ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

Web Page Classification in Web Mining Research - A Survey

E.Suganya[1], Dr.S.Vijayarani[2]

Ph.D Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, India[1]

Assistant Professor, Department of Computer Science, Bharathiar University, Coimbatore, India [2]

ABSTRACT: The web is a collection of documents which contains more information like textual content, images, audio, video, etc. A web page is a document which can be suitable for web browsers like Mozila Firefox, Google Chrome, Microsoft Internet Explorer, etc. A website is a collection of web pages which are grouped and connected together in various ways. At present, the numbers of web pages on the World Wide Web are increasing significantly. Many search engines like Google, Yahoo, and Bing are available to users to find the specified web page. Generally the search engines work through keyword inputs. However, web pages are retrieved in this manner usually include invalid links and irrelevant web pages. Hence, there is a need for categorizing web pages or web documents to facilitate the indexing, searching and retrieving of web pages and it also useful for finding its category. Classification plays a vital role in many information retrieval tasks. Web page classification is an essential for crawling, the development of web directories, topic-specific web link analysis, contextual advertising, and analyzing of the topical structure of the Web. It can also help to improve the quality of Web search. The main problem in web mining is to classify the web pages and it can be done using optimization algorithms, classification algorithms, feature selection and feature extraction algorithms. Web mining techniques are used to fetch knowledge from web data. Web mining is used to extract relevant information from web pages or web documents. Detailed study and survey have been conducted and this paper provides the basic concepts of web page classification, different algorithms and techniques used for web page classification, web directories, types, approaches, applications, features and research issues.

KEYWORDS: Web Mining, Web Page Classification, Web Directories, Research Issues

I. INTRODUCTION

The fast development on the networking technologies have increased the popularity of the web which contains large amount of information. Web contents of the web page include online documents, e-books, articles, technical reports, digital libraries which have been rapidly exploring all time. Categorizing the web pages is very useful for efficient contents browsing, managing and spam filtering. This process is also difficult for search engine to identify a particular web page [3]. The web pages are retrieved based on the content and structure of a web page using web mining techniques. Web mining is used to extract useful patterns from web pages. It can be classified into three types, namely content mining, structure mining and usage mining. Web content mining is used to extract the web information based on content which includes text, images, audio, video, etc. Web structure mining is used to extract the information through hyperlinks. Web usage mining is used to find the behavior of online users. In web page classification, the information is extracted based on content, links and usage of web data using web mining techniques. It is a process of categorizing the web pages with meaningful predefined category labels, for example news, sports, business, etc. It is used to search, manage and retrieve the web data. A web page belongs to any of the above category. Let C = {c_1, c_2, …c_k} be a set of predefined categories, D = {d_1, d_2,….d_n}be a set of web pages to be classified and A= D×C be a decision matrix:

Web Pages	**Categories**			
	c_1	c_2	…	c_k
d_1	a_{11}	a_{12}	…	a_{1k}
d_2	a_{21}	a_{22}	…	a_{2k}
...	…	…	…	…
d_n	a_{n1}	a_{n2}	…	a_{nk}

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

A web page has some significant characteristics, first one is web pages are semi-structured documents that are usually written in HTML. Second one is web pages are connected to each other using hyperlinks. Third, web pages are often short, hence it may be insufficient to analyze. Finally the sources of the web pages are frequently changing, heterogeneous, distributed and dynamically changing [8].

Web page classification can be divided into multiple sub problems, namely the subject based classification, function based classification, sentiment classification and other types of classification. In subject based classification, web pages are classified based on their contents or topic of a web page hence it is also known as topic based classification [1]. For example Science domain used Yahoo.com such as Agriculture, Astronomy, Biology, Chemistry, Cognitive Science, Complex Systems and Computer Science. The subject based classification can be applied to build topic orders of web pages related to specific subjects. In functional based classification, web pages are categorized based on functional factors such as a personal home page, course page, admission page and etc. This type of classification is called style based classification [4]. Sentiment classification is based on sentimental polarities of opinions, for example, favorable or unfavorable, positive or negative.

