Comprehensive Review on Web Content Mining

Er. Reena , Er.Shaveta Angurala
Asstt.Professor, DAVIET, Jal.

Abstract

Web is composed of huge and diverse information. Information exists in the form of Hyperlinks having structured tables, semi-structured and unstructured texts and multimedia (audio, video, images). Web mining aim is to drill useful information from hyperlinks, page contents and usage data. Web Mining is categorized into three categories: Web Structure Mining, Web Content Mining, and Web Usage Mining. Web Content Mining mines useful information from Web Page Contents and discover patterns in Web Pages. Web Content Mining is based on text mining and information retrieval (IR) techniques and is used to discover what a web page is about and how to uncover new knowledge from it. We broadly classify Data mining into Web Mining, Text Mining because most of the Web part is unstructured data i.e. text and some structured data i.e. Database Tables. Web Content Mining is also referred to as Web Text Mining. The technologies which are used in Web Content Mining are: Natural Language Processing and Information Retrieval.

1. Introduction

Combination of Text Mining, Data Mining, Information Retrieval and Machine Learning provides the framework for Web Content Mining. Its basic purpose is to filter the information, improve the information finding and to integrate the data on the web so that more sophisticated queries can be performed on the internet. In simple terms mining, extraction and integration of useful data, information and knowledge from Web page contents. It is related to data mining and text mining because web has almost free text and data mining techniques are applied to them. Content data corresponds to the facts for which the web page was designed to convey to users.

It is processing of Information or resource discovery from contents of millions of sources across the World Wide Web. Web Pages are made up of Hypertext Mark up Language. Processes involves in Web Mining as follows:

- ✓ Retrieval of Web Documents
- ✓ Pre-Processing
- ✓ Discovery of Patterns
- ✓ Analysis

1.1) Retrieval

The process of retrieving the available data on the internet such as newsletters, newsgroups , text contents of HTML documents by removing their tags ,text contents and text databases.

1.2) Pre-processing

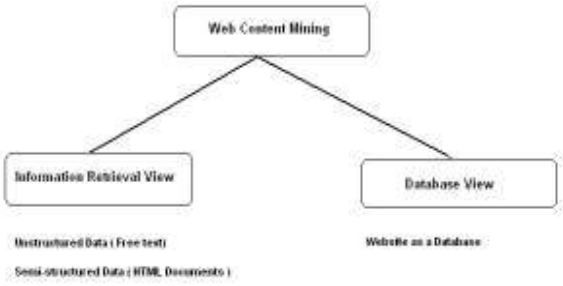
It is the process of transformation the original data which we have retrieved in previous stage. Its aim is to transform the information into some logic form. The transformations can be stemming, stop-word removal.

1.3) Pattern Discovery

Discovers Pattern at individual web sites and the machine learning and data mining techniques are used for it e.g. Sequential Patterns, Association rules. It is really important stage for the interpretation and validation of the discovered patterns.

1.4) Analysis

It is the interpretation and validation of mined patterns. The goal of this process is to eliminate irrelevant rules and extract interesting patterns from pattern discovery output.



1.1Web Content Mining View of Data

Various techniques are applied in it for Web pages representation and content mining. These are as follows:

A) Pre-processing
B) Document Representation
C) Web Mining Operations.

2. Pre-Processing

It is to filter the data which we have retrieved from the web by navigation and through browsing. It makes the web data suitable for mining.

2.1) Pre-processing for Text Most of the web Part is unstructured data and to transform that original data we have some techniques:

Perceptual Grouping
Tokenization
POS Tagging } NLP Tasks
Parsing
Stemming
Stop word elimination and filtering
Latent Semantic Indexing

I) Perceptual Grouping extracts some document level fields such as Author, Title and labeling the text zones. It is used to extract some internal text zones such as paragraphs, columns or tables.

II) Bag of words takes single words and order is discounted. Single words result can be Boolean means whether they are present or not. Each document is treated as a bag of words or terms. These terms whose semantics helps to remember document themes.

III) Tokenization means breaking the text into sentences and words. This is done at different levels and resulting data is referred as tokens which mean character stream is divided into meaningful constituents.

IV) Part of Speech Tagging divides words into categories based on the role in the sentence. It provides the information about the semantic content of word. It determines Part of speech tag e.g. noun, verb, adjective for each term.

V) Stop word elimination and filtering means to remove words that bear little or no content information like articles, prepositions and conjunctions. In Filtering elimination of the common words takes place or the words that occur too often.

