(7) **Exeute Job:** It is a MapRedue job. The query executes the job

(8) **Metadata Operation:** Meanwhile the exeution engine can execute the meta data operations with Metastore.

(9) **Fetch Results:** Execution engine Receives the results from Data nodes.

(10) **Send Results:** Exeution engine sends result to Driver

(11) **Send Results:** Driver sends the result to HIVE Interfaces.

MODULE-5

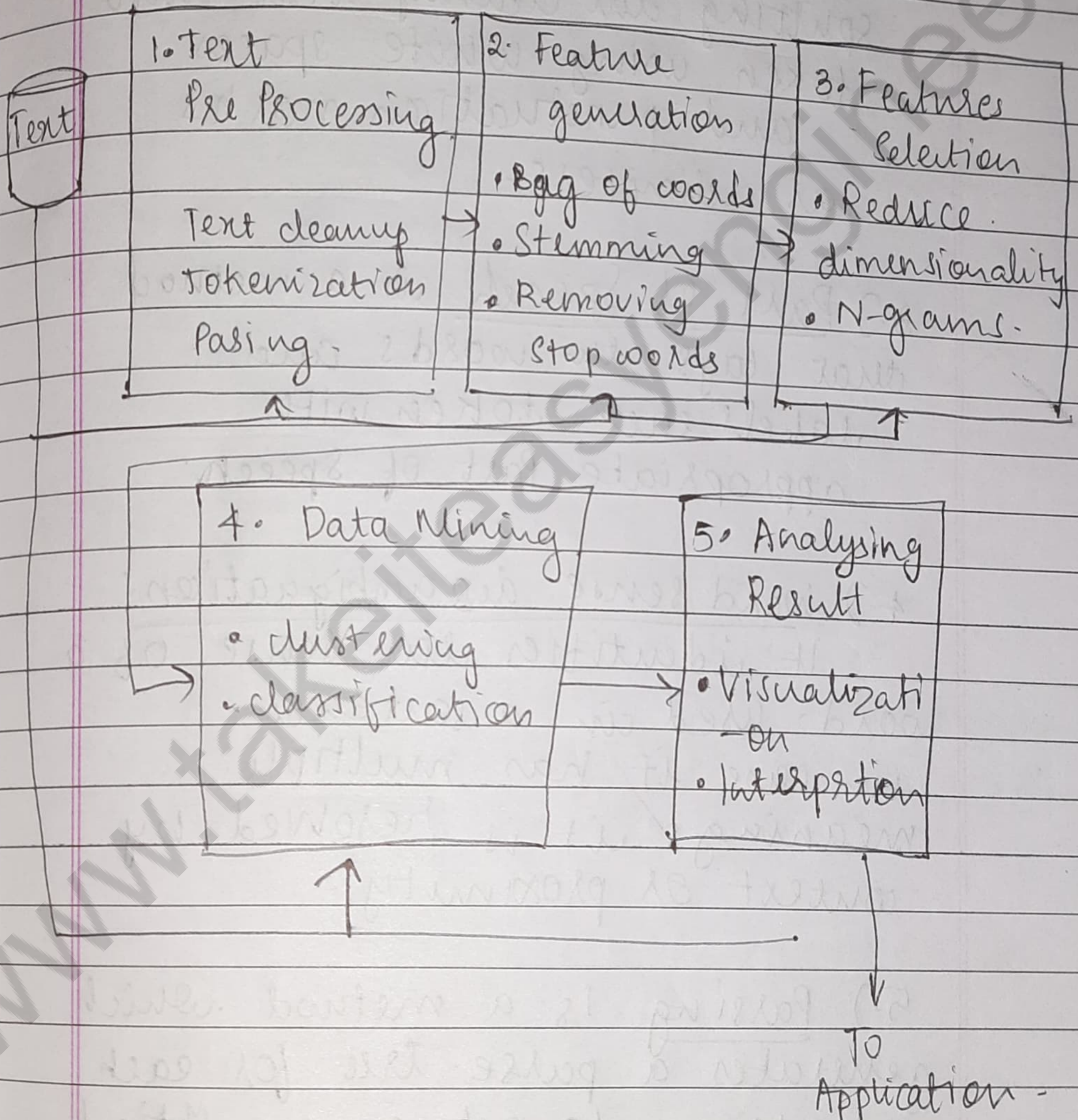
(1) **Phases of text mining:**

- Text is most commonly used for information exchange.
- Text mining is the process that analyses a text, to ~~a~~ extract information useful for a specific purpose.

