

A Study on Web Content Mining Algorithms And Techniques

RajKumar.R¹, Rahul R.², AfsarAhmed.A³

^{1,2,3} Dept of Computer Application

^{1,2,3} Sri Krishna Arts and Science College, Coimbatore

Abstract- Data Mining pertains to the study of observational datasets to find relationships and to precis the data. Data mining is a World Wide Web process. Web mining is a part of data mining which relates to various research communities such as information revival, Artificial intelligence and database management systems. There Web mining has its three categories web content mining, web structure mining, web usage mining. This study attention on various web content mining types, Algorithms and techniques.

Keywords- Data mining, Web mining, Web content mining

I. INTRODUCTION

The previous decade experienced a dramatic development of computer technology, such that with the press of finger information about the particular topic appeared in monitors within seconds. As time passed by, a complexity of web increased due to enormously large amount of data. So extraction of data according to users need became a tedious task. As a result mining became an essential technique to extract valuable information from internet. And this technique was named as web mining [1]. Web mining has three categories web content mining, web structure mining, web usage mining. This paper we mainly concentrate on Web content mining. web content mining. Is the process of extracting knowledge from the content of documents or their descriptions [2]. Web Content Mining comprises of excavating structured data, semi structured data or non structured data.[3]

The structure of this paper is as follows: Section 2 presents the overview of web mining and categories, section 3 deals with web content mining Algorithms and techniques.

II. WEB MINING OVERVIEW

Discovering useful patterns or knowledge from data repositories such as in the form of databases, texts, images, the Web, etc .The data repositories should be valid, potentially useful, and understandable[8]. Web is a popular and interactive medium with intense amount of data freely available for users to access. It is a collection of documents, text files, audio, video and other multimedia data [3]. Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc., i.e. Web Content means the content may be text, audio, video, structure records etc., Web Structure means the structure data like hyperlink, document structure and Web Usage is used to web server, application logs and application level logs data [4].

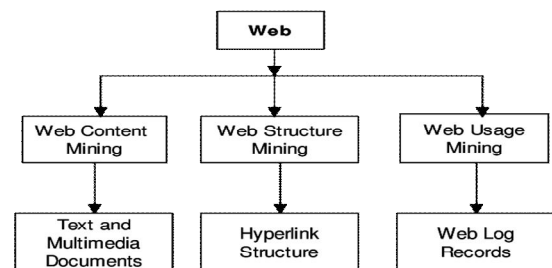


fig:1.web mining categories

2. A1. Web Mining Tasks:

Web mining consists of different essential tasks, which are described in a fig.2. Below.

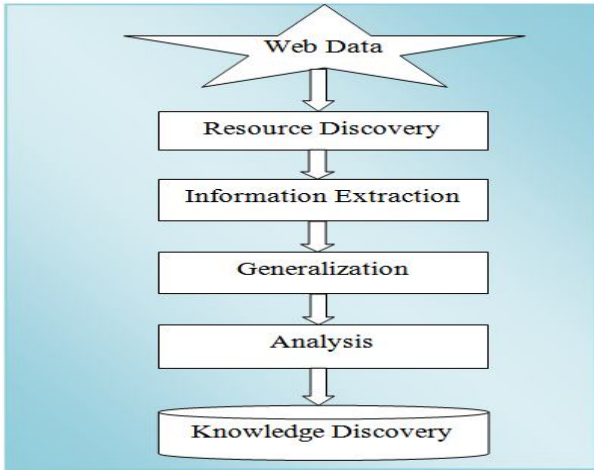


Fig.2.webmining tasks

1 Information Retrieval

It is the task of retrieving the intended information from the Web. It locates the unfamiliar documents and services on the Web.

2 Pre-processing

It is the task of automatically selecting and pre-processing specific information from retrieved Web resources.

3. Pattern Recognition & Machine Learning

It is the task to automatically discover general patterns of individual Web sites as well as across multiple sites.

4. Analysis

It is the task of analyzing, validating and interpreting the mined patterns.

III. 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 different from these because web data is semi structured in nature and text mining focuses on unstructured text.[2]. Fig.3 describes the taxonomy of web content mining.

3.A.Web content mining algorithms

There are two common tasks involved in web mining through which useful information can be mined. They are Clustering and Classification. Here various classification Algorithms used to fetch the information are described.