VI) Stemming reduces the words to a common root e.g. analysis, analyze, analyzing ---- analy.In stemming we remove the suffixes that seems to be identical and we can create the database of the words and their relationships.

VII) Parsing produces a parse tree of a sentence which helps in defining relation of one word to another and its function in the sentence.

VIII) Latent Semantic Indexing Document is represented by D = t1 w1
Where t (term) = keywords or content descriptions
w (weight)= measure of the importance of term

Any document is represented as its terms and its associated weights. Term frequency represents how repeated words are strongly related to content and on the basis of this an index is maintained. That index includes the number of terms and is known as Latent Semantic Indexing. This is efficient approach and we can drop low frequency words.

2.2) Pre-Processing for Semi-structured Data

Besides the text Web Pages also has hyperlinks and anchor texts which do not exist in free text. We follow some other pre-processing Techniques for the HTML documents.

- **Identification of Different Text fields** - In HTML we have different text fields e.g. title, metadata, body. Title has higher weight than any other fields because it is a small description of the page. Other fields also have weights. These terms are treated differently.

- **Identification of Anchor Text** - Anchor text associated with a hyperlink is treated specially in search engines because the anchor text often represents a more accurate description of the information contained in the page pointed to by its link.
- ✓ **Removal of HTML tags-** The removal of HTML tags can be dealt with similarly to punctuation. One issue needs careful consideration, which affects proximity queries and phrase queries.

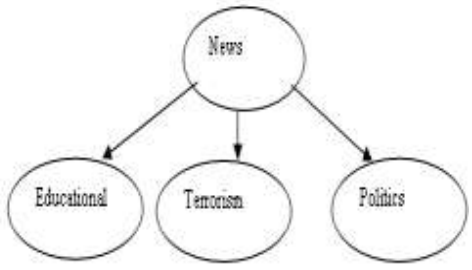
- ✓ **Identify Main content block** In Web Pages we have some information like advertisements and banners which are not main content. Two techniques are used to find main content block.

(a) **Partitioning based on visuals**: This method uses visual information to help find main content blocks in a page. Visual or rendering information of each HTML element in a page can be obtained from the Web browser. For example, Internet Explorer provides an API that can output the X and Y coordinates of each element. A machine learning model can then be built based on the location and appearance features for identifying main content blocks of pages

(b) **Tree matching**: HTML has a nested structure, so a Tag Tree is built for page. By this technique we can be able to find some hidden templates that may contain main content blocks when they are quite different in different web pages in same template

3. Web Page Document Representataion

3.1 Graph Representation of Document -Web document content in the form of graphs. The most occurring terms are represented as nodes of the graph. Edges represent the Links and nodes represent terms. Stop word removal and stemming represents vertex of graph. Each vertex represents unique word.



3.1 Graph Representation of a Document

$|G|=|V| + |E|$

G : Size of Graph
V : Vertices
E : Edges

3.1.1 Traditional Representation using Vector Space Model Relevance ranking of the documents is obtained using this model. Document is represented by a weight vector where weight is associated to each term.
Term frequency TFi denotes number of times tw (weight of term) appears in document D.

$$\text{Normalized term Frequency NF} = \frac{TFi}{\text{Max (TF1, TF2,……..TFV)}}$$

V=Vocabulary i.e. size of Collection

As we know web pages also have hyperlinks and other tags .Its document representations are different from free text documents.

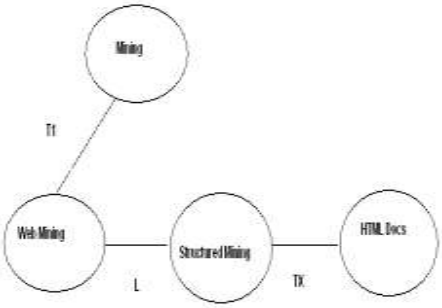
3.1.2 Standard Graph Representation of Web page – Various Sections we are defined for HTML documents for its representation are:

Title which contains the text related to the document's title and any provided keywords (meta-data)

Link which is text that appears in clickable hyper-links on the document

Text, which comprises any of the readable text in the document (this includes link text but not title and keyword text).

Ovals indicate terms and edges are labeled Title (T1), link (L), Text (TX).



3.2 Standard Graph Representation

Web Document Data sets We represent HTML documents in data sets which are useful to represent it in graphs and multiple classes represent various

grouping which actually help in finding the content of documents.

