Comparison Between Data Mining, Web Mining and Text Mining

Outline

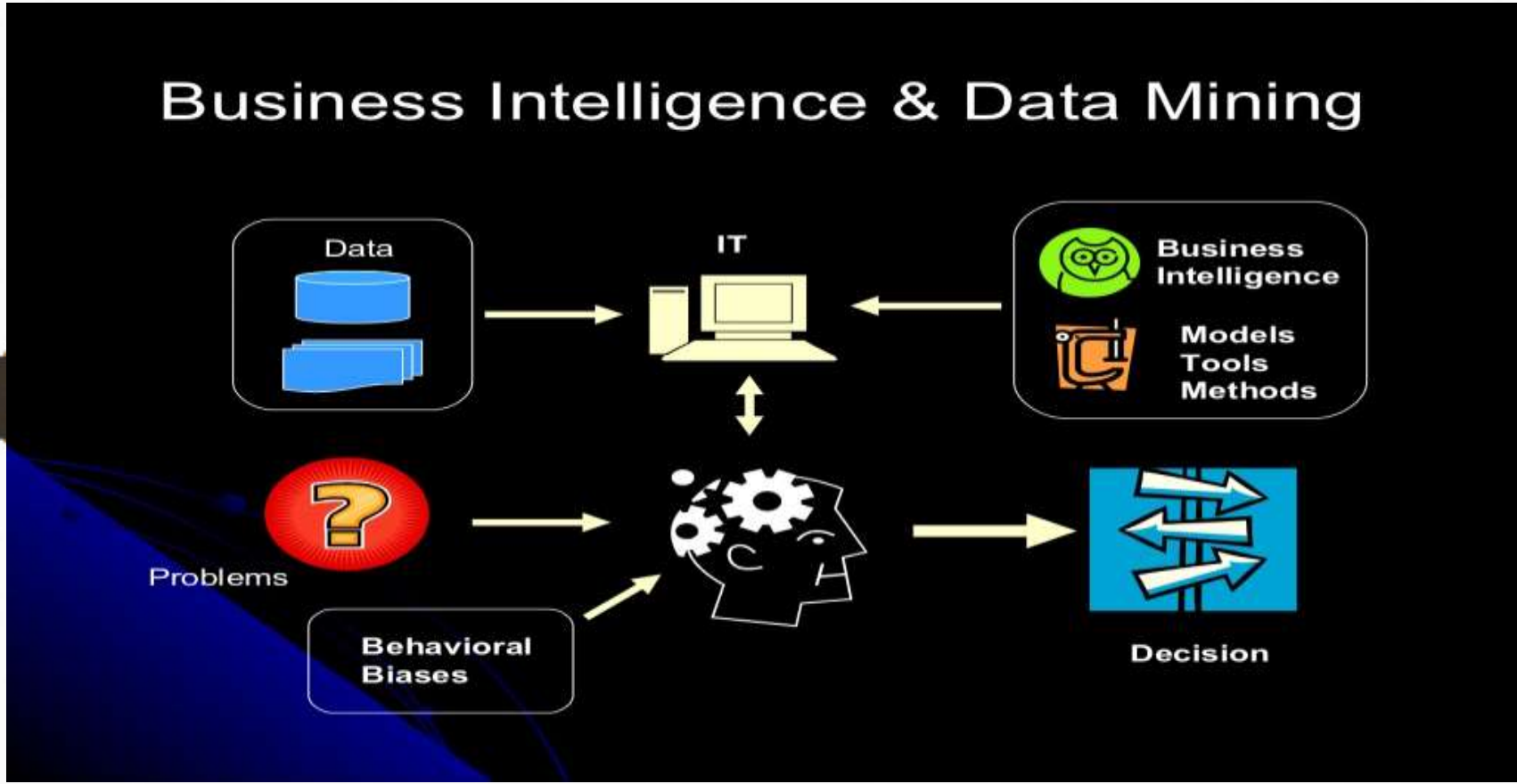
- Data Mining
- Text Mining
- Web Mining
- Differences
- Similarities
- Shared Techniques
- Conclusion

DATA MINING

OVERVIEW

Data Mining: The task of discovering interesting patterns from large amounts of data.