- Text mining steps are recognizing, extracting and using the information present in words.

Along with searching of words, mining involves search for semantic patterns as well.

- It consist of a process pipline.
- Mining uses the iterative and interactive processes.



Phase 1:

Text pre processing enables Syntatical/ Semantic text analysis and does the following:

1.) Text cleanup is a process of removing unnecessary or unwanted info.

2.) Tokenization is a process of splitting the cleanup text into token using white spaces and punctuation marks as delimiters.

3.) Part of Speech is a method that tags the words and labels each token with appropriate Part of Speech.

4.) Word sense disambiguation: It identifies the sense of a word used in a sentence. in case it has multiple meanings it is resolved by context or proximity.

5.) Parsing is a method which generates a parse tree for each sentence. to get a grammatical relationship b/w different words.

Phase 2: Feature Generation

1.) Bag of words:

Order of words is not important but the frequency of words is.
It provides a document with a bag of words. Document classification methods which use the occurrence of each word as a feature for training a classifier.

2.) Stemming: It identifies a word by its root

i) Normalizes or unifies variations of the same concept such as variations

ii) Removes plurals and normalizes verb tenses.

3.) Removing stop words from the feature space, they are the common words like or, it, in etc

4.) Vector Space Model: It represents a text document as a vector of identifiers, word frequencies or terms in the document index.

Phase 3: Feature Selection.

i) Dimensionality reduction:
Objective is to eliminate irrelavent and redundant data
Redundant features are those which provide no extra information
Correlation helps in finding the redundancy of the feature.
Two features are redundant to each other if their values correlate with each other.

ii) N-gram evaluation:
Finding the number of consecutive words of interest and extract them.

iii) Noise detection: and evolution:
Helps cleaning the data.
It reduces dimensionality that not only improves the performance also but reduces storage requirement for a dataset.

Phase 4: Data mining techniques.

i) Unsupervised Learning:
- For unlabeled data:
- Groups or cluster the data.

ii) Supervised Learning
- The training data is labeled

Date
Page

- New data is classified based on training set.

Phase 5: Analysing Results:

i) Evaluate the outcome of the complete process.

ii) Interpretation of Result:
If acceptable the results are used else they are discarded & try to understand what & why it failed.

iii) Visualization: Create visuals & build prototype.

iv) Use the results for further improvement in activities.

Definition of Web Mining

Web mining refers to the use of techniques and algorithms that extract knowledge from the web data available in the form of web documents and services. Web mining applications are as follows:

(i) Extracting the kagment from a web document that represents the full web document

(ii) Identifying interesting graph patterns or pre-processing the whole web graph to come up with metrics, such as PageRank

(iii) User identification, session creation, malicious activity detection and filtering, and extracting usage path patterns

9.3.2 Web Content Mining

Web Content Mining is the process of information or resource discovery from the content of web documents across the World Wide Web. Web content mining can be (i) direct mining of the contents of documents or (ii) mining through search engines. They search fast compared to direct method.Web content mining relates to both, data mining as well as text mining. Following are the reasons:

(i) The content from web is similar to the contents obtained from database, file system or through any other mean. Thus, available data mining techniques can be applied to the web.

(ii) Content mining relates to text mining because much of the web content comprises texts.

(iii) Web data are mainly semi-structured and/or unstructured, while data mining is structured and the text is unstructured.

App1ications

Following are the applications of content mining from web documents:

1. Classifying the web documents into categories
2. Identifying topics of web documents
3. Finding similar web pages across the different web servers

4. Applications related to relevance:

(a) Recommendations - List of top “n” relevant documents in a collection or portion of a collection

(b) Filters - Show/Hide documents based on some criterion

(c) Queries — Enhance standard query relevance with user, role, and/or task-based relevance.

9.3.3 Web Usage Mining

Web usage mining discovers and analyses the patterns in click streams. Web usage mining also includes associated data generated and collected as a consequence of user interactions with web resources. Figure s.7 shows three phases for web usage mining.