3.2 DOM Tree Representation A well-formed HTML document is a properly nested hierarchy of regions that is represented by a tree-structured Document Object Model or DOM. In a DOM tree, internal nodes are elements and some of the leaf nodes are segments of text. Some other nodes are hyperlinks to other Web pages, which can in turn be represented by DOM trees. In DOM trees BODY of HTML represents root of Tree .The rectangular boxes represent the other attributes as a tag node. Shaded rectangle shows actual content

<BODY bgcolor=RED>

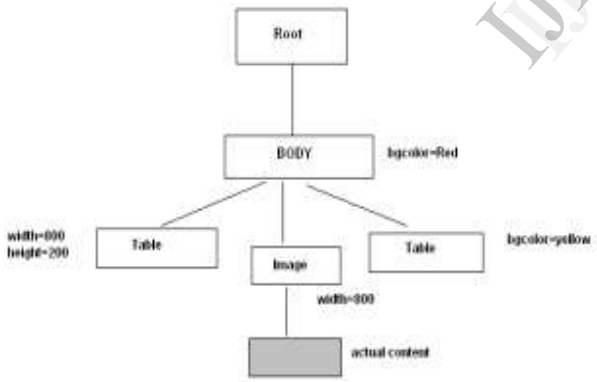
<TABLE width=800 height=200>
.
</TABLE>

<IMG src="sun..gif" width=800>

<TABLE bgcolor=YELLOW>

</TABLE>

</BODY>



3.3 DOM Tree Representation

If we view a Website as a database then we have one more representation for it to simplify its overall structure and to precisely give the overview of its structure. Database which has various relationships with various objects and interlinks.

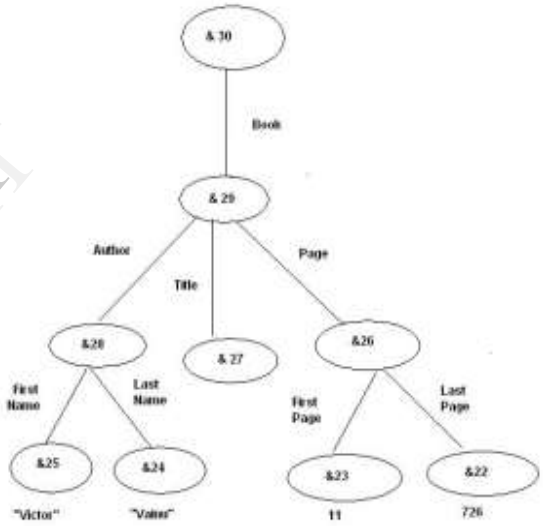
3.3 Object Exchange Model – Represent semi-structured data by this model and proposed it for modeling the data. It represents semi-structured data by a labeled graph with objects as vertices and labels as graphs. Each object in OEM has certain structure

Label – What the object Represents

Type--Data type of object value .it can be atomic i.e. string, number etc.

Object -id –Unique identifier

Unique identifier identifies the objects; labels connect the objects and its sub objects which describes the relationship among them. It can also be treated as graph where nodes are objects and labels are on the edges. For the extraction of structures this model can produce the most specific schema. Schematic information is embedded on labels and it is self describing model. Every object has some value i.e. atomic e.g. string, html, gif or complex in the form of objects references.



3.4 Object Exchange Model

The main title (on the first page) should begin 1-3/8 inches (3.49 cm) from the top edge of the page, centered, and in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two 12-point blank lines after the title.

4. Web Mining Operations

These operations are used to discover patterns using various algorithms and techniques which are predictive and descriptive. It consist of various mechanisms for discovering patterns of concept occurrence within a

given document collection or subset of a document collection.

4.1) Classification or Text Categorization It gives set of categories and collection of text documents, the process of finding the correct topic for each document. It is also knows as Supervised learning or Inductive learning in machine learning.

Four step scenario applied for unstructured documents in Supervised Learning

a) Data Collection
b) Building the Model
c) Testing and Evaluating the Model
d) Using the model to classify new documents

There are several types of supervised learning tasks. This section will discuss the Supervised learning Techniques for text and Hypertext.