II. LITERATURE SURVEY

In the literature review, the authors discussed and implemented number of techniques, approaches and algorithms such as machine learning algorithms, optimization algorithms, classification algorithms and feature reduction algorithm etc. these information are discussed in the table given below.

Authors Name	Title	Description	Techniques	Inference
Poonam Asawara, Dr Amit Shrivastava and Dr Manish Manoria – 2017	Hybrid Approach for Web Page Classification Based on Firefly and Ant Colony Optimization	The authors surveyed a detailed review of useful web-specific features for classification. Major applications of web page classification and future research directions are discussed.	Random Forest classification, Ant Colony Optimization algorithm and Firefly algorithm.	The authors concluded that the hybrid approach of Firefly and Ant Colony optimization algorithm gives the better result than random forest classification algorithm [13].
P.Kameshwari, E.Salini Varma 2017	Web Page Classification Approach	The authors discussed a method that Cross Training based Corrective approach (CTC) is a new method for web page classification that acquires data from the test set in order to fix primarily assigned by a text classifier on that test set.	Cross training based corrective approach (CTC), Support Vector Machine, Naive Bayes and K-Nearest Neighbors (KNN).	The authors collected the dataset from Open Directory Project (ODP). It containing around 5 million webpages and is controlled into 765,282 groups and subgroups. The four binary classification tasks are "Male" vs. "Other" (1700 web pages), "Female" vs. [11]. Based on the observation the proposed CTC approach gives better result than SVM, Naïve Bayes and KNN.
Dr.Tamer Anwar Ahmed Alzohairy, Dr. Mohamed Taha Abu-Kresha and Ahmed Nagy Ramadan Bakry – 2017	An efficient method for web page classification based on text.	In this paper, the authors applied a different feature extraction technique with different web page classification methods to find an efficient method for web page classification. Each web page is represented by the three feature extraction methods. The principal component analysis (PCA) is used to select the most relevant features for the classification as the number of unique words in the collection of the set is big. Then the final output of	The three feature extraction methods used in the study are Term Occurrence, Term Frequency, Term Frequency-Inverse Document Frequency and the three classifiers used in the study are K-nearest neighbor (KNN), Naive Bayes (NB) and Decision Tree (DT).	The authors compared the classification algorithms, the results shows that Naive Bayes is the best compared to other classification methods. Term Frequency Inverse Document Frequency made an improvement to all classifiers used. Naive Bayes classification method with Term Frequency-Inverse Document Frequency is proposed to handle this domain of problems. In future research they will improve the accuracy of the proposed classifier. Studying web page classification problem with different features like photos, links, etc. Applying

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

		the PCA is sent to the three different classification methods to find the best method for web page classification.		the proposed methods in other applications and domains [12].
Zinah A. Abbas, Prof .Dr.Shawkat K. Guirgui, Dr .LaithR. Fleih, and Magda M. Madbouly – 2016	Web Pages Classification Using Rule-Based System	In this paper a rule based system is used to contract the system classifier for solving Web online classification by assigning each scanned HTML document to their class. The aim of this work is to design and implement an HTML document classification system that is able to classify the HTML documents according to their class (category).	Association rule mining, Naïve Bayes, SVM, Logistic Regression, decision trees and the Flexi Rank algorithm.	The authors developed web online classification. HTML documents are classified through online based on rule based system. [14].
Deepshikha Patel, Dr. Prashant Kumar Singh – 2013	A Web Page Classification Technique with Textual Content Analysis Using NN-PCA for Objectionable Web Page Classification	The authors proposed a new content based classification scheme for objectionable web documents filtering using neural network with input obtained by Principal Component Analysis (PCA). Each web page is represented by the term weighting scheme. They have conducted results by taking four datasets containing different web pages to test the performance of a classifier in various scenarios. The experimental evaluation demonstrates that the proposed method provides outstanding classification accuracy for objectionable document classification.	Machine Learning algorithm, PCA feature reduction algorithm	In this paper, the authors designed and implemented a web page filtering system for objectionable web documents. They proposed OWPCM for classifying objectionable and non-objectionable web documents. This system was experimented using the NN machine learning algorithm, entropy term weighting indexing algorithm, and PCA feature reduction algorithm. They have taken four different types of data sets. The experimental results have achieved great accuracy for classifying web documents [15].
Ajay S. Patil, B.V. Pawar - 2012	Automated Classification of Web Sites using Naive Bayesian Algorithm	In this paper the authors have attempted to classify web sites based on the content of their home pages using the Naïve Bayesian machine learning algorithm.	Naïve Bayesian machine learning algorithm	The Naive Bayes approach for classification of home pages for the ten categories such as Academic institution, hotels, book sellers/publishers, health care, sports, automobiles, tours and travels, computer, banking and domestic appliances are considered which yielded 89.05% accuracy. However, in this results, only distinct and non-hierarchical categories are considered [16].

