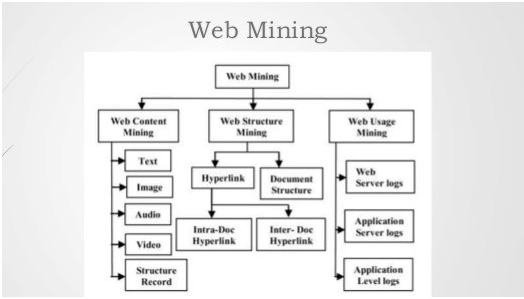
Web Mining and Text Mining – An In-Depth Mining Guide

Web Mining:

Web mining is the process which includes various data mining techniques to extract knowledge from web data categorized as web content, web structure and data usage. It includes a process of discovering the useful and unknown information from the web data.

Web mining can be classified based on the following categories:

**1. Web Content  
2. Web Structure  
3. Web Usage**



Let’s understand concepts of various categories included in web mining.

Web Content Mining:

Web content mining is defined as the process of converting raw data to useful information using the content of web page of a specified web site.

The process starts with the extraction of structured data or information from web pages and then identifying similar data with integration. Various types of web content include text, audio, video etc. This process is called as text mining.

Text Mining uses Natural Language processing and retrieving information techniques for a specific mining process.

Web Structure Mining:

Web graphs include a typical structure which consists of web pages such as nodes and hyperlinks which will be treated as edges connected between web pages. It includes a process of discovering a specified structure with information from the web.

This category of mining can be performed either at document level or hyperlink level. The research activity which involves hyperlink level is called hyperlink analysis.

**Terminologies associated with Web structure:**

**1. Web graph:** It is a directed graph which represents the web.  
**2. Node:** Each web page includes a node of the web graph.  
**3. Link:** Hyperlink is a type of directed edge of the web graph.  
**4. In-degree:** In-degree specifies the number of distinct links that point to a specified node.  
**5. Out-degree:** Out-degree specifies the number of distinct lakes originating at a node that points to other nodes.  
**6. Directed path:** Directed path includes a sequence of links starting from a specified node that can be followed to reach another node.  
**7. Shortest Path:** The shortest path will be the shortest length out of all the paths between p and q.  
**8. Diameter:** The maximum of the shortest path between a pair of nodes p and q for all pairs of nodes p and q in the web graph.

Web Usage Mining:

Web includes a collection of interrelated files with one or more web servers. It includes a pattern of discovery of meaningful patterns of data generated by the client-server transaction.

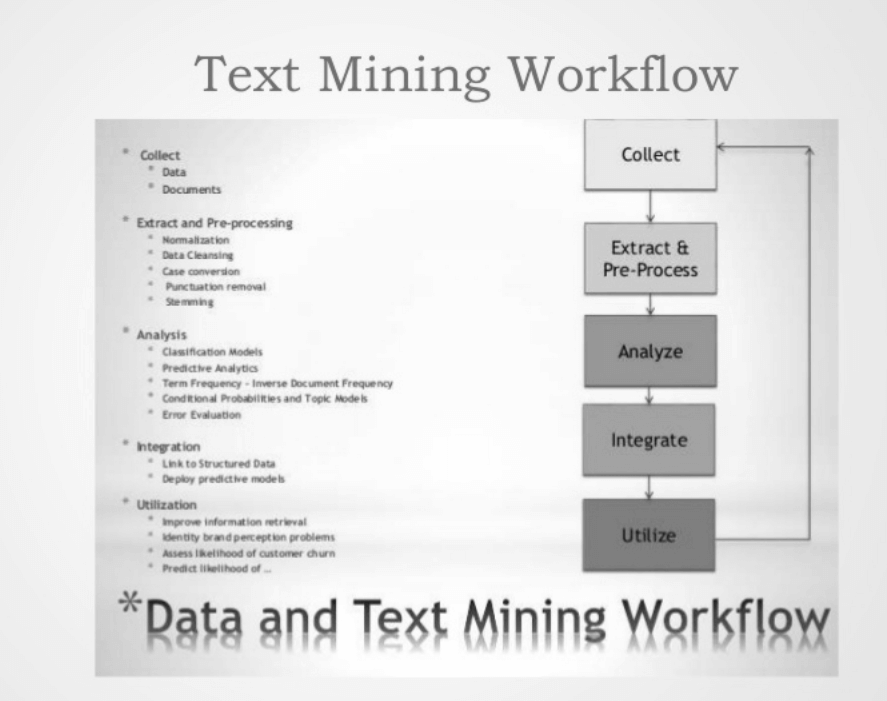
**The typical sources of data are mentioned below:**  
1. Data which is generated automatically is stored in server access logs, referrer logs, agent logs and client-side cookies.  
2. Information of user profiles.  
3. Metadata which includes page attributes and content attributes.

Web server log:

Server logs created by the server record all activities. The page forwarded to the web server includes every single piece of basic information about URL.

Text Mining:

The objective of text mining is to exploit information which is included in textual documents in various patterns and trends in association with entities and predictive rules.  
The results are manipulated and used for:  
1. The analysis of a collection  
2. Providing information about intelligent navigation and browsing method.



Data mining and Text Mining:

1. Both processes seek novel and useful pattern.  
2. Data Mining and Text mining are semi automated process.  
3. The basic difference is the nature of data. Structured data include databases and unstructured data includes word documents, PDF and XML files.  
4. Text Mining imposes a structure to the specified data.

Technology premise of Text Mining:

**1. Summarization:** It is the process of creating a summary of any document which consists of a large amount of information while the theme or main idea of a document is maintained.  
**2. Information Extraction:** It is the process of utilizing relations within the text format. It uses pattern matching format.  
**3. Categorization:** Categorization is the supervised learning technique which places the document according to content. Document categorization is largely used in libraries.  
**4. Visualization:** Visualization is computer graphics used to represent information and visualizing relationships. It is beneficial to depict a more clearer output.  
**5. Clustering:** Clustering involves document’s textual similarity based on the unsupervised technique used for data analysis to divide the text into a manual exclusive group.  
**6. Question Answering:** It includes natural language queries with questions and answers and finding an appropriate solution from the list of patterns.  
**7. Sentiment Analysis:** Sentiment analysis is also known as opinion mining which is configured based on user’s emotion with various categories such as positive, negative, neutral and mixed. It is used to get people’s view and attitude towards anything with services and products.

**Conclusion: –**  
Text and data mining are considered as complementary techniques required for efficient business management. Data mining and text mining tools have gathered its primary location in the marketplace. Natural Language processing is a subset of text mining tools which is used to define accurate and complete domain specific taxonomies. This helps in effective metadata association. Text mining is more mature and efficient in comparison with data mining process. 80 percent of the information is made of text.