retrieval and management of large amount of data.

→ In mid-1980s advanced data models and ~~advanced applications~~ application oriented database systems such as
→(geographical data, medical, satelite images)
spatial, multimedia, active, stream
↳ space
and sensor, scientific and engineering databases, knowledge bases ~~and~~ are came into picture.

→ During late 1980's the technique of data mining and data warehouse came into puse.

Data warehouse

→ A repository of multiple heterogeneous databases.

→ In the datawarehouse there is large amount of data coupled with the need for powerful data analysis tools.

→ This situation is called Data Rich but information poor situation.

→ If we search for data manually, this will consume more time and also prone to
leads
error.

So Data mining tools are required to perform data analysis and uncover the important data patterns.

Difference between database and datawarehouse.

→ A datawarehouse is a repository of information collected from multiple sources over a history of time, stored under a unified schema, and used for data analysis and decision support.

→ A database is a collection of interrelated data that represents the current status of the stored data. A database systems supports ad-hoc query and on-line transaction processing.

→ similarities. → Both are repositories of information, storing huge amounts of data.

Steps involved in data mining when viewed as a process of knowledge discovery :-

The steps are.

Data preprocessing

(1) Data Cleaning → a process that removes noise and inconsistent data.

(2) Data Integration → where multiple data sources may be combined.

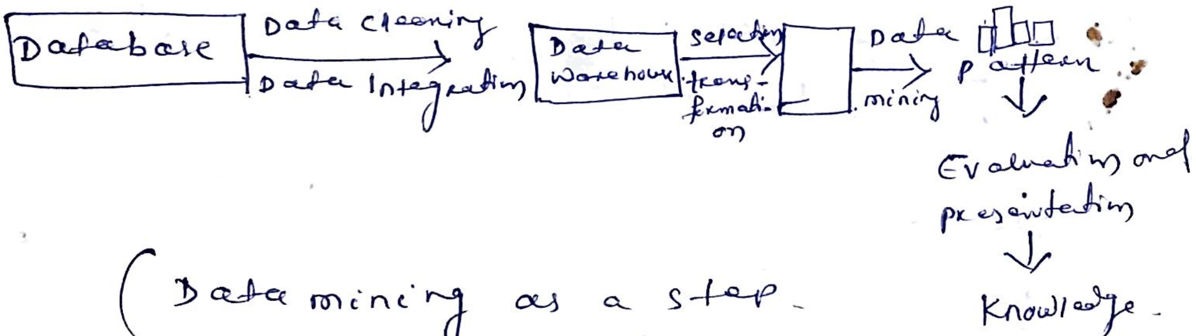
(3) Data selection → where data relevant to the analysis task are retrieved from the database.

(4) Data transformation → where data are transformed into forms appropriate for mining.

(5) Data mining → an essential process where intelligent and efficient methods are applied in order to extract patterns.