1. Decision Tree: [8]:The decision tree is one of the powerful classification techniques. Decision trees take the input as its Features and output as decision, which denotes the class information. Two widely known algorithms for building decision trees are Classification and Regression Trees and ID3/C4.5. The tree tries to infer a split of the training data based on the values of the available features to produce a good Generalization. This split at each node is based on the feature that gives the maximum information gain. Each leaf node Corresponds to a class label. The leaf node reached is considered the class label for that example. The algorithm can naturally handle binary or multiclass classification problems. The leaf nodes can refer to either of the K classes concerned.

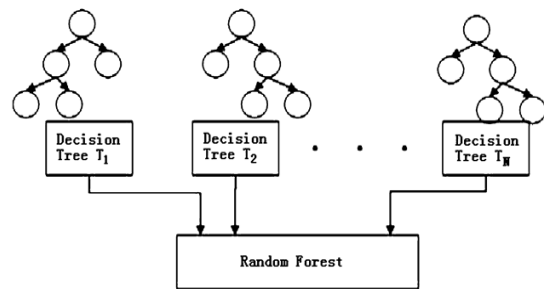


Fig11.decision tree algorithm

2. k-Nearest Neighbor [8]:KNN is considered among the oldest nonparametric classification algorithms. To classify an unknown example, the distance (using some distance measure e.g. Euclidean) from that example to every other training example is measured. The k smallest distances are identified, and the most represented class in these k classes is considered the output class label. The value of k is normally determined using a validation set or using cross-validation.

K-NEAREST NEIGHBOR

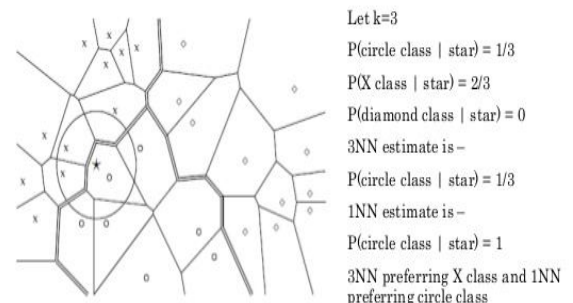


Fig12.k-nearest neighbor algorithm

3. Neural Network [8]:The most popular neural network algorithm is back propagation which performs learning on a multilayer feed forward neural network. It contains an input layer, one or more hidden layers and an output layer. The basic unit in a neural network is a neuron or unit. The inputs to the network correspond to the attributes measured for each training tuple. The inputs fed simultaneously into the units making up the input layer. It will be weighted and fed simultaneously to a hidden layer. Number of hidden layers is arbitrary, although usually only one. Weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction. As network is feed-forward in that none of the weights cycles back to an input unit or to an output unit of a previous layer.

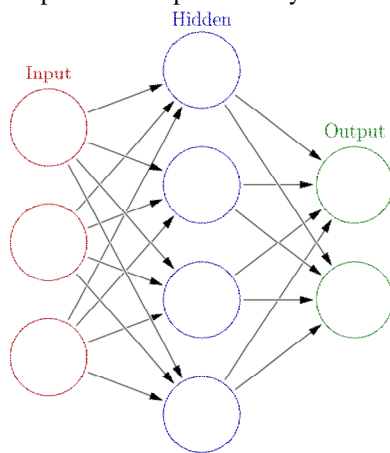


fig:15:neural network algorithm

4.Comparison of algorithm

Table2. Shows in this flowing table we present some of the web content mining algorithm. problem ,algorithm and applicability


Problem	Algorithm	Applicability
Classification 	Logistic Regression (GLM)	Classical statistical technique
	Decision Trees	Popular / Rules / transparency
	Naive Bayes	Embedded app
	Support Vector Machine	Wide / narrow data / text

Table2:comparison of algorithm

3.B.Web content mining techniques

It identifies the useful information from the web contents/data/documents, however, such a data in its broader form has to be further narrowed down to useful information. Web content data consist of structured data such as data in the tables, unstructured data such as free texts, and semi-

structured data such as HTML documents.unstructured, structured, semi structured and multimedia data.

