Web Mining

Web Mining

- Web mining is a technique to discover and analyze the useful information from the Web related data and services.
- According to Etzioni [2], web mining can be divided into four subtasks:(i)
- Information Retrieval/Resource Discovery (IR):
  - Find all relevant documents on the web. The goal of IR is to automatically find all relevant documents, while at the same time filter out the non relevant ones.
  - Search engines are a major tool people use to find web information.

Web Mining

(ii) Information Extraction (IE):

- Automatically extract specific fragments of a document from web resources retrieved from the IR step. Building a uniform IE system is difficult because the web content is dynamic and diverse.
- Most IE systems use the “wrapper" [3] technique to extract specific information for a particular site.
- Machine learning techniques are also used to learn the extraction rules.

(iii) Generalization:

Discover information patterns at retrieved web sites. The purpose of this task is to study users' behaviour and interest.

Data mining techniques such as clustering and association rules are utilized here

(iv) Analysis/Validation:

Analyze, interpret and validate the potential information from the information patterns.

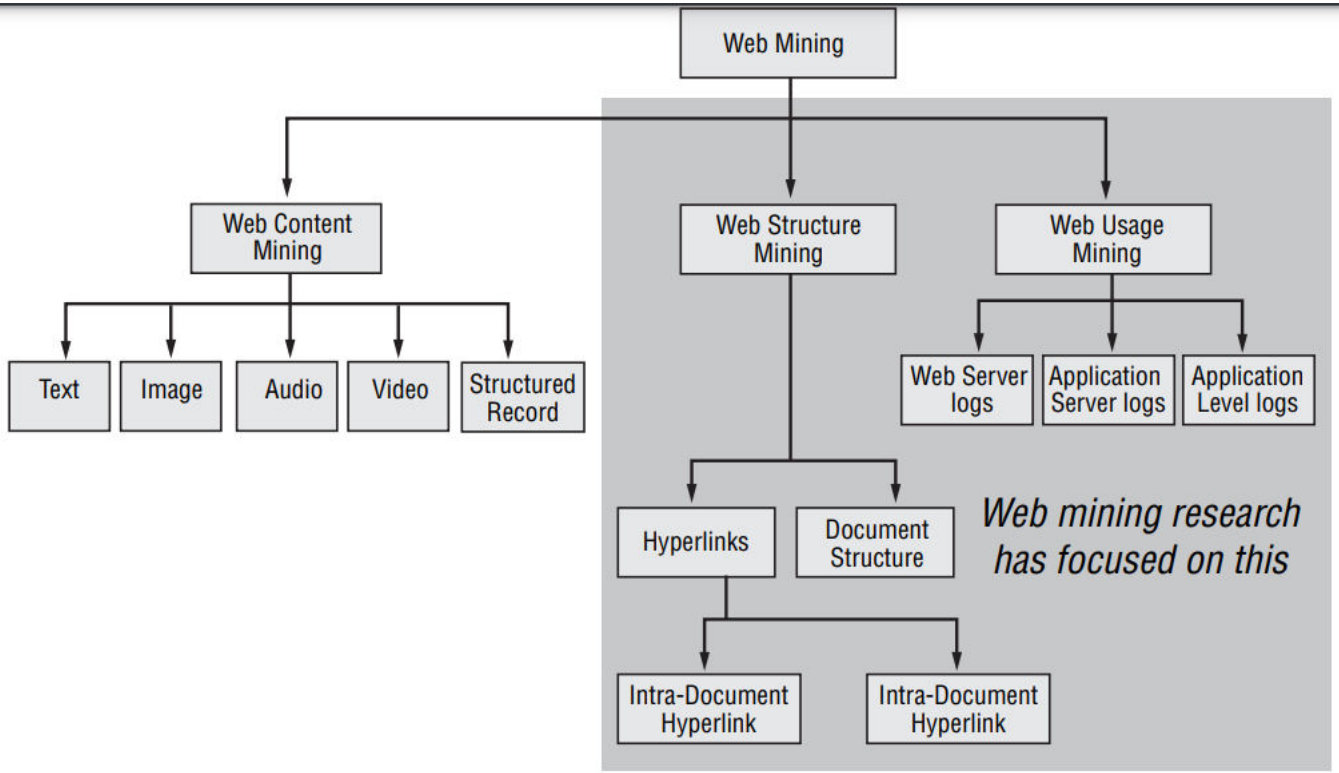
The objective of this task is to discover knowledge from the information provided by former tasks. Based on web data, we can build models to simulate and validate web information.

Challenges in web mining

Along with opportunities, there are serious challenges in web mining

- Web is noisy : A Web page typically contains a mixture of many kinds of information, e.g., main contents, advertisements, navigation panels, copyright notices, etc.
- Web is dynamic : Information on the Web changes constantly. Keeping up with the changes and monitoring the changes are important issues.
- Web is a virtual society : It is not only about data, information and services, but also about interactions among people, organizations and automatic systems, i.e., communities.
- Many other such restrictions pose a pretty big challenge for mining the web.

Web Mining Taxonomy



Web mining Taxonomy

Web Mining : View , Data, Method and application

Web mining: view, data, method and application

	Web Mining			
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR view	DB view		
View of data	Unstructured , Semi-structured	Semi-structured Web site as DB	Links structure	Interactivity
Main Data	Text Document , Hypertext document	Hypertext document	Links structure	Server Log, browser Log
Representation	Bag of words, n-grams, Terms, Phrases, Concepts or ontology, relational	Edge- labeled graph (OEM), Relational	Graph	Relational table, Graph
Method	TFIDF and variants, Machine learning, Statistical	Proprietary algorithms, ILP, association rules	Proprietary algorithms [1]	Machine learning, Statistical, association rules
Application Categories	Categorization, clustering, Finding extraction rules, finding patterns in text, user modeling	Finding frequent sub structures, web site schema discovery	Categorization, Clustering	Site construction, adaption, and management, Marketing, user modeling

Different approaches for web content mining

Different approaches for web content mining

Web Content Data Type	Mining Method	Techniques
Unstructured Data	Unstructured text mining	Information Extraction
		Topic Tracking
		Summarization
		Categorization
		Clustering
		Information Visualization
Structured Data	Structured Data Mining	Web Crawler
		Wrapper Generation
		Page content Mining
Semi-Structured Data	Semi-Structured Data Mining	Object Exchange Model (OEM),
		Top Down Extraction
		Web Data Extraction language
Multimedia Data	Multimedia Data Mining	SKICAT
		Color Histogram Matching
		Multimedia Miner
		Shot Boundary Detection