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

J. Alamelu Mangai, Dipti D. Kothari and V. Santhosh Kumar – 2012	A Novel Approach for Automatic Web Page Classification using Feature Intervals	The authors proposed a new web page classification algorithm WVFI (Weighted Voting of Feature Intervals). Experiments done on a benchmarking data set called WebKB has shown good classification accuracy when compared with many of the existing classifiers.	Supervised Discretization algorithm	In this paper a new web page classifier which combines the predictions made by all features is proposed. Each feature votes for a particular class using the class distribution of its intervals. From the results it is inferred that the performance of this classifier is good when compared with other existing web page classifiers. The future work is to run this classifier for multiple categories of web pages [17].
J. Alamelu Mangai, V. Santhosh Kumar – 2011	A Novel Approach for Web Page Classification using Optimum features	This paper suggested a method to classify a Web page with only minimum number of representative features or terms extracted from it without using the entire Web page. The optimum number of features is selected using a three step procedure, by filtering the features in each subsequent step. The experimental results have shown good classification accuracy with these optimum features.	Machine learning classifiers such as decision tree, KNN, OneR, Multilayer perceptron and RBF.	The results shown that the proposed method of feature selection helps to Identify a less number of features, but with a high information gain, which contribute more to the classification accuracy [18].

III. WEB DIRECTORIES

A web directory is a listing of websites organized in a hierarchy of interconnected list of categories and subdirectories which is usually maintained by humans. A Web directory is also known as link directory. Unlike, the search engine database automatically gathered the links or entries by web crawlers, most of the web directories are created and maintained manually. There are two ways to find information on the web namely searching or browsing. Web directories provide links in a structure form which leads to make browsing easier. With the help of the search engine, many web directories combine searching and browsing to search the directory. The datasets are collected from the web directory for classifying the web pages. Figure 1 shows the web directory [10].

Figure 1: The Web Directory

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

Most popular web directories are Yahoo Directory, DMON (from directory.mozila.org), Best of the web, the World Wide Web Virtual Library, Martindale's the Reference Desk, Jasmine Directory, Hotfrog, Incrawler, Family Friendly Sides, Enquira and etc.

IV. TYPES OF WEB PAGE CLASSIFICATION

Based on the number of classes in the problem, classification can be divided into binary classification and multiclass classification which is divided into single label-hard classification, multi label-hard classification and soft classification [11]. Figure 2 shows the types of classification.

