Appendix A

Text Mining

Text mining, also referred to as *text data mining*, roughly equivalent to **text analytics**, refers to the process of deriving **high-quality information** from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output.
Source: https://en.wikipedia.org/wiki/Text_mining

Text mining is the analysis of data contained in natural language text. The application of **text mining techniques** to solve business problems is called **text analytics.**
Source: http://searchbusinessanalytics.techtarget.com/definition/text-mining

Text mining, also known as **text data mining** or **knowledge discovery** from textual databases, refers to the process of extracting interesting and non-trivial patterns or knowledge from text documents.
Source:http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf

Text mining is a new and exciting area of computer science research that tries to solve the crisis of information overload by combining techniques from **data mining**, **machine learning**, **natural language processing**, **information retrieval**, and **knowledge Management**. In the case of text mining, however, the data sources are document collections, and interesting patterns are found not among formalized database records but in the **unstructured textual data** in the documents in these collections.
Source: http://www.roelsbeestenboel.nl/text.pdf

The enormous amount of information stored in unstructured texts cannot simply be used for further processing by computers, which typically handle text as simple sequences of character strings. Therefore, specific **(pre-)processing method**s and **algorithms** are required in order to extract useful **patterns**. Text mining refers generally to the process of extracting interesting information and knowledge from **unstructured text.**

Source: http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf

Link detection

Link detection relies on a process of building up networks of interconnected objects through various relationships in order to discover **patterns and trends**. The main tasks of link detection are to extract, discover, and link together sparse evidence from vast amounts
of data sources, to represent ånd evaluate the significance of the related evidence, and to learn patterns to guide the extraction, discovery, and linkage of entities.

Source: http://www.roelsbeestenboel.nl/text.pdf

ASSOCIATION AND LINK ANALYSIS

The goal of the techniques often described as *association mining*, *link analysis*, or *sequence analysis* is to detect relationships or associations between specific values of categorical variables in large data sets (although the message can also be applied to continuous variables partitioned into discrete intervals). This is a common task in many data mining projects and often in text mining. For example, these methods may uncover frequently co-occurring terms or phrases in a document corpus and thus detect themes, names, places, and so on that are often mentioned in the same document.

Source:
https://books.google.nl/books?hl=nl&lr=&id=-B6amxqygTMC&oi=fnd&pg=PP2&dq=Link+detection+text+mining&ots=FOgg2xXQ2B&sig=BuWbw_rJMzTG0u9YLtcLEybqzEs#v=onepage&q&f=false

Link Mining

“Links," or more **generically relationships**, among data instances are ubiquitous. These links often exhibit patterns that can indicate properties of the data instances such as the importance, rank, or category of the object.

Link mining refers to data mining techniques that explicitly consider these links when building predictive or descriptive models of the linked data. Commonly addressed link

mining tasks include **object ranking**, **group detection**, **collective classification**, **link prediction** and **subgraph discovery**.
Source: Link

A key challenge for data mining is tackling the problem of mining **richly structured datasets**, where the objects are linked in some way. Links among the objects may demonstrate certain patterns, which can be helpful for many data mining tasks and are usually hard to capture with traditional statistical models.
Source: Link

Information retrieval

Information retrieval (IR) has most usually been construed as the problem of selecting texts from a database in response to some more-or-less well-specified query.
Source: https://pdfs.semanticscholar.org/781d/1cb85b0cb9d4ecdd8c1aee171a57cd5b0008.pdf

Information extraction

We can view IR systems as combine harvesters that bring back useful material from vast fields of raw material. With large amounts of potentially useful information in hand, an IE system can then transform the raw material, refining and reducing it to a germ of the original text
Source: Link

Of particular importance is information extraction (IE), the task of locating specific pieces of data from a natural language document, allowing one to obtain useful structured information from unstructured text.
Source: http://www.aaai.org/Papers/AAAI/1999/AAAI99-048.pdf

Knowledge Discovery from Data(bases)

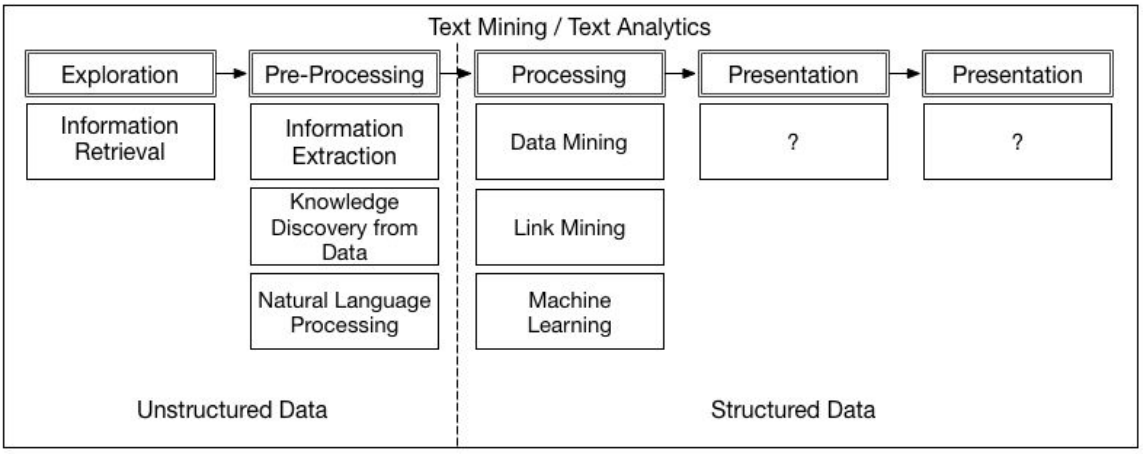
The basic problem addressed by the KDD process is one of mapping low-level data (which are typically too voluminous to understand and digest easily) into other forms that might be more compact (for example, a short report), more abstract (for example, a descriptive approximation or model of the process that generated the data), or more useful (for example, a predictive model for estimating the value of future cases). At the core of the process is the application of specific data-mining methods for pattern discovery and extraction.
Source: http://www.aaai.org/ojs/index.php/aimagazine/article/download/1230/1131