Business Intelligence & Data Mining



Why Data Mining

Data, Data everywhere yet…….

I can't find the data I need

I can't get the data I need

I can't remember the data I found

I can't use the data I found

Data format

- Text file –
  - Dataset is a text file containing a transaction per line.
- Table format –
  - Dataset is stored in a two columns table.
- Custom Format –
  - Many data mining packages use a custom format for the input data and example of this is ARFF(Attribute relation File Format).

Data storage

- Most of our data's used in Data Mining solutions are stored in a Data warehouse

- A Data Warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process.

Potential Application

Database analysis and decision support

– **Market analysis and management**

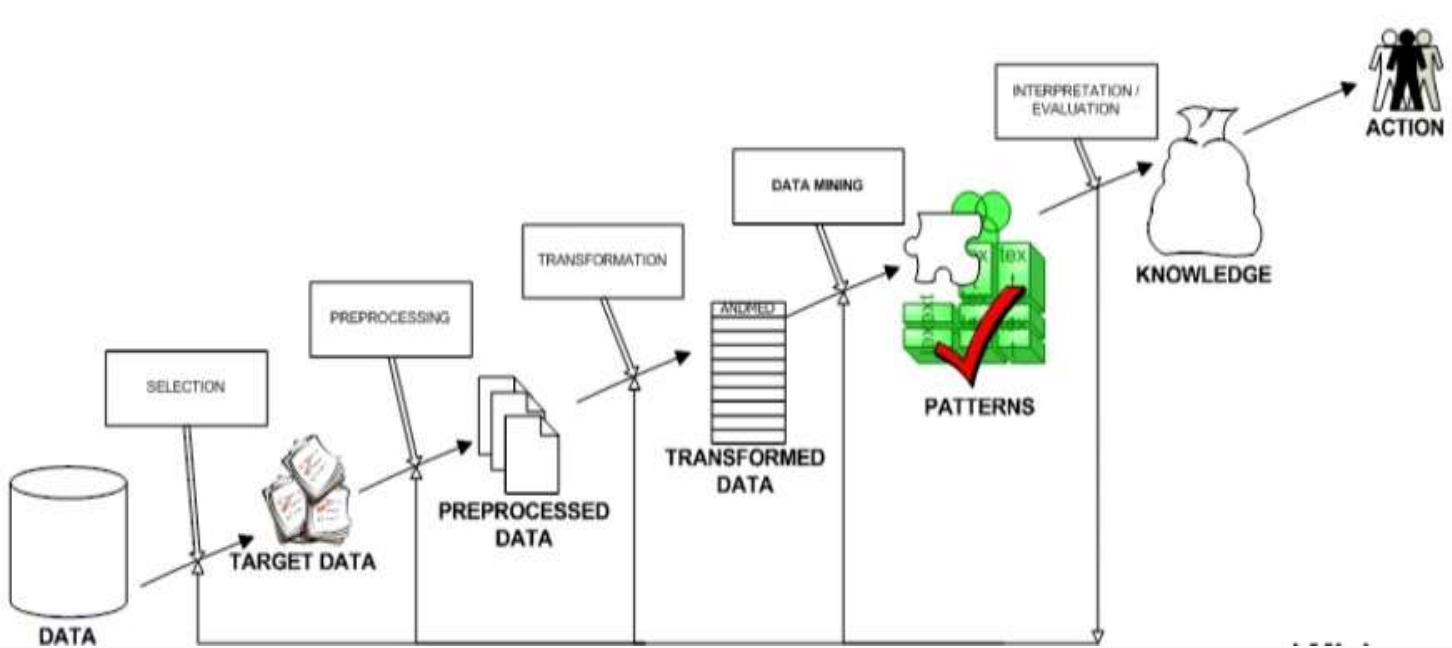
• target marketing, customer relation management, market basket analysis, cross selling, market segmentation