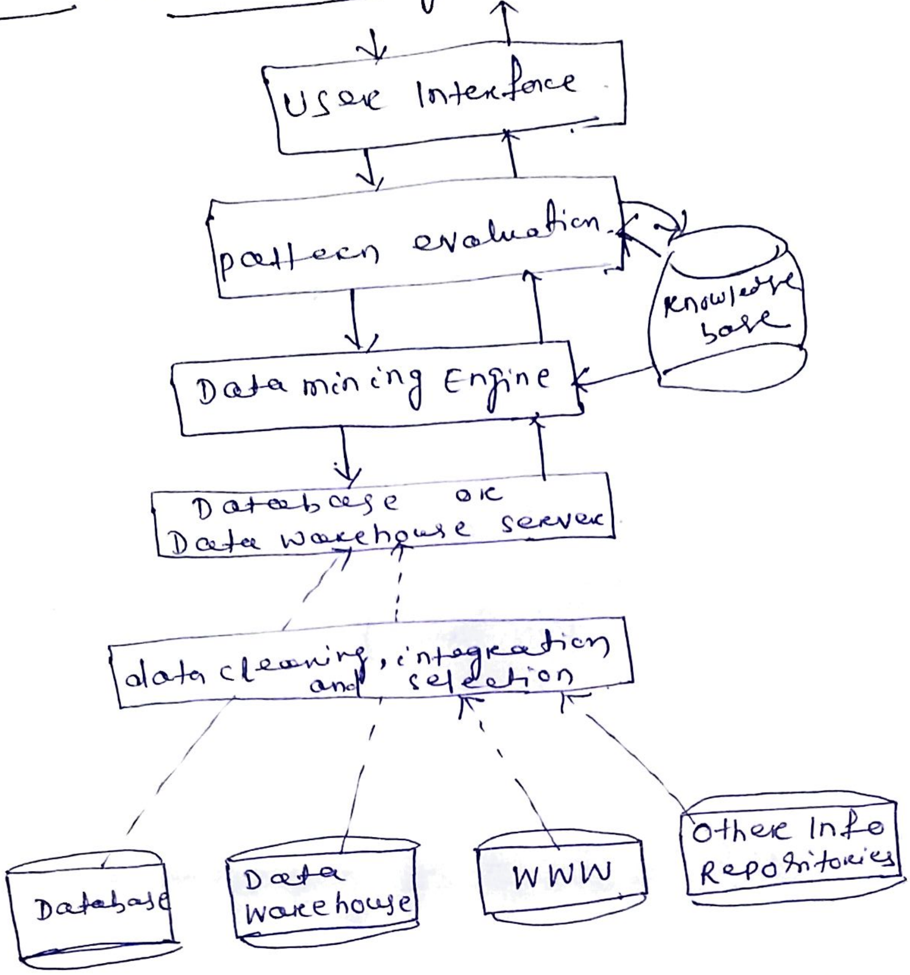
(6) pattern evaluation → a process that identifies the truly interesting patterns representing knowledge base.

(7) Knowledge presentation → where visualization and knowledge representation techniques are used to present the mined knowledge to the user.



(Data mining as a step in the process of knowledge discovery).

Architecture of a typical data mining system or Data mining components.



① Database, Data warehouse, WWW or other information repository.

→ Data cleaning
Data integration } performed here.

② Database/Dataware house server —

→ Required for fetching the relevant data, based on the user's data mining request.

③ Knowledge base —

→ This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns.

(4) Data mining engine —

It consists of a set of functional modules for several tasks such as

(a) characterization → summarization of the general features of a target class of data.

(b) association.

(c) correlation analysis.

(d) classification → is the process of finding a model that distinguishes data classes, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.

(e) prediction.

(f) clustering.

(g) outlier analysis → Analysis of noisy data.

(h) evolution analysis → Describes and models trends for objects whose behavior changes over time. Ex → stock market analysis.

Association

buys(X, "computer") → buys (X, "s/w") [support = 1%, confidence = 50%.]

X → is a variable representing a customer.

confidence = 50% → means that there is a 50% chance that the customer will buy s/w when he/she buys a computer

1% support → means, 1% of all the transactions under analysis showed that computer and s/w were purchased together.

prediction → used to predict missing or unavailable numerical data.

Clustering → The objects are clustered or grouped based on the principle of max. the intraclass similarity and minimizing the interclass similarity.

(5) pattern evaluation module -

→ This component typically employs interestingness measures and interacts with the data mining modules to focus on the search towards interesting patterns.

→ It may use interestingness thresholds to filter out discovered patterns.

(6) User Interface →

It communicates the user with the data mining system, allowing the user to interact with the system by specifying a data mining query.

Database	Data Mining
1) Find all credit applicants with last name of smith	1) Find all credit applicants who are poor credit risks. (classification)
2) Identify customers who have purchased more than Rs. 10,000 last month	2) Identify customers with similar buying habits (clustering).
3) Find all customers who have purchased milk.	3) Find all items which are frequently purchased with milk (association rules)