A survey on techniques for web page classification for text mining

Mrs Anusha Meti
Assistant Professor, Dept of IT
DYPCOE,AKURDI ,PUNE, India

Dr Kalyan D Bamane
Associate Professor, Dept of IT
DYPCOE, AKURDI PUNE, India

***Abstract*— The rapid growth of the World Wide Web (www) is demanding for an automated assistance for Web page classification and categorization. In most existing Web page classification tasks, Web pages are classified into topical categories based on their content regardless of the possible relationships among them. In this paper, a comprehensive survey on classification of Web pages is presented. The features for creating tag information, classifiers and datasets used for experimentation are also discussed. It also gives comparative analysis of all Web page classification techniques. The challenges/Issues involved in developing Web page classification are also discussed. This would help researchers to take up new work on Web page classification and address most of the important challenges/issues.**

Keywords*— *Web page Classification; Web pages; Web content mining; Text mining.

I. INTRODUCTION

In recent years, the World Wide Web (WWW) has become a global data centre, which permits people to store and distribute their information. The information in Web pages may be related to be personal, official, commercial and business. The users of Web would like to access such information for their needs. Hence the Web page classification methods must properly organize Web pages so that the relevant information is supplied for the user queries. Such Web page classification methods use Web content mining, also known as text mining that aims on raw data exists in the Web pages. The Web Content mining is also the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query.

Web content mining is directed toward specific information provided by the customer search information in search engines. This allows for the scanning of the entire Web to retrieve the cluster content triggering the scanning of specific Web pages within those clusters. The results are pages relayed to the search engines through the highest level of relevance to the lowest. Though, the search engines have the ability to provide links to Web pages by the thousands in relation to the search content, this type of web mining enables the reduction of irrelevant information.

Web content mining is very effective when used in relation to a content database dealing with specific topics. For example online universities use a library system to recall articles related to their general areas of study. This specific content database enables to pull only the information within those subjects, providing the most specific results of search queries in search engines.

The main uses for this type of data mining are to gather, categorize, organize and provide the best possible information available on the WWW to the user requesting the information. This approach is imperative to scanning the many HTML documents, images, and text provided on Web pages. The resulting information is provided to the search engines in order of relevance giving more productive results of each search.

In short, the ability to conduct Web content mining allows results of search engines to maximize the flow of customer clicks to a Web site, or particular Web pages of the site, to be accessed

Numerous times in relevance to search queries. The clustering and organization of Web content in a content database enables effective navigation of the pages by the customer and search engines. Images, content, formats and Web structure are examined to produce a higher quality of information to the user based upon the requests made. Businesses can maximize the use of this text mining to improve marketing of their sites as well as the products they offer. The web content mining techniques are very useful for classification web pages. In this paper a comprehensive survey of various web page classification methods is presented.

The rest of the paper is organized in to three sections. Section 2 describes the existing techniques for classification of web pages. Section 3 highlights the challenges involved in developing new methods for web page classification. Section 4 gives the conclusion. Table 1 gives the comparative study of web page classification.

II. DESCRIPTION OF EXISTING TECHNIQUES FOR WEB PAGE CLASSIFICATION

Significant amount of work has gone into the research related to development of techniques for web page classification. Some of the important techniques are summarized in the following.

[1] **Dou Shen, Qiang Yang, Zheng Chen (2007)** proposes a technique to improve the web page classification performance by removing the noise through summarization process. The method uses empirical evidence that ideal web page summaries generated by human editors can indeed

improve the performance of web-page classification algorithms. Later, the method put forward a new web-page summarization algorithm, which is based on web page layout and the method is evaluated along with several other state of the art text summarization algorithms on the look smart web directory. Here, around 2 million web pages are used which are crawled from the ***look smart web directory (i.e http:// search.looksmart.com)***. The experimental results show that the classification algorithms i.e support vector machine augmented by any summarization approach can achieve an improvement by more than 5.0% as compared to pure text based classification algorithms. To improve the pure text based methods, an ensemble method is introduced to combine the different summarization algorithms. The ensemble summarization method achieves 12.0% improvement over pure text based methods.

[2] **Rung-ching chen, Chung-Hsun Hsieh (2005)** describe a web page classification based on support vector machine(SVM) using a weighted vote schema for various features. The system uses both latent semantic analysis and web page feature selection training and recognition by the SVM model. The method extracted text features from web page content. The dataset used here is ***sports news*** especially to test system performance, sports news was downloaded from ***udndata*** website. Data sets include various games such as basketball, baseball, golf, tennis, volley ball, soccer, billiards, football and formula 1 racing. So based upon the output of the support vector machine, a voting schema is used to determine the category of the web page. The weighted vote support vector machine yields a better accuracy even with small data set.