– **Risk analysis and management**

• Forecasting, customer retention, improved underwriting, quality control, competitive analysis

– **Fraud detection and management**

Process of Data Mining



Process of Data Mining

- **Data Integration** –
  - where multiple data source may be combined.
- **Data Selection** –
  - where data relevant to the analysis task are retrieved from the database.
- **Data Cleaning-**
  - to remove noise and inconsistent data

Process of Data Mining

- **Data Transformation**
  - where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance.
- **Data Mining**
  - an essential process where intelligent methods are applied in order to extract data patterns.
- **Pattern Evaluation**
  - (to identify the truly interesting patterns representing knowledge based on some interestingness measures
- **Knowledge Presentation**
  - where visualization and knowledge representation techniques are used to present the mined knowledge to the user

Techniques Involved in Data Mining

- Classification and prediction
  - Finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown and it use SVM, trees, decision tree etc for algorithm
- Cluster Analysis
  - Analyzes data objects without consulting a known class label and uses K mean algorithm.
  - An example is the Market segmentation.
- Outlier Analysis
  - Detect data objects that do not comply with the general behavior or model of the data.
  - It has its usage in fraud detection.

Techniques Involved in Data Mining

- Association Analysis
  - Discovery of interesting associations and correlations within data
  - Uses Apriori and FP- growth algorithm
  - Example is the Market Basket Analysis
- Evolution Analysis
  - Describes and models regularities or trends for objects whose behavior changes over time.
  - Forecasting stock market index using Time series analysis

Data Mining Tools

- There are variety of tools to all techniques or processes related to Data mining and some of which are;
  - Oracle Data Mining
  - Microsoft SQL SERVER
  - DBMiner

Conclusion

TEXT MINING

Overview

- Text miming is the extraction of useful information from a collection of documents.
- Textual data makes up huge amounts of data found on World Wide Web WWW, aside from multimedia.
- The phrase “text mining” is used to denote any system that analyzes large quantities of natural language text by parsing it and then detects lexical or linguistic usage patterns in an attempt to extract correct information (Sebastian, 2002).

Data format

- Textual data can be unstructured i.e. they are in a free-style text with no format just texts all through

- It could be weakly/ semi structured i.e. adheres to some pre-specified format. Example are scientific papers, business reports.

Data storage

- In text mining, data are stored in Document warehouse or document collection.
- Document is a unit of discrete textual data within a collection, representing some real word documents like emails, research paper, newspaper, business reports.
- Document collection is a grouping or a repository for Business Intelligence keeping variety of document types from different sources to automatically extract and store the salient features of documents.
  - It could be static or dynamic (grows over time)