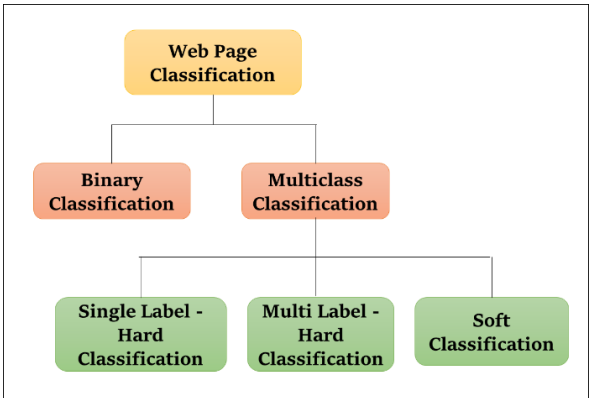

Figure 2: Types of Classification

Binary classification categorizes the instances into exactly two classes, namely commercial and non-commercial. Multiclass classification deals with more than two classes such as single label and multi label classification. In single label classification, one and only one class is to be assigned to each instance, while in multi-label classification, more than one class can be assigned to an instance [10]. For example, four class classification, i.e. four classes are involved, such as Arts, Business, Education and Science. Based on the type of a class assignment, classification can be divided into hard classification and soft classification. In hard classification, an instance can either be or is not being in a particular class, without an intermediate state, while in soft classification, an instance can be predicted to be in the same class with some probability.

V. TECHNIQUES FOR WEB PAGE CATEGORIZATION

There are several categories are used to classify the web pages. Commonly used techniques are manual categorization, clustering approaches, META tags based categorization, content based categorization and structure or link based categorization.

5.1 Manual Categorization

The traditional manual approach involved the analysis and classification of the contents of the web page by a number of domain experts and this type of classification is based on the textual content. But this approach is inappropriate in case of web page classification because vast number of web pages available on the web. Manual categorization is too slow to keep a list up to date. New documents are published, old ones are either removed or updated, new categories developed, and old ones fade away new meanings. Keeping up-to-date of this evolution by manual is practically impossible. Moreover, such a classification would be subjective.

5.2 Clustering Approaches

Clustering algorithms have been used to form clusters of related web pages to make classification easier and faster. The web pages are categorized using clustering algorithms like K-Means, etc. The similar web pages are grouped together and extracting information from them [9]. Moreover, clustering approaches are computationally more expensive.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

5.3 Meta Tags Based Categorization

In this classification technique is based on the attributes of the META tags like <META name="keywords"> and <META name="description">. However, there is a possibility of the web page author to include keywords which do not reflect the content of the page. The disadvantage of this method is that web page vendors may specify keywords that are irrelevant to the contents of their web pages just to increase the hit ratio of their own pages.

5.4 Content Based Categorization

In text content based categorization, to analyze the frequency of the occurrence of words in a large collection of text [6]. The stop words, i.e. the commonly occurring words are removed from web pages. The remaining words are the keywords which is represented as a feature vector. The document is then classified into an appropriate category using classification algorithms like the K-Nearest Neighbor. These text based algorithms do not make use of other relevant features like the structure of the document, images etc. for classification.

5.5 Structure Based (or) Link Based

The structure based categorization is an automatic web page classification technique based on the fact. The web page refers to a document must contain enough hints about its content to induce someone is to read it. Such hints given by the linking page can be used to classify the document being referred.

VI. APPLICATIONS OF WEB CLASSIFICATION

There are several applications for web page classification. They are:

- Constructing, maintaining or expanding web directories (web hierarchies)
- Improving quality of search results
- Helping question answering systems
- Building efficient crawlers or vertical search engines
- Web content filtering
- Assisted web browsing
- Knowledge based construction and etc.

VII. FEATURES

Web pages contain additional information, such as HTML tags, hyperlinks, and anchor text, other than the textual content visible in a Web browser [5]. These features can be divided into two broad classes: on-page features, which are directly located on the page to be classified, and features of neighbors, which are found on the pages related in some way to the page to be classified.

