Experiment No 10

Aim: To compare Web Data Mining Techniques, Tools and Algorithms

Theory: Web mining is one of the types of techniques use in data mining. The main purpose of web mining is to automatically extract information from the web. For discovering useful data (videos, tables, audio, images etc.) from the web different techniques and tools are used. Information over the internet is huge and increasing with passage to time due to which size of data bases are also growing. Digging knowledgeable information and analyzing the data sets for relevant data is much difficult because data over the internet in not in plain text. It could be unstructured data, multimedia, table, tag.

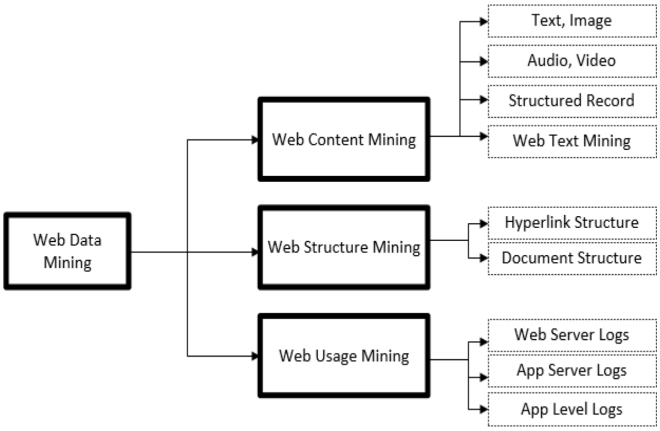
Web mining is actually an area of data mining related to the information available on internet. It is a concept of extracting informative data available on web pages over the internet . Users use different search engines to fetch their required data from the internet, that informative and user needed data is discovered through mining technique called Web Mining. Different tools and algorithms are used for extraction of data from web pages that includes web documents, images etc. Web mining is rapidly becoming very important due to size of text documents increasing over the internet and finding relevant patterns, knowledge and informative data is very hard and time consuming if it is done manually. Structure (Hyperlinks), Usage (visited pages, data use), content (text document, pages) are included in information gathered through Web mining . Term World Wide Web is related to the combination of web documents, videos, audios etc.

Web Mining is sub categorized in to three types as shown in Fig. 1:

A. *Web Content Mining*

B. *Web Structure Mining*

C. *Web Usage Mining*



A. *Web Content Mining:*

Content Mining is a process of Web Mining in which needful informative data is extracted from web sites (WWW). Content includes audio, video, text documents, hyperlinks and structured record . Web contents are designed to deliver data to users in the form of text, list, images, videos and tables. Over last few decades the amount of web pages (HTML) increases to billions and still continues to grow. Searching query into billions of web documents is very difficult and time consuming task, content mining extracts queried data by performing different mining techniques and narrow down the search data which become easy to find required user data.

B. Web Structure Mining

Now a day's massive amount of data is increasing on web. World Wide Web is one of the most loved resources for information retrieval. Web mining techniques are very useful to discover knowledgeable data from web. Structure mining is one of the core techniques of web mining which deals with hyperlinks structure [14]. Structure mining basically shows the structured summary of the website. It identifies relationship between linked web pages of websites. Continues growth of data over the internet become a challenging task to find informative and required data [15]. Web mining is just a data mining which digs data from the web. Different algorithmic techniques are used to discover data from web. Structure mining analyzes hyperlinks of the website to collect informative data and sort out in categories like similarities and relationship. Intra-page is a type of mining that is performed at document level and at hyperlink level mining is known as inter-page mining. Link analysis is an old but very useful method that is way its value increases in the research area of web mining – Structure analysis is also called as Link-mining.

C. *Web Usage Mining Techniques:*

Following three techniques are described in detail with their sub approaches use in web usage mining. Each technique performs different tasks in a hierarchy.

- Data Preprocessing
- Pattern Discovery
- Pattern Analysis

Web Mining Categories	**Techniques**	**Tools**	**Algorithms**
Web Content Mining	- Unstructured Data Mining - Structured Data Mining - Semi – Structure Data Mining - Multimedia Data Mining	- Screen Scaper - Mozenda - Automation Anywhere7 - Web Content Extractor - Web Info Extractor - Rapid Miner	- Decision Tree - Naive Bayes - Support Vector Machine - Neural Network
Web Structure Mining	- Link-based Classification - Link-based Cluster Analysis - Link Type - Link Strength - Link Cardinality	- Google PR Checker - Link Viewer	- Page Rank Algorithm - HITS algorithms (Hyperlink Induced Topic Search) - Weighted Page Rank Algorithm - Distance Rank Algorithm - Weighted Page Content Rank Algorithm - Webpage Ranking Using Link Attributes - Eigen Rumor Algorithm - Time Rank Algorithm -Tag Rank Algorithm - Query Dependent Ranking Algorithm
Web Usage Mining	**- Data Preprocessing** Data Cleaning User & Session Identification **- Pattern Discovery** Statistical Analysis Association Rules Clustering Classification Sequential Patterns **- Pattern Analysis** Knowledge Query Mechanism OLAP (Online Analytical processing) Intelligent Agents	**- Data Preprocessing Tools** Data Preparator Sumatra TT Lisp Miner SpeedTracer **- Pattern Discovery Tools** SEWEBAR-CMS i-Miner Argunaut MiDas(Mining In-ternet Data for As-sociative Sequenc-es) **- Pattern Analysis Tools** Webalizer Naviz WebViz WebMiner Stratdyn	**- Association Rules** Apriori Algorithm Maxi-mal Forward References Markov Chains FP Growth Prefix Span **- Clustering** Self-Organized Maps Graph Partitioning Ant Based Technique K-means with Genetic algorithms Fuzzy c-mean Algorithm **- Classification** Decision Trees Naïve Bayesian Classifiers K-nearest Neighbor Classifiers Support Vector Machine **- Sequential Patterns** MIDAS (Mining Internet Data for Association

			Sequences) algorithm

USAGE MINING TECHNIQUES COMPARISON

Usage Mining Techniques	Methods	Data Gathering	Data Store	Advantages	Important Algorithms
Data Preprocessing	- Web status codes	- Data logs - Website - Users login information - Web access logs - Cache - Cookies etc.	- Web logs	- Convert raw data to understandable Common LF and Extended CLF for recording	- Apriori algorithm - FP Growth
Pattern Discovery	- Frequency, median, mode used to show length, recently accessed, view time ofpages	- Filtered data from preprocessing section	- Session logs	- Extract useful information from discovered patterns correlations	- K-means with Genetic algorithms - Fuzzy c-mean Algorithm
Pattern Analysis	- Roll-up - Drill Down/Up	- Pattern discovery	- Data cube (mult	- Irrelevant rules and patterns are	- SQL Language - OLAP

			i-dimensional database)	separated	

Conclusion: Thus the Web Data Mining Techniques, Tools and Algorithms are studied and compared.

Viva Voce:

1. What is Web mining? How does it differ from regular data mining or text mining?
2. What are the three main areas of Web mining?
3. What is Web content mining? How can it be used for competitive advantage?
4. What are the three Web usage mining techniques