IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014.
www.ijiset.com
ISSN 2348 – 7968

# A Survey on Web Content Mining Techniques and Tools

[1]Ms.S.Valarmathi,MCA.,M.Phil, [2]Mr.P.Purusothaman.MCA

[1]Asst. Professor, Department of Computer Science
[2]System Administrator, Library
[1,2]SAAS College (Autonomous) Trichy-5

## ABSTRACT:

This paper discus on the issues of web content mining. World Wide Web has rich source of voluminous and heterogeneous information which continues to expand in size and complexity.web content mining is the process of identifying user specific data from text, image, and audio (or) video data which already persist on the web. Much web content is structured like data in the tables, unstructured such as free texts and semi structured like HTML documents. This web content mining aims to mine useful information (or) knowledge from web page contents.

## KEYWORDS:

Web Mining, Web Content Mining, Decision Tree Learning, Text Mining, Web Content Mining Tools.

## 1. INTRODUCTION:

Web content mining is used for extracting useful information from web pages, content mining is related to semi structured and unstructured documents. Data mining and text mining techniques can be applied in web content mining. Unstructured web pages data's are easy to extract when compared with structured and semi structured data's.

The structure of this paper is as follows: Section 2 Presents the overview of web mining, web content mining techniques and tools section 3 deals with literature work and section 4 deals with conclusion.

## 2. WEB MINING:

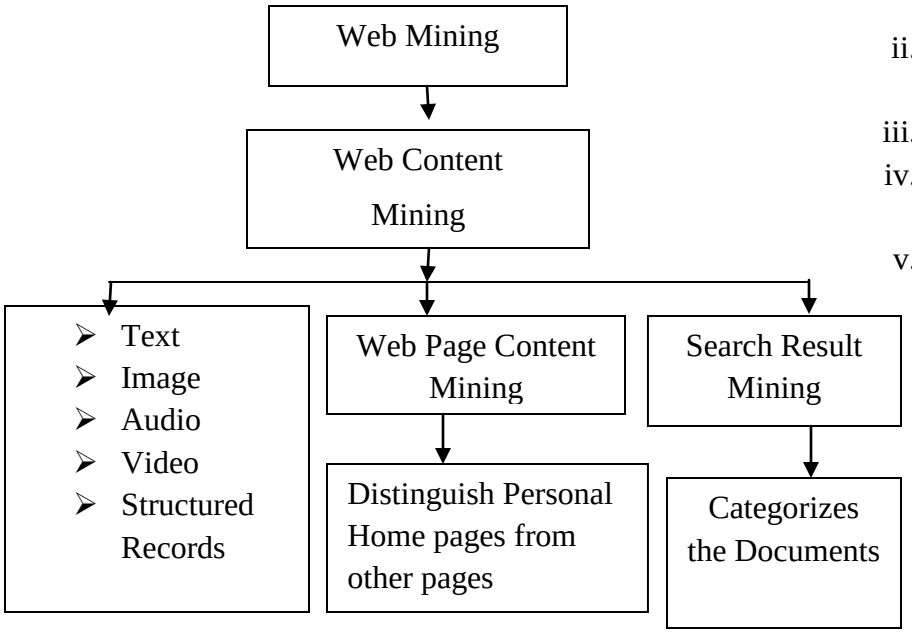
### 2.1 Overview:

Web Mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the web data. Web Mining is used to capture relevant information, rating new knowledge out of the relevant data, learning about the different types of users. Web mining uses data mining techniques to automatically discover and extract information from web documents and services. Several other techniques like information retrieval, information extraction and machine learning have been used in the past to discover the new knowledge from the huge amount of data available in the web. These techniques have been compared with web mining, Information retrieval works by indexing text and then select useful information. Web mining comprises of two systems like information retrieval system and information extraction system.

### 2.2 Web Content Mining:

Web content mining is the process of extracting useful information, from the contents of web documents. Content of web documents may consist of text, image, audio, and video (or) structured records such as lists and tables. Web content mining is related to data mining because many data mining techniques can be applied in web content mining. It is also related to text mining because much of web contents are text based. It is different from these because web data is semi structured in nature and text mining focuses on unstructured text.

ISSN 2348 – 7968




Taxonomy of web content mining

### 2.2.1: Unstructured Data Mining Techniques:

Content mining can be done on unstructured data such as text. Mining of unstructured data gives unknown information. Text Mining is extraction of previously unknown information extracting information from different text sources. Content mining requires application of data mining and text mining techniques

## 3. LITERATURE WORK:

Text is a kind of data in which the word attributes are sparse, and high dimensional with less frequencies on most of the words. Hence, it is difficult to apply the classification methods to the text. The methods which are commonly used for text classification are follows:

i. Decision Tree Classification
ii. Neural Networks Classification
iii. Naïve Baye's Classification
iv. SVM (Support Vector Method) Classification.
v. Association Rule Based Classification

**Decision Tree Classification:**

A decision Tree is used for decision making purpose. Decision tree has root and branch node. From the root node users split each node recursively based on decision tree reasoning algorithm. The final result of decision tree consists of branches and each branch represents a possible scenario of decision and its consequences. C4.5 is a decision tree algorithm which comprises of two branches coming out of the node. The algorithm checks every partition and chooses the best one. The complexity of the algorithm is reduced by pruning techniques (i.e.) cutting the branches. It also decides about attribute selection and classification with the lowest amount of information needed. Finally, the algorithm chooses optimal split, the one with the biggest information gain.