Natural Language Processing

Natural language processing is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction.
Source:https://en.wikipedia.org/wiki/Natural_language_processing

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things.
Source:https://en.wikipedia.org/wiki/Natural_language_processing

Source:https://en.wikipedia.org/wiki/Natural_language_processing



Data Mining

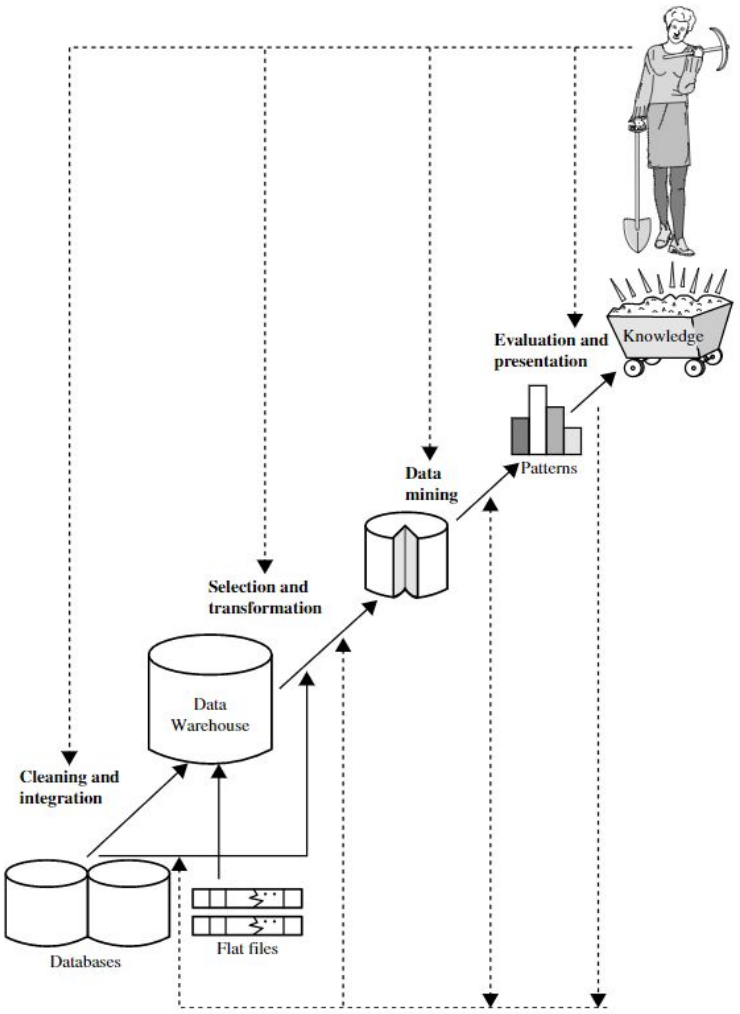
Data mining is the application of specific algorithms for extracting patterns from data.

The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data. Blind application of data-mining methods (rightly criticized as data dredging in the statistical literature) can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns.

Source: http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/1230

Data mining turns a large collection of data into knowledge

An essential process where intelligent methods are applied to extract data patterns



Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

Source: http://ccs1.hnue.edu.vn/hungtd/DM2012/DataMining_BOOK.pdf

Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The data is invariably present in substantial quantities.

Source: Link

Data warehouse

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.
Source:

Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions.
Source: Link

Machine learning

Source:

Supervised Learning

Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances.
Source: https://en.wikipedia.org/wiki/Supervised_learning

Unsupervised Learning

The goal in such *unsupervised learning* problems may be to discover groups of similar examples within the data, where it is called *clustering*, or to determine the distribution of data within the input space, known as *density estimation*, or to project the data from a high-dimensional space down to two or three dimensions for the purpose of *visualization*.

Source: Bishop - Pattern Recognition And Machine Learning - Springer 2006.pdf

Reinforcement Learning

The technique of reinforcement learning (Sutton and Barto, 1998) is concerned with the problem of finding suitable actions to take in a given situation in order to maximize a reward.
Source: Bishop - Pattern Recognition And Machine Learning - Springer 2006.pdf

Reinforcement learning is an area of machine learning inspired by behaviorist psychology, concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward
Source: https://en.wikipedia.org/wiki/Reinforcement_learning

Source:

Knowledge discovery in databases (KDD)

There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data.
http://www.aaai.org/ojs/index.php/aimagazine/article/download/1230/1131/

Clustering

Clustering is a powerful technique for large-scale topic discovery from text. It involves two phases: first, feature extraction maps each document or record to a point in high-dimensional space, then clustering algorithms automatically group the points into a hierarchy of clusters.
Source:http://delivery.acm.org.proxy.library.uu.nl/10.1145/320000/312186/p16-larsen.pdf

Temporal Text Mining

Temporal Text Mining (TTM) is concerned with discovering temporal patterns in text information collected over time
Source:http://delivery.acm.org.proxy.library.uu.nl/10.1145/320000/312186/p16-larsen.pdf

Polypharmacy

Polypharmacy, or the chronic use of multiple medicines, poses significant threats to patients' health. A consensual definition of polypharmacy is lacking, but it is often described as the concurrent use of five or more different chronically used drugs
Source: Link