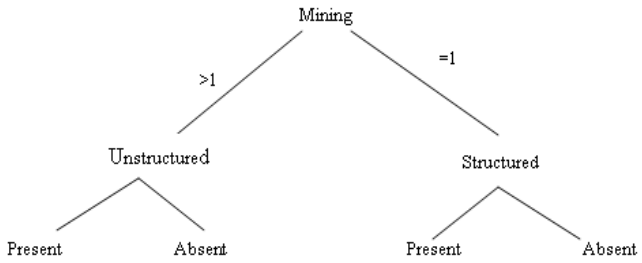
4.1.1 Nearest Neighbor Algorithm It is a non-parametric method. For deciding whether that document D has a similarity with previously defined training documents in the set .The proportion of neighbors having same class may be taken as estimator for the probability of that class and class with largest proportion is assigned to document D.

4.1.2 Naïve Bayes Algorithm This approach assumes that we can compute or estimate the distribution of terms within the documents assigned to these categories. Firstly transform the probability of term occurrence given a category into probability of category given a term occurrence

P (Ci | TD) – Probability of category C given terms T in document D

$$P(Ci|D) = \frac{P(D|Ci)\,P(Ci)}{P(D)}$$

4.1.3 Decision Trees construct a tree that incorporates those feature tests need to discriminate between objects of different classes. Leaf nodes represent the categories non-uniquely because more leaf nodes can represent same category. It works best with large data sets. They are one of the most popular machine learning techniques currently used. Decision trees are is a probabilistic classifier – confidence (class) represents a probability distribution. They are easy to interpret, however they require a number of model parameters which is usually hard to find, and error estimates are difficult. Decision trees consist of a series of simple decision rules, often presented in form of a graph. In the following Representation Decision tree tells us that if "mining" term is present more than once then there should be one more term present i.e. "Unstructured" and if it is present only once then "structured "term should be there.



2.3.1) Decision Tree Representation

4.1.4 Decision lists are strictly ordered and contain only Boolean values. A decision list for a document, D, with respect to a category, C, is essentially a list of rules of the form

if w1 E D & . . .& wn E D then D E C

Finally, we turn to a discussion of techniques to address supervised learning for hypertext. Apart from plain text, HTML, the dominant form of hypertext on the Web, contains many different kinds of features.

Web Page Classification –In this we use hierarchical Web page Categorization for the automatic classification of Web Pages. This is useful because search is restricted to particular pages including those topics. It constraints the number of documents belonging to particular category. If it exceeds then it is further divided into sub-categories. Another feature of the problem is the hyper textual nature of the documents. The Web documents contain links, which may be important sources of information for the classifier because linked documents often share semantics. It can be dealt with by using a separate classifier at every branching point of the hierarchy.

Relational Learning -- In Traditional Methods HTML structure, interlinking is ignored. Hyperlink-based classification can be implemented by using graph models (e.g., sub graph isomorphism). It is also known as First order learning or Inductive logic programming.

which uses the language of logic programming (or Prolog) as a representation language for learning.

Let we have documents D and terms t. If term t is there in document D then in means contains (D,t)is true. Background knowledge is used by relational learning to derive clauses.

Class A (D):-Contains (D, Art)
Class B (D):-Contains (D, Biology)

It means if Document D contains "Art" then it belongs to class A. If D contains "Biology" then it belongs to class B.

Rule Induction – It is basically based on algorithmic approach of Relational Learning and also known as First order logic Algorithm. In background knowledge we have positive examples from which algorithm will generate negative examples by using closed word assumption approach. It covers the entire negative and positives covered which will be used for later on evaluation. Information gain evaluation will give the best literal.

Hyperlink Ensembles --The pages which we want to classify are predecessor pages of target pages and they may contain better information. In this approach we make separate predictions of the various predecessor pages so that feature sets are treated differently. We represent a page with a set of instances and each instance consists of features of predecessor page and class label of target page. A classifier is able to predict a class for links between the pages. All predictions that involve same target page uniformly predict same class.

4.2) Clustering is a process through which objects are classified into groups called clusters. It is also known as Unsupervised learning. The problem is to group the unlabelled collection into some meaningful clusters. Clustering tasks in text analysis: Improving search recall, Search Precision, Query specific clustering. The utility of clustering for text and hypertext information retrieval lies in the so-called cluster hypothesis.

4.2.1 K-means Algorithm It is also known as partitional clustering algorithm. It proceeds as follows:
k-seeds from the core of k-clusters (C1,C2……Ck) which is a collection of vectors (x1,x2,……xn).
Every vector is assigned to cluster of closest seed. Centroid of clusters are computed

$$M_i = |C_i|^{-1} \sum_{x \in c_i} x.$$

When no more changes occur then it converges.