[3] **A.J.Shaikh,V.L.Kolhe (2013)** report a framework for Web Content Mining Using Semantic Search and Natural Language Queries. Here the method implements a framework for semantic based web content mining system using semantic ontology and SPARQL. It converts natural language queries into SPARQL quires using NLP query processing module. For testing, the method considers a framework for cricket domain in which it shows a better improvement over traditional keyword based searching. The implemented framework is for ***cricket domain for ICC World Twenty20, 2012-13 Series***. In which OWL API where used for semantic mapping. Protégé is used for ontology design. The performance of keyword based search is than compared with SPARTQL query search. SPAQRL query search gives more precise results compared to keyword based search.

[4] **Ali Ahmadi, Mehran Fotouhi, Mahmoud Khaleghi (2010)** proposes an Intelligent Classification of Web pages using Contextual and visual features. The method is applied for classification of pornographic Web pages. The filtering of unwanted Web content is achieved based on blocking a specific Web address via searching it in a reference list of black URLs or by doing a plain contextual analysis on the page by searching special keywords in the text. In this paper, the intelligent approach which is based uses textual, profile and visual features in a hierarchical structure classifier. The ID3 classifier is used for textual and profile features. The textual features contain information about keywords and black-words. The profile features contains structural information like number of links, meta-tags, pictures etc. The algorithm is applied on a dataset consisting of ***1295 web pages*** as training set which includes 700 porn pages (which includes text, image or both) in both English and Persian and also includes 595 non-porn pages which again includes pages with ***medical, health and sports***. The neural network model is used for skin color and visual features. The model attains 95% accuracy by using test dataset with 290 web-pages.

[5] **Chih-Ming Chen,Hahn-Ming Lee,Yu-Jung Chang (2007)** employed a two novel feature selection approaches for web page classification. In this model a fuzzy ranking analysis paradigm has been described together with a novel relevance measure and discriminating power measure (DPM) to effectively reduce the input dimensionality from tens of thousands to a few hundred with zero rejection rates and small decrease in accuracy. It emphasizes classification in parallel order, rather than classification in serial order. The dataset is obtained from ***China-Times Web site (http://news.chinatimes.com/) and Reuter-21578***.Thus the result obtained is that the DPM can reduce both redundancy and noise features to set up a better classifier.

[6] **Inma Hernandez, Carlos R.Rivero, David Ruiz, Rafael Corchuelo** (2013) proposed to automatically generate URL-based web page classifiers which can be used in the context of enterprise web information systems. It builds a number of URL patterns that represent the different classes of pages in a web site, so further pages can be classified by matching their URLs to the patterns. The system is experimented on top ***40 alexa websites*** written in English. And have achieved an average precision of 98% and average recall of 90%.The system is mainly used for real-world web page classification.

[7] **Selma Ayse Ozel (2010)** describes a web page classification system based on a genetic algorithm using tagged-terms as features. In this method, automatic Web page classification system has been used, which uses both HTML tags and terms as classification features. The system classifies Web pages by simply computing similarity between the learned classifier and the new Web pages. The system is tested on datasets such as ***conference, course, and student web pages***. The classification accuracy is 95%.

[8] **R.Etemadi, N.Moghaddam (2010)** employs an approach in web content mining for clustering web pages. The algorithm uses data content and new similarity criterion for classification of web pages. For evaluating the accuracy of algorithm some pages with ***five subjects of software engineering, computerized networks, and architecture of computer, parallel processing and operating system*** are taken as datasets and have been investigated. The results obtained from simulation show high efficiency of the algorithm in separating web pages and their clustering.

[9] **Tarique Anwar,Muhammad Abulaish (2012)** proposes a Markov Clustering(MCL) based text mining approach for namesake disambiguation on the web. The technique represents the collection of web pages as a weighted graph and apply MCL to determine different clusters. The system mainly focuses on three broad and realistic aspects to cluster web-pages retrieved through search engines which includes content overlapping, structure overlapping, and local context overlapping. For experimentation, two different datasets are used namely ***Bekkerman and McCallum.*** It is found that the computational complexity of the proposed method is quite satisfactory in comparison to other state-of-art techniques.

[10] **Aixin Sun,Ying Liu,Ee-Peng Lim (2011)** describes Web classification of conceptual entities using co-training. Web pages are described on physical