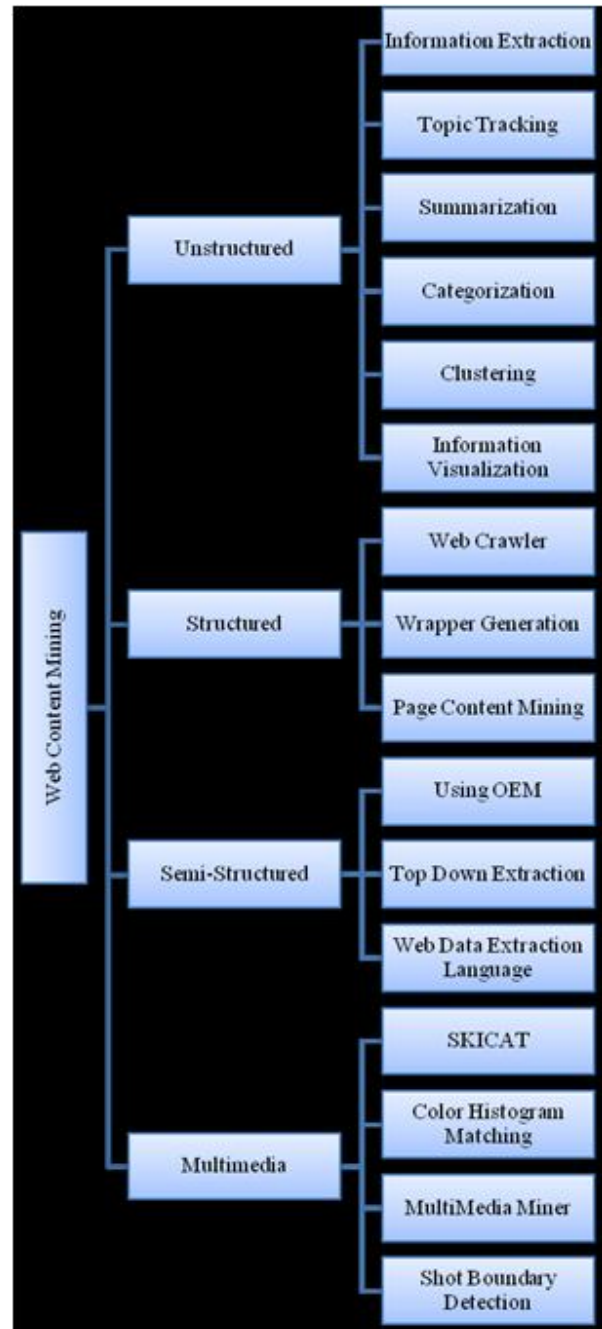


Fig.12. Taxonomy of web content mining

3.B.1. Unstructured Data Mining Techniques:

1. Information Extraction:

pattern matching is used to extract information from unstructured data[4]. It traces out the keyword and phrases and then finds out the connection of the

keywords within the text. This technique is very useful when there is large volume of text.

2. Topic Tracking

In Topic Tracking applied by yahoo, user can give a keyword and if anything related to the keyword pops up then it will be informed to the user. Same can be applied in the case of mining unstructured data.

3. Categorization

This technique counts the number of words in a document. It decides the main topic from the counts. It ranks the document according to the topics. Documents having majority content on a particular topic are ranked first. Categorization can be used in business and industries to provide customer support.

4. Clustering:

Clustering is a technique used to group similar documents. Some documents can appear indifferent group. Clustering helps the user to easily select the topic of interest.

5. Information Visualization

It utilizes feature extraction and key term indexing to build a graphical representation. The documents having similarity are determined using Information Visualization. Large textual materials are represented as visual hierarchy or maps where browsing facility is allowed. It helps the user to visually analyze the contents. User can interact with the graph by zooming, creating sub maps and scaling. This technique is very useful to find out related topic from a very large amount of documents [5].

3.B.2. Structured Data Mining Techniques:

1. Web Crawler:

Crawlers computer programs that are traverse the hypertext structure in the web.

2. Wrapper Generation:

In Wrapper Generation, it provides information on the capability of sources. The sources are what query they will answer and the output types. The wrappers will also provide a variety of Meta information. E.g. Domains, statistics, index look up about the sources. Page Content Mining: Page Content Mining is structured data extraction technique which works on the pages ranked by traditional search engines.