7.1 Using on Page Feature

a. Textual Content and Tags

N-gram representation is an approach, each document is represented by a vector of features, which includes not only single terms, but also up to five consecutive words. The advantage of using n-gram representation is that it is able to capture the concepts expressed by a sequence of terms (phrases), which are unlikely to be characterized using single terms. For example, one document contains the phrase Computer Science. The other contains the terms computer and science, but the two terms appear far apart [1]. The N-gram approach has a significant drawback: it usually generates a space with much higher dimensionality than the bag-of-words representation does. Therefore, it is usually performed in combination with feature selection.

b. Visual Analysis

Each Web page has two representations, if not more. One is the text representation written in HTML. The other is the visual representation render by a Web browser. They provide different views of a page [11]. Most approaches focus on the textual representation while ignoring the visual information. Yet the visual representation is useful as well. The web page is represented as a hierarchical visual adjacency multigraph. In the graph, each node represents an HTML object and each edge represents the spatial relation in the visual representation. Based on the results of visual analysis,

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

heuristic rules are applied to recognize multiple logical areas, which correspond to different meaningful parts of the page.

7.2 Using Neighbors

Web page these features are sometimes missing, misleading, or unrecognizable for various reasons. For example, some Web pages contain large images or flash objects, but little textual content. In such cases, it is difficult for classifiers make reasonable judgments based on the features on the page [2]. In order to address this problem, features can be extracted from neighboring pages that are related in some way to the page to be classified to supply supplementary information for categorization. There are a variety of ways to derive such connections between pages. One obvious connection is the hyperlink. However, other types of connections can also be derived, and some of them have been shown to be useful for Web page classification. These types of connections are using assumption, neighbor selection and utilizing artificial link.

VIII. RESEARCH ISSUES OF WEB PAGE CLASSIFICATION

8.1 Web Page Content Preprocessing

In most experimental work, preprocessing is performed before the content of Web pages. HTML tags are usually eliminated. However, the content of META keyword, META description, and "ALT" fields of image tags is usually preserved. Although stemming may be used in web search, it is rarely utilized in the classification. The stemming is used in indexing mainly in order to improve recall.

8.2 Dataset Selection and Generation

The manual labeling can require an excessive amount of human effort, many researchers use subsets of the existing Web directories that are publicly available. The two most frequently used Web directories are the Yahoo! Directory and the DMOZ (from directory.mozila.org) ODP (Open Directory Project) [7]. The use of different datasets makes it difficult to compare performance across multiple algorithms. Therefore, the Web classification research community would benefit from a standard dataset for Web page classification, such as the TREC (Text REtrival conference) datasets for information retrieval. Although the "WebKB" dataset is one such dataset, it is small and provides only limited functional classes.

8.3 Blog classification

The word blog was originally a short form of Web log, which, as defined by the Merriam Webster Dictionary, is a Web site that contains an online personal journal with reflections, comments, and often hyperlinks provided [5]. As blogging has gained in popularity in recent years, an increasing amount of research about blogs has also been conducted. Research into blog classification can be broken into three types: blog identification, mood classification, and genre classification. First category aims at identifying blog pages from a collection of Web pages, which is essentially a binary classification of blog and non-blog. The second category includes identification of the topic, mood or, sentiment of blogs. The third category focuses on the genre of blogs. For example, three types of blogs are news, commentary, and journal.

IX. CONCLUSION

This paper discussed about the concepts of web page classification in detailed form. Web page classification is an important process to categorize the web pages based on the content and structure, to remove the HTML tags and finally used to extract the content from web pages. Mostly the researchers collected the dataset of web directories which is managed by humans. The web directories, web page classification, techniques, applications, features and research issues are also discussed. It study is helpful for researchers and students who are doing their research in the field of web page classification and web content filtering.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