4.2.2 Probabilistic Algorithm –The basic idea is the assignment of probabilities for the membership of a document in a cluster. Each document can belong to more than one cluster according to the probability of belonging to each cluster. Compute the probability or likelihood of all objects x.Then relabel all objects according to computed probabilities.

$$L = \sum_{i=1}^{n} \log \sum_{A} P(d_i \mid A) P(A)$$

4.2.3 Hierarchical Clustering –Trees of clusters are formed known as dendogram. To merge or split subsets of points rather than individual points, the distance between individual points has to be generalized to the distance between subsets. Such derived proximity measure is called a linkage metric. It describes the concept of closeness and connectivity. Linear algebra methods, based on singular value decomposition (SVD) are used for this purpose in collaborative filtering and information retrieval . SVD application to hierarchical divisive clustering of document collections resulted in the PDDP (Principal Direction Divisive Partitioning) algorithm.

4.2.4 Partitioning Relocation Clustering Relocation schemes that iteratively reassign points between k-clusters. One approach to data partitioning is to take a conceptual point of view that identifies the cluster with a certain model whose unknown parameters have to be found. More specifically, probabilistic models assume that the data comes from a mixture of several populations whose distributions and priors we want to find.

4.2.5 Fuzzy Clustering Fuzzy clustering approaches, on the other hand, are non-exclusive, in the sense that each document can belong to more than one clusters. Fuzzy algorithms usually try to find the best clustering by optimizing a certain criterion function. The fact that a document can belong to more than one cluster is described by a membership function. The membership function calculates for each document a membership vector, in which the i-th element indicates the degree of membership of the document in the i-th cluster

Other approaches of clustering for Hypertext or semi-structured data.

4.2.6 Suffix Tree Clustering Suffix Tree Clustering (STC) is a linear time clustering algorithm that is based on identifying the phrases that are common to groups of documents. A phrase is an ordered sequence of one or more words. The base cluster is a set of documents that share a common phrase. This algorithm was developed by Zamir and Etzioni in 1998.It has three steps: Document cleaning, Identify base clusters using suffix tree and combining these base clusters into clusters. Firstly HTML tags are marked and stripped. Suffix tree is generated with linearity of time and size of document collection .Each node of tree represents cluster and suffix tree represents documents whose phrases are common. In the end to avoid any overlapping merging of base clusters is done.

4.2.7 Clustering based on Structure - Syntactic structure of Web pages represented by abstract tree syntax. Distance between two trees is determined by tree edit distance and pages are clustered if they are at minimum distance using agglomerative algorithm.

4.2.8 Clustering based on Connectivity- The hyperlinks connecting the Web pages are used as entity descriptions. When a group of pages is highly connected by hyperlinks, the related cohesion metric is given a high score. Clusters are determined which maximize "quality of clustering metric"

4.2.9 Clustering based on Keywords -The content of the Web pages is analyzed in order to determine a set of keywords that characterize each of them. Two pages are clustered together when they share a large number of keywords

4.2.10 C-LINK Clustering –Similar pairs generated as edges in graph and nodes represent URL's.This approach divides the clusters in such a way that they are close to Centre and all other nodes are close to enough. It is used to form Flat clusters

4.3) Associations

Finding the information from the collection of indexed documents using association rules. Its objective is to find all co-occurrence relationships, called associations, among data items. Let Wi and Wk be the set of keywords.

Support –The percentage of transactions which includes keywords and can be seen as estimate of probability.

$$\text{Support} = \frac{\text{Support count of } Wi \text{ U } Wk}{\text{No. of documents}}$$

Confidence – The percentage of transactions that contain Wi also contain Wk

$$\text{Confidence} = \frac{Wi \text{ U } Wk}{Wi \text{ . count}}$$

Association rule mining problem is broken into steps:

1) Generate all the keyword combinations keyword sets whose support is greater than the user specified minimum support (called minsup). Such sets are called the frequent keyword sets.

2) Use the identified frequent keyword sets to generate the rules that satisfy a user specified minimum confidence (called minconf). The frequent keywords generation requires more effort and the rule generation is straightforward.

These rules for hypertext identify groupings between set of items with some minimum specified confidence and support.

(4.3.1) Extended Concept Hierarchy – Concept Hierarchies define a sequence of mappings from low level concepts to higher level counterparts. It can be represented as Linked list and as a lattice or arbitrary graph structure. We generate rules that relate a structured attributes with concepts in ECH. User provides a structured component and minimum confidence and generation of the rules according to the neighbors that have minimum support and structured attribute values. Once the ECH is created, attribute values of different attributes in the database can be associated with different subsets of concepts in the ECH.It was applied to maintain parent, child and sibling relationships between concepts so that generated rules can create concepts in ECH.