Text Mining Process

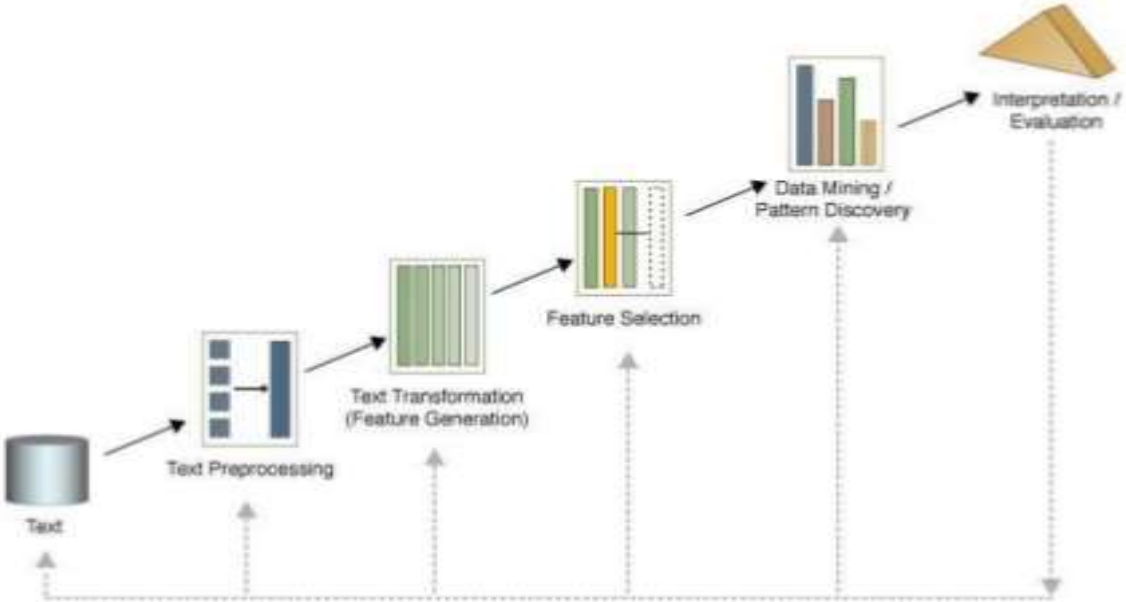
Text preprocessing
- **Syntactic/Semantic text analysis**

Features Generation
- **Bag of words**

Features Selection
- **Simple counting**
- **Statistics**

Text/Data Mining
- **Classification- Supervised learning**
- **Clustering- Unsupervised learning**

Analyzing results
- **Mapping/Visualization**
- **Result interpretation**



Techniques involved in Text mining

- Information Retrieval
  - Users find documents that satisfies their information needs.
  - Utilizes techniques of vector space model by calculating Euclidian distance between vectors representing documents and query.
- Categorization
  - It is a kind of supervised learning that categorizes volumes of documents into “topics” or “themes”
  - Techniques like Naïve Bayesian classifier, Nearest Neighbor classifier, Decision Tree, and Support Vector Machines can be used to categorize text

Techniques involved in Text mining

- Clustering
  - The technique used in order to find groups of documents with similar content.
  - The K-means algorithm is used here
- NLP/ computational Linguistics.
  - Technique of text mining in which computer gets the ability to process, analyze and deduce patterns from language enabling algorithms for problems like parts of speech tagging e.t.c.
  - The Hidden Markov Model algorithm is used here.

Techniques involved in Text mining

- Summarization.
  - Text summarization is to reduce the length and detail of a document while retaining most important points and general meaning
- Visualization
  - To represent individual documents or groups of documents text flags are used to show document category and to show density colors are used. Like scikit-learn, scipy, numpy, pandas, matplotlib.

Text Mining Tools

- There are variety of tools to all techniques or processes related to Text mining and they range from Commercial tools to Open source tools
- Commercial Tools
  - SAS text miner, Poly vista.
- Open Source Tools
  - Rapid Miner text mining, GATE, NLTK, OpenNLP
- Comprehensive list of tools can be found here
  - https://en.wikipedia.org/wiki/List_of_text_mining_software
  - http://www.kdnuggets.com/software/text.html

Conclusion.