3. Web Page Content Mining

Web page content mining aims to extract/mine useful information or knowledge from a web page contents. By comparing page Content rank it classifies the pages. It identifies information within web page and distinguishes home page from other pages. Web page content mining uses two types of approaches: Database approach and Agent based approach. The first approach aims on modeling the data on the Web into more structured form in order to apply standard database querying mechanism and data mining applications to analyze it. The second approach aims on improving the information finding and filtering.

3.B.3. Semi-Structured Data Mining Techniques:

1. Object Exchange Model (OEM):

A main feature of object exchange model is self describing; there is no need to describe in advance the structure of an object.

2. Top down Extraction:

It extracts complex objects from a set of rich web sources and converts into less complex objects until atomic objects have been extracted.

3. Web Data Extraction Language

Web data extraction language converts web data to structured data and delivers to end users

3.B.4. Multimedia Data Mining Techniques

Some of the Multimedia Data Mining Techniques are SKICAT, Multimedia Miner, Color Histogram Matching and Shot Boundary Detection.

1. SKICAT

SKICAT is a Successful Astronomical Data Analysis and Cataloging System that produces digital catalog of sky object. It uses machine learning technique to convert these objects to human usable classes. It integrates technique for image processing and data classification which helps to classify very large classification set [2].

2. Color Histogram Matching

The paper[9] that Histogram matching is a method in image processing of color adjustment of two images using the image histograms. It is possible to use histogram matching to balance detector responses as a relative detector calibration technique. It can be used to normalize two images, when the images were acquired at the same local illumination (such as shadows) over the same location, but by different sensors, atmospheric conditions or global illumination.

3. Multimedia Miner

Multimedia Miner is a prototype of a data mining system for mining high-level multimedia information and knowledge from large multimedia databases. It includes the construction of multimedia data cubes which facilitate multiple dimensional analysis of multimedia data, and the mining of multiple kinds of knowledge, including summarization, classification, and association, in image and video databases.

4. Shot Boundary Detection

Shot transition detection (or shot boundary detection) also called cut detection or shot detection, is a field of research of video processing. Its subject is the automated detection of transitions between shots in digital video with the purpose of temporal segmentation of videos. It is used to split up a film into basic temporal units called shots; a shot is a series of interrelated consecutive pictures taken contiguously by a single

camera and representing a continuous action in time and space. Shot detection methods can be classified into many categories: pixel based, statistics based, transform based, feature based and histogram based [9].

IV. CONCLUSION

As the Web has become a major source of information, techniques and methodologies to extract quality information is of paramount importance for many Web applications and users. Web mining and knowledge discovery play key roles in many of today's prominent Web applications such as e-commerce and computer security. This paper declares various Algorithms and techniques of web content mining.

REFERENCES

- [1] Faustina Johnson, Santosh Kumar Gupta, "Web Content Mining Techniques", A Survey International journal of computer applications(0975-888), June 2012.
- [2] ms.s.Valarmathi,mca., m.phil, mr.P.Purusothaman .mca "A Survey on web content mining techniques and tools" ijiset issn 2348 – 7968 august 2014.
- [3] ShipraSainiHari Mohan Pandey, " Review on Web Content Mining Techniques International Journal of Computer Applications" (0975 – 8887) Volume 118 – No. 18, May 2015
- [4] T.Shanmugapriya1, P. Kiruthika2," Survey on Web Content Mining" and Its Tools International Journal of Scientific Engineering and Research (IJSER) www.ijser.inISSN (Online): 2347-3878 Volume 2 Issue 8, August 2014
- [5] Arvind Kumar Sharma1, P.C. Gupta2,"Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining"
- [6] M.Karpaga R.Sasikala, " Analysis of Web Content Mining Tools "International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Availableonline at: www.ijarcsse.com
- [7] IR.Malarvizhi,"WebContentMiningTechniquesTools & Algorithms AComprehensiveStudy"2K. Saraswathi International Journal of Computer Trends andTechnology (IJCTT) – volume 4 Issue 8– August

2013ISSN:22312803<http://www.ijcttjournal.org>Page
2940

- [8] Shyam Nandan Kumar “World towards Advance Web Mining” A ReviewM.Tech-Computer Science and Engineering, Lakshmi Narain College of Technology-Indore (RGPV, Bhopal), MP, India