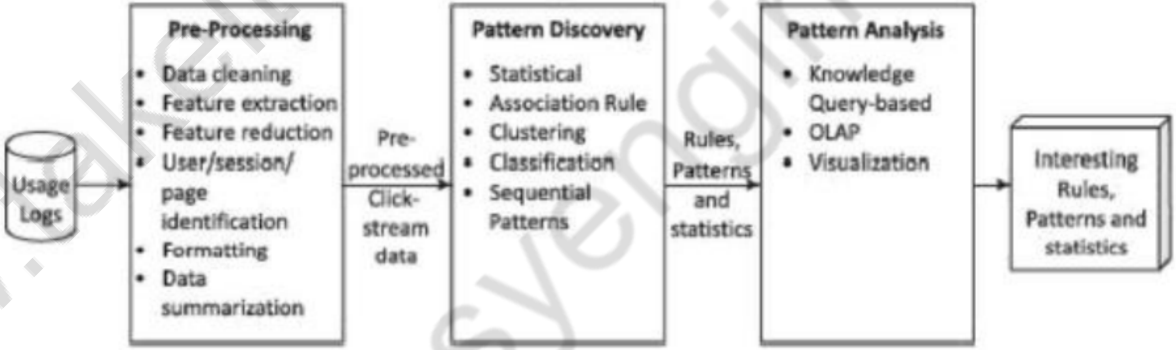


Figure 9.7 Process of web usage mining The phases are:

1. Pre-processing - Converts the usage information collected from the various data sources into the data abstractions necessary for pattern discovery.

2. Pattern discovery — Exploits methods and algorithms developed from fields, such as statistics, data mining, ML and pattern recognition.

3. Pattern analysis - Filter outs uninteresting rules or patterns from the set found during the pattern discovery phase.

single-user, multi-user, single-site access and multi-site access patterns.

9.3.3.1 Pre-processing

The common data mining techniques apply on the results of pre-processing using vector space model (Refer Example 9.2). Pre-processing is the data preparation task, which is required to identify:

(i) User through cookies, logins or URL information

(ii) Session of a single user using all the web pages of an application

(iii) Content from server logs to obtain state variables for each active session

(iv) Page references.

The subsequent phases of web usage mining are closely related to the smooth execution of data preparation task in pre-processing phase. The process deals with (i) extracting of the data, (ii) finding the accuracy of data, (iii) putting the data together from different sources, (iv)transforming the data into the required format and (iv) structure the data as per the input requirements of pattern discovery algorithm.

Pre-processing involves several steps, such as data cleaning, feature extraction, feature reduction, user identification, session identification, page identification, formatting and finally data summarization.

9.3.3.2 Pattern Discovery

The pre-processed data enable the application of knowledge extraction algorithms based on statistics, ML and data mining algorithms. Mining algorithms, such as path analysis, association rules, sequential patterns, clustering and classification enable effective processing of web usages. The choice of mining techniques depends on the requirement of the analyst. Pre-processed data of the web access logs transform into knowledge to uncover the potential patterns and are further provided to pattern analysis phase.

Some of the techniques used for pattern discovery of web usage mining are:

Statistical techniques They are the most common methods which extract the knowledge about users. They perform different kinds of descriptive statistical analysis (frequency, mean, median) on variables such as page views, viewing time and length of path for navigational.

Statistical techniques enable discovering:

(i) The most frequently accessed pages

(ii) Average view time of a page or average length of a path through a site

(iii) Providing support for marketing decisions

Association rule The rules enable relating the pages, which are most often referenced together in a single server session. These pages may not be directly connected to one another using the hyperlinks.

Other uses of association rule mining are:

(i) Reveal a correlation between users who visited a page containing similar information. For example, a user visited a web page related to admission in an undergraduate course to those who search an eBook related to any subject.

(ii) Provide recommendations to purchase other products. For example, recommend to user who visited a web page related to a book on data analytics, the books on ML and Big Data analytics also.

(iii) Provide help to web designers to restructure their websites.

(iv) Retrieve the documents in prior in order to reduce the access time when loading a page from a remote site.

The objective of pattern analysis is to filter out uninteresting rules or patterns from the rules,patterns or statistics obtained in the pattern discovery phase.

Pattern Analysis

The most common form of pattern analysis consists of:

(i) A knowledge query mechanism such as SQL

(ii) Another method is to load usage data into a data cube in order to perform Online Analytical Processing (OLAP) operations

(iii) Visualization techniques, such as graphing patterns or assigning the colors to different values, can often highlight overall patterns or trends in the data

(iv) Content and structure information can filter out patterns containing pages of a certain usage type, content type or pages that match a certain hyperlink structure.