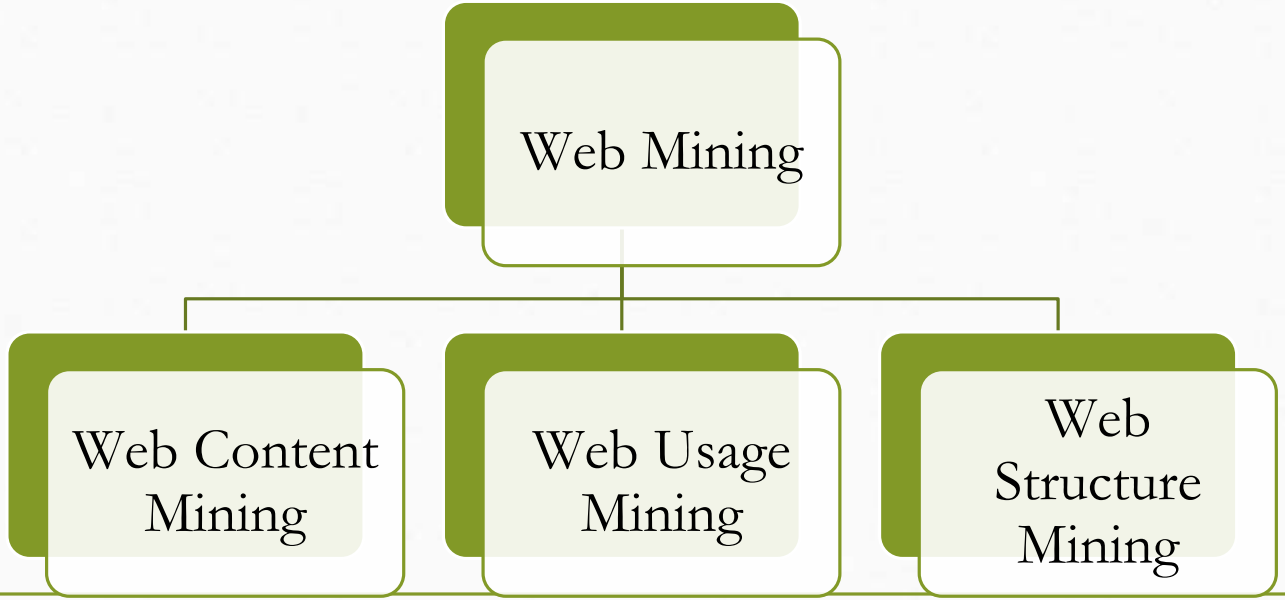
Text Mining is an important aspect of Business Intelligence that helps users and enterprises in analyzing stored text in a better way so as to make better decisions, improve customer satisfaction and gain competitive advantage. It is better than data mining as it provides deeper insight into the expanding business domain and extracts more fruitful data for business intelligence.

WEB MINING

Web Data

- Web pages
- Usage Data
- Intra-page structures
- Inter-page structures.
- Supplemental Data
  - Profiles
  - Registration Information
  - Cookies

Types of Web Mining



Web Content Mining

- Here we have web page content mining and Search result mining.
- It extends the work of basic search engines.
- It helps in discovering useful information from web data
- Text, multimedia, metadata and hyperlinks
- Its techniques involves
- Classification,
- Clustering
- Association.

Web Structure Mining

- This part mine structure like link and graphs of the web.
- Its techniques used are PageRank and Clever.
- It could be combined with content mining to retrieve important information and pattern.

Web Usage Mining

- Web Usage Mining is used for mining social networks, namely online blogs or web data in order to understand and better serve the needs of Web-based applications.
- The web usage mining can also be classified under the different kind of data involved.
  - Web server data
  - Application server data
  - Application level data
  - The Technique used in WUM is clustering having user clusters and page clusters.

DIFFERENCES

Data and Data Format

Data Mining	Text Mining	Web Mining
Mostly the data are in a categorical and numerical order but multimedia data are also included. The data is formatted in a text file/tabular format or custom format. They are mostly structured data	The data are in text formats It could be in a MS word or a an xml files. Textual data are mostly unstructured or weakly structured	Web logs, cookies and web pages are the data used in this solution. The datas are gotten from the WWW and could be structured unstructured or semi-structured. Depending on which type of Web mining is being done.

Data Storage

Data Mining	Text Mining	Web Mining
In data mining, after integration cleaning and transformation, they are all being stored in a **Data Warehouse.** Then our mining process continues from there.	In text mining, our textual data which are in documents are all collected and stored in a document warehouse could also be called corpus.	In web mining, our datas are stored on the web in web log files on a server.