(4.3.2) Direct Association Rules in the Web
Let d be the independent webpage and D website content .A set of Pages X is called page set and number of pages in it represents its length. The page set X is the body and Y is the head of the rule X →Y. It represents regularities discovered from large data set.

5) Database view of Web Page

Database consists of several interlinked relations where each represents certain objects and its relationships. Real life data which is interlinked through some linkages and that is stored in Relational Databases, XML files and other data repositories. The various algorithms can be applied on single table or an independent table. But in Multi-relational Databases it's not so easy. In this field we have some challenges. It is not easy to model such table which is interlinked to many objects and finding a such kind of hypothesis that fits the data. Database view tries to convert a website into some kind of database so that a sophisticated query retrieval should be there.

"DATABASE VIEW" means abstract layer that hides the web under a virtual database view. It is mainly concerned with two objectives:

(I) Schema Extraction from semi-structured view
(II) Query languages for semi-structured View

5.1) Schema Discovery - Its basic purpose is to find all of the patterns. It explores an object by navigating from one object to another and keeps tracks of the object references traversed using Object Model Exchange. It can be explained as follows:

Suppose we have a transaction database .Let T be some tree expression for transaction. Support of T is number of transactions tn.T will be the schema pattern if it is maximal frequent and it is not weaker than any other tree expression.
To compute various patterns in the increasing size of tree expressions. Containing k-occurrences of no schema information. It represents the path that is a sequence of alternating objects and labels.

5.2) Schema Extraction - It is the task of automatically discovering implicit structural information from semi-structured data. Web data belongs to semi-structured data class, i.e., data with self-contained schema.

Two approaches

We can extract schema by the occurrence of each object in the graph which is related to some concept name and concept term .The number of concepts can be reduced by grouping together equivalent concepts. But after this we have still large specific schema.

We can extract frequent substructures from graph. Frequency of substructures with respect to threshold value that is equal to the number of number of objects which can satisfy a structure and we have some relevant structures.

5.3) Query languages for the Web – Query languages give querying power to the databases. Standard query languages, such as SQL for relational databases and OQL for object databases, are too constraining for querying semi-structured data, because they require data to conform to a fixed schema before any data is stored into the database.

5.3.1) WebSQL proposes to model the web as a relational database composed of two (virtual) relations: Documents and Anchor. The document relation has one tuple for each document in the web and the Anchor relation has one tuple for each anchor in each document in the web.

5.3.2) WebOQL The main data structure provided by WebOQL is the hypertree. Hyptertrees are ordered arc labeled trees with two types of arcs, internal and external. Internal arcs are used to represent structured objects on a page and external arcs are used to represent hyper links. Sets of related hypertrees are collected into structures called "webs". Both hypertrees and webs could be manipulated through WebOQL.

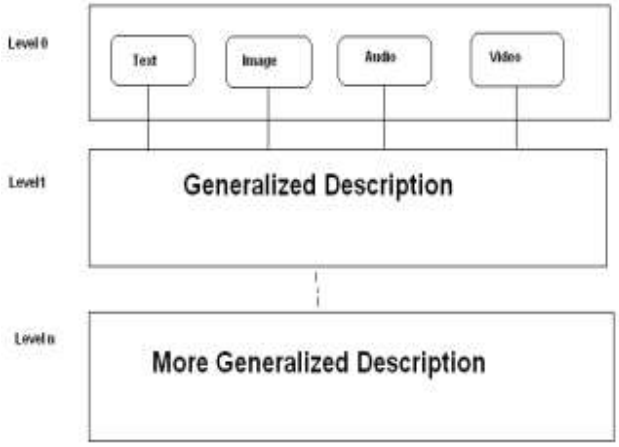
5.3.3) WebML. A weakness of both WebSQL and WebOQL is that both languages do not provide any means of knowledge discovery. WebML allows resource discovery as well as knowledge discovery from a subset or from web as a whole. WebML takes advantage of the MLDB (Multi Level Databases) model in which each layer is obtained by successive transformation and generalizations of the lower layers.

5.3.4) W3QL combines structure queries, based on the organization of hypertext documents, and content queries, based on information retrieval techniques.
5.3.5) Web Log Logic-based query language for restructuring extracted information from Web information sources.