**Pseudo Code of C4.5 algorithm:**

- ✓ Initially check for base cases
- ✓ For each attribute "a"

i. By splitting the attribute a, the normalized information gain is found.
ii. Let a_best be the attribute with the highest normalized information gain.
iii. Then create a decision node which splits on a_best.
iv. Recurse on the sub list obtained by splitting on a_best and add those nodes as children of node.

**Neural Network Classification**:

Neural networks are used in a wide variety of domains for the purposes of classification. Neural networks classifier is implemented in the text data with the use of word features.

The basic unit in a neural network is a neuron (or) unit. Each unit receives a set of inputs which is denoted by the vector $\overline{X_n}$, which relates to the item frequencies in the nth document. Each neuron is associated with a set of weights A which is used to compute the function. In neural network the linear function is calculated as follows.

$$P_n = A \cdot \overline{X_n}$$

**Naïve – Baye's Classification:**

Naïve Baye's classifier also termed as generative classifier models the distribution of the documents in each class with a probabilistic approach. Two classes of models are commonly used in this method. Both models focus on posterior probability of a class, based on the distribution of the words in the document. These models work with the "bag of words" but ignore the actual position of the words. The main difference in these two models is based on the word frequencies.

**Support Vector Machine Classification:**

Support Vector Machine technique is accurate learning method for classification problem. It tries to find an optimal hyper Plane within the input space in order to classify the text data in the binary form. the linear separable space for the hyper plane is given by

$$W \cdot X + b = 0$$

Where X stands for the arbitrary object to be classified, the vector $W$ and constant b are learned from a training set of linearly separable data SVM separates the positive and negative training example with a maximum margin.

## Association Rule Based Classification:

Classification based on association integrates classification and association rule mining. Classification association rules (CAR) are association rules with the class on the right hand side of the rules and conditions on the left side of the rules. These rules are extracted from the available training data and an accurate "association classification model" is built.

In text classification, classification based association is used to classify text documents into topic hierarchies. Rules are extracted using Apriori Algorithm. More than one the final class will be derived from the association rule. A new association classification method called Classification Multiple Association Rule (CMAR) which performs Classification based on Multiple

Association Rules. CMAR consist of two phases: rule generation and classification. In the first phase, rule generation, CMAR computes the complete set of rules in the form of R: P -> C, where P is a pattern in the training data set, and c is class labels such that sup (R) and conf ( R ) pass the given support and confidence thresholds, respectively. Furthermore, CMAR prunes some rule and only selects a subset of high quality rules for classification. In the second Phase, classification, for a give data object and predicts the class label of the object by analyzing this subset of rules. In this section, we develop method to generate rules for classification; CMAR first mines the training data set to find the complete set to find the complete set of rules passing certain support and confidence threshold. This is a typical frequent pattern or association rule mining task. To make mining highly scalable and efficient, CMAR adopts a variant of FP-growth method. FP-growth is a frequent pattern mining algorithm which is faster than conventional Apriori-like methods, especially in the situation where there exist large data sets, low support threshold, and/or long patterns.

## 4. CONCLUSION:

This paper discusses the techniques of web content mining. Text classification, the assignment of text documents with predefined classes based on their content is an essential component in many management tasks.

Compared to all the classifiers, Decision Tree learning performs well. It works efficiently based on greedy algorithm. C4.5 algorithm adopts the pruning techniques and the list attributes are selected on the basis of heuristic (or) statistical measure. It classifies data set's attributes with reasonable speed. This classification could be recommended as it supports faster learning speed, ease of use and accuracy with other methods.

***References:***

1. Kosla,R.and Blockeel,H.2000.Web Mining Research: A Survey. SIG KDD Exporations.Vol.2, 1-15.
2. MitChell,T.1997, Machine Learning McGraw Hill.
3. Han J, Kamber M, "Data Mining Concepts and Techniques", Second edition, Morgan Kaufmann Publishers, 2006.Pp. 628-648.[Accessed on Feb.18,2013]
4. Ajoudanian, S. and Jazi,M.D. 2009. Deep Web Content Mining. World Academy of Science, Engineering and Technology 49.
5. http://en.wikipedia.org/wiki/Decision_tree_learning
6. http:en.wikipedia.org/wiki/C4.5_algorithm
7. Wen Zhang,Taketoshi Yoshida and Xijin Tang, 2008, "Text classification based on multi-word with support vector machine". Journal Knowledge Based Systems-KBS . vol 21, no. 8, pp.879-886.
8. Christoph Goller,Joachim Loning,Thilo Will and Werner Wolff, 2009,"Automatic Document Classification: A thorough Evaluation of various Methods."
9. Steve R.Gun,1998. "Support Vector Machines for Classification and Regression", University of Southampton.