Approaches in Mining

Data Mining	Text Mining	Web Mining
1. Classification according to the kinds of databases mined, 2. Classification according to the kinds of knowledge mined, 3. Classification according to the kinds of techniques utilized, 4. Classification according to the application adapted.	1. The keyword-based approach: where the input is a set of keywords or terms in the documents. 2. The tagging approach: where the input is a set of tags 3. The information-extraction approach: which inputs semantic information, such as events, facts, or entities uncovered by information extraction.	1. It is based on Internet and agent technologies that utilize soft computing and fuzzy logic techniques. 2. After the agent finds their targets, they are programmed to apply the mining technique. 3. These agents then analyse the gathered data, using DM techniques, to understand the habits, patterns found in the WWW

Peculiar Characteristics

Data Mining	Text Mining	Web Mining
Data Mining has a peculiar characteristics of incorporating many algorithms for both prediction and description	Text mining peculiar characteristics is the implementation of the NLP processes which is only peculiar to this mining technique	The data in web mining are its peculiar characteristics in that sometimes the data grow dynamically (changes over time)

SIMILARITIES

Techniques

Data, text and web Mining both shared the same mining technique like Classification, Clustering and Association.

Clustering - in Text Mining use clustering in order to find groups of documents with similar content. Also in Data Mining technique used to place data elements into related groups without advance knowledge of the group definitions and lastly the Web Usage Mining use this techniques for Users cluster or Page Cluster.

Techniques

CLASSIFICATION(Supervised Techniques) - FINDING A MODEL (OR FUNCTION) THAT DESCRIBES AND DISTINGUISHES DATA CLASSES OR CONCEPTS, FOR THE PURPOSE OF BEING ABLE TO USE THE MODEL TO PREDICT THE CLASS OF OBJECTS WHOSE CLASS LABEL IS UNKNOWN.

ASSOCIATION RULE- DISCOVERY OF INTERESTING ASSOCIATIONS AND CORRELATIONS WITHIN DATA AND USES APRIORI AND FP GROWTH ALGORITHM.

Data Processing

Data Mining, Text Mining and Web Mining all shared the same data processing stage such has :

Data selection

Data Cleaning

Data Integration

Data Transformation

Data Processing

Data selection - where data relevant to the analysis task are retrieved from the database.

Data Cleaning - to remove noise and inconsistent data.

Data Integration -where multiple data source may be combined.

Data Transformation - where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

DATA

Data mining, Text Mining and Web Mining all accept large volume of data and involve integration of techniques unlike other machine learning system that does not handle large amount of data.

Data mining, Text Mining and Web Mining have a major relationship in finding new data or knowledge previously unknown to the system.