5.3.5) Lorel and UnQL query heterogeneous and semi-structured information on the Web using a labeled graph data model

5.4) Multilevel Databases Several researchers have proposed a multilevel database approach to organizing

Web-based information. The main idea behind these proposals is that the lowest level of the database contains primitive semi-structured information stored in various Web repositories, such as hypertext documents. At the higher level(s) Meta data or generalizations are extracted from lower levels and organized in structured collections such as relational or object-oriented databases.



5.4.1) Multilevel Databases

Level 0 – Unstructured and massive information base.

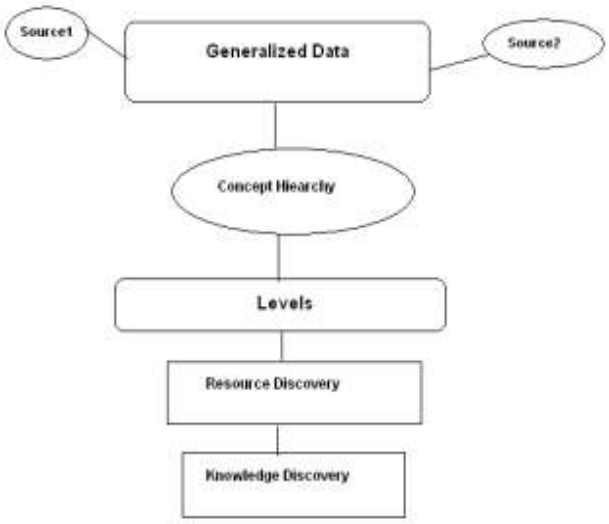
Level 1- Relatively structured and obtained by transformation and generalization

Level 2 – Further generalizations for better structured data

In a multilayered database each layer is obtained through generalization and transformation operations performed on the lower layers. Summarization of the large data sets related to a general description is known as generalization

5.4.1) Data Generalization which abstracts large data sets of relevant data in a database from lower level concept to higher ones.

5.4.2) Data Classification which finds the common properties among set of objects in a database and classifies them into different classes. Suppose we have training set data in which each tuple consists of same set of multiple attributes .Classification task is to analyze training data and develop an accurate description



5.4.2) Architecture of MLDB

6) Multimedia Mining

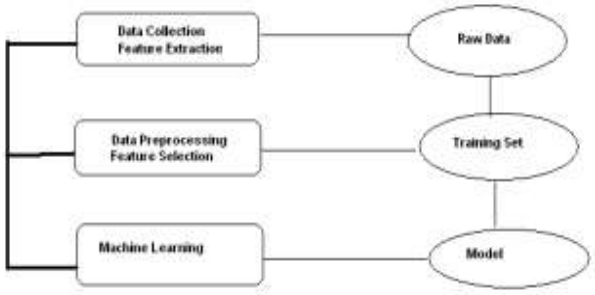
Multimedia Data essentially refers to data such as text, numeric, images, video, audio, graphical, temporal, relational and categorical data. Multimedia Mining is discovering knowledge from large amounts of different types of multimedia data. It involves the extraction of implicit knowledge, multimedia data relationships, or other patterns not explicitly stored in multimedia databases. Relevant Information and analysis is key aspect of this field. Recent growth of image databases has created a need for effective and efficient retrieval from large databases.

Generally, multimedia database systems store and manage a large collection of multimedia objects, such as image, video, audio and hypertext data. Thus, in multimedia documents, knowledge discovery deals with non-structured information. For this reason, we need tools for discovering relationships between objects or segments within multimedia document components, such as classifying images based on their content, extracting patterns in sound, categorizing speech and music, and recognizing and tracking objects in video streams Multimedia files undergo various operations to extract important features from it.

need tools for discovering relationships between objects or segments within multimedia document components, such as classifying images based on their content, extracting patterns in sound, categorizing speech and music, and recognizing and tracking objects in video streams Multimedia files undergo various operations to extract important features from it.

6.1) Multimedia Mining Process

Data collection is the starting point of a learning system, as the quality of raw data determines the overall achievable performance. Then, the goal of data pre-processing is to discover important features from raw data. Data pre-processing includes data cleaning, normalization, transformation, feature selection, etc. Learning can be straightforward, if informative features can be identified at pre-processing stage. The product of data pre-processing is the training set.