REFERENCES

[1] Xiaoguang Qi and Brian D. Davison "Web Page Classification: Features and Algorithms", ACM computing surveys (CSUR), 2009,
[2] Zinah A. Abbas, Prof .Dr.Shawkat K. Guirgui, Dr .Laithr. Fleih, Magda M. Madbouly, "Web Pages Classification Using Rule-Based System" IOSR Journal Of Mechanical And Civil Engineering, E-ISSN: 2278-1684, P-ISSN: 2320-334X, Volume 13, Issue 3.
[3] T. Xia, Y. Chai, "Improving SVM On Web Content Classification By Document Formulation", In Proceedings Of The 7th International Conference On Computer Science & Education (ICCSE '12), Australia, 14-17 July 2012, Pp. 110-113.
[4]. S. D. Vaghela, P. Patel, "Web Page Classification Techniques-A Comprehensive Survey", International Journal of Engineering Development and Research (IJEDR), Dec. 2014, Vol.1, Issue 3, Pp.229 – 232.
[5] A. P. Asirvatham, K. K. Ravi, "Web Page Categorization Based on Document Structure", International Institute of Information Technology, Hyderabad, India 500019, 2001.
[6] P. Manchanda, S. Gupta And K. K. Bhatia, "On the Automated Classification of Web Pages Using Artificial Neural Network", IOSR Journal Of Computer Engineering IOSRJCE, Vol. 4, Issue 1, Sep-Oct. 2012, PP. 20-25.
[7]. M. I. Devi, Dr. R. Rajaram And K. Selvakuberan, ―Automatic Web Page Classification By Combining Feature Selection Techniques and Lazy Learners", International Conference on Computational Intelligence and Multimedia Applications (ICCIMA -07), Pp. 33- 37, Sivakasi, Tamil Nadu, 13-15 Dec. 2007, Vol. 2.
[8] Shawn C. Tice, "Classification of Web Pages in YIOOP with Active Learning", May 2013, Master's Projects.Paper 297, San Jose State University.
[9] Patil, Ajay S, Pawar, B. V. Automated classification of web sites using Naive Bayesian algorithm. In: Proceedings of the international multiconference of engineers and computer scientists. 2012. p. 14-16.
[10] Henderson, Lachlan. 2009. "Automated Text Classification in the DMOZ Hierarchy"
[11] P.Kameshwari, E.Salini Varma, "Web Page Classification Approach, SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE) – volume 4 Issue 2–February 2017.
[12] Dr.Tamer Anwar Ahmed Alzohairy," An efficient method for web page classification based on text", International Journal of Engineering and Computer Science ISSN: 2319-724 Volume 6 Issue 6 June 2017
[13] Poonam Asawara, Dr Amit Shrivastava and Dr Manish Manoria, "Hybrid Appproach for Web Page Classification Based on Firefly and Ant Colony Optimization", International Journal of Computer Engineering and Applications, Volume XI, Issue I, Jan. 17, ISSN 2321-3469
[14] Zinah A. Abbas, Prof .Dr.Shawkat K. Guirgui, Dr .LaithR. Fleih, Magda M. Madbouly, "Web Pages Classification Using Rule-Based System", Journal of Mechanical and Civil Engineering, e-ISSN: 2278-1684, P-ISSN: 2320-334X, Volume 13, Issue 3, May - June 2016
[15] Deepshikha Patel and Dr. Prashant Kumar Singh, "A Web Page Classification Technique with Textual Content Analysis Using NN-PCA for Objectionable Web Page Classification", International Journal of Electronics Communication and Computer Engineering, Volume 4, Issue 2, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209
[16] Ajay S. Patil and B.V. Pawar, "Automated Classification of Web Sites using Naive Bayesian Algorithm", Proceedings of the International MultiConference of Engineers and Computer Scientists, March 14-16, 2012
[17] J. Alamelu Mangai, Dipti D. Kothari and V. Santhosh Kumar, "A Novel Approach for Automatic Web Page Classification using Feature Intervals", International Journal of Computer Science Issues, Vol. 9, Issue 5, No 2, September 2012, ISSN (Online): 1694-0814
[18] J. Alamelu Mangai, V. Santhosh Kumar, "A Novel Approach for Web Page Classification using Optimum features", International Journal of Computer Science and Network Security, VOL.11 No.5, May 2011.