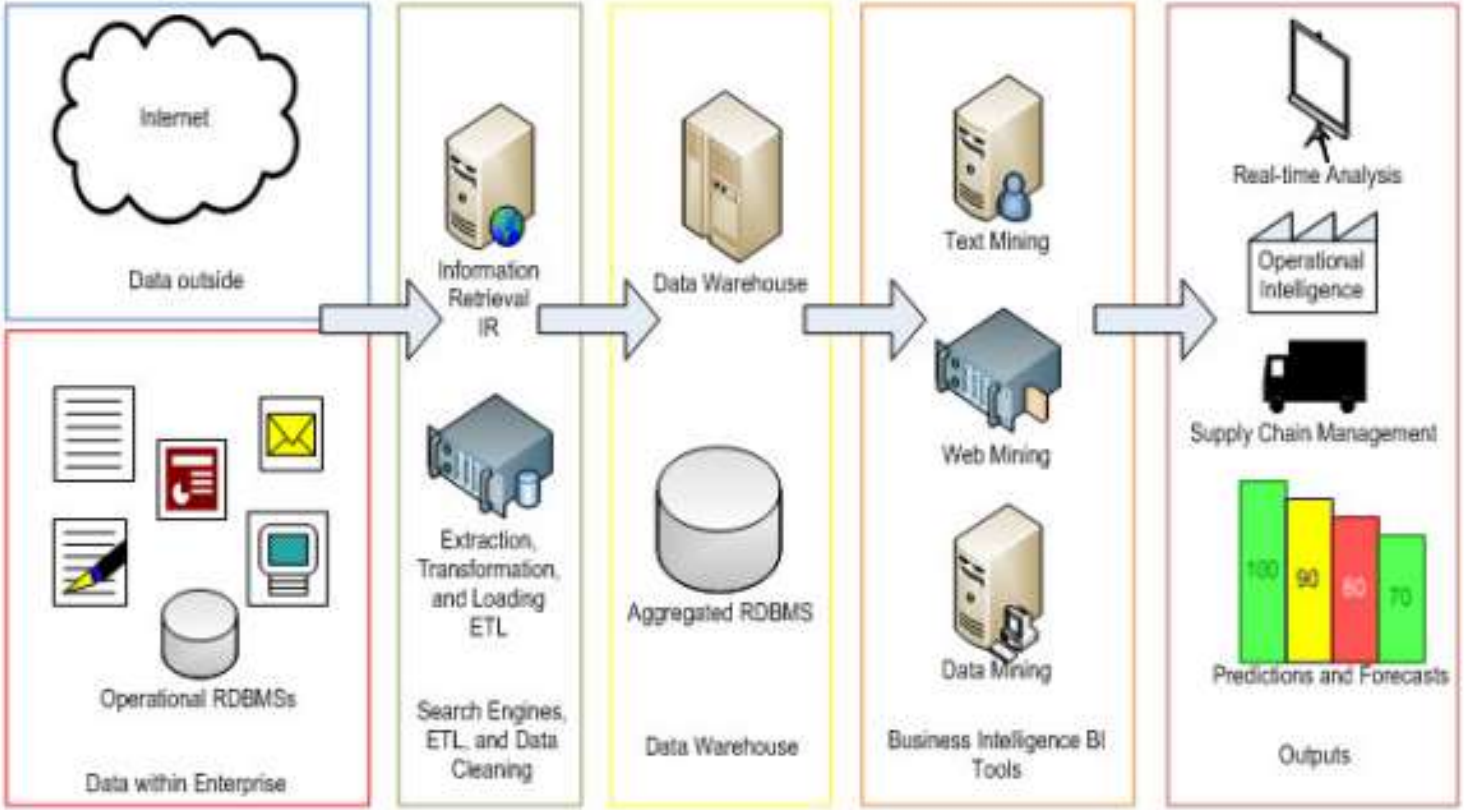
Data mining, Text Mining and Web Mining also known to be viable in retrieving known data, document, web content effectively and efficiently.

SHARED TECHNIQUE

Technique	Data Mining	Text Mining	Web Mining
Clustering	Here, clustering partitions a set of **data** (or objects) into a set of meaningful sub-classes, called **clusters**. Help users understand the natural grouping or structure in a **data** set. Used either as a stand-alone tool to get insight into **data** distribution or as a preprocessing step for other algorithms	Clustering in text mining is done by first converting text data to numeric data using a Document-Term Matrix. K Means clustering is used here.	Clustering is used in Web Mining (Web Usage Mining) dividing it into user clusters and Page clusters.

SHARED TECHNIQUE

Technique	Data Mining	Text Mining	Web Mining
Classification	Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.	Text classification makes us of Supervised Latent Dirichlet Allocation (SLDA) to classify texts into a specified set/ class. They could be classified according to subjects or other needed attributes.	Web Usage mining make use of this technique to classify usage pattern on the web. We could be looking at the Entry and Exit point of the user.


Internet
Data outside
Operational RDBMSs
Data within Enterprise
Information Retrieval IR
Extraction, Transformation, and Loading ETL
Search Engines, ETL, and Data Cleaning
Data Warehouse
Aggregated RDBMS
Data Warehouse
Text Mining
Web Mining
Data Mining
Business Intelligence BI Tools
Real-time Analysis
Operational Intelligence
Supply Chain Management
100
90
80
70
Predictions and Forecasts
Outputs