6.1.1 Multimedia Mining process

6.2) Feature Extraction

There are two kinds of features: description-based and content-based. The former uses metadata, such as keywords, caption, size and time of creation. The later is based on the content of the object itself. Image categorization classifies images into semantic databases that are manually precategorized. In the same semantic databases, images may have large variations with dissimilar visual descriptions. Three types of feature vectors for image description:
1) pixel level features
2) region level features
3) tile level features

Perceptual features such as loudness, brightness, pitch etc. to represent sound clips.Apart from These Total energy, Bandwidth, Pitch period are also used for audio classification. The first stage for mining raw video data is grouping input frames to a set of basic units, which are relevant to the structure of the video. In produced videos, the most widely used basic unit is a shot, which is defined as a collection of frames recorded from a single camera operation. Shot detection methods can be classified into many categories: pixel based, statistics based, transform based, feature based and histogram based.

6.3) Data Pre-processing

In a multimedia database, there are numerous objects that have many different dimensions of interests. Selecting a subset of features is a method for reducing the problem size [21]. This reduces the dimensionality of the data and enables learning algorithms to operate faster and more effectively. The problem of feature interaction can also be addressed by constructing new features from the basic features set. This technique is called feature construction/transformation

7) Conclusion

In this paper we reviewed representation issues, processes involved in we content mining. Changes in the nature of web poses new challenges in web mining tasks and how algorithms can be changed according to changes. Now Webpage is also seen as a database not only text. This paper describes multimedia content and how they can be mined and preprocessed .Mining is not only limited to text it's application areas are increasing day by day.Security,Sensor networks,Arificial Intelligence, Weather Forecasting are upcoming areas for implementing web mining.

10. References

[1]Arvind Arasu , Hector Garcia-Molina (2003) Extracting Structured Data fromWeb Pages , Proceedings of the 19th International Conference on Data Engineering
[2] Bernard J Jansen and Amanda Spink (2003) An Analysis of Web Documents Retrieved andViewed ,The 4th International Conference on Internet Computing. p. 65-69.
[3]Dr. Abdallah Alashqur (2008) Mining Association Rules : A Database Perspective. International J. Computer Science and Network Security 8(12): 69-73
[4]Hany Mahgoub, Dietmar Rösner, Nabil Ismail and Fawzy Torkey (2007) A Text Mining Technique Using Association Rules Extraction : International Journal of Computational Intelligene 4(1): 21-27
[5]Manisha Marathe, Dr. S.H.Patil, G.V.Garje,M.S.Bewoor (2009) Extracting Content Blocks from Web Pages. J. Recent Trends in Engineering 2(4): 62-64
[6]Mirela Pater, Daniela E. Popescu and Daniela Maştei (2008) Pattern discovery techniques in Web mining. J. Computer Science and Control Systems. 1(1): 77-81
[7]Przemyslaw Kazienko (2009) Mining indirect association rules for Web .J. Appl. Math. Comput. Sci 19 (1) :165-186
[8]Raymond Kosala and Hendrik Blockeel (July 2000) Web Mining Research: A Survey SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM 2(1): 2-9
[9]Samia Jones and Omprakash K. Gupta (2006) Web Data Mining : A Case Study . Communications of the IIMA 6(4) :59-62

[10] Zdravko markov,Data mining the Web,Wiley Series
[11] David L.Olson, Dursun Delen ,Advanced Data Mining Techniques,Springer.
[12]Soumen Chakarborti ,Mining the Web , Morgan Kaufmann Series in Data Management Systems
[13] Jackson, Peter, Natural language processing for online applications : text retrieval, extraction, and categorization / Peter Jackson, Isabelle Moulinier. p.cm. (Natural Language Processing, issn 1567–8202 ; v.5)
[14] Michael W. Berry ,Survey of Text Mining,Springer
[15] Adam Schenker ,Horst Bunke,Mark Last,Abraham Kandel,Graph Theoretic Techniques for Web Content Mining Series in Machine Perception and Artificial Intelligence — Vol. 62
[16] Bing Liu,Web DataMining, Springer
[17] Min Song and Yi-Fang Wu ,Handbook of research on text and web mining techologies ,Information Science Reference
[18] Hsinchun Chen and Michael Chau, Web Mining: Machine learning for Web Applications ,Ch -6, Annual Review of Information Science and Technology.
[19] Xindong Wu, Chengqi Zhang, Shichao Zhang (2005) Database classification for multi-database mining . Information systems 30(1): 71-88
[20]Zhao Li, Wee Keong Ng , Aixin Sun (2005) Web data extraction based on structural similarity, Knowledge and Information Systems 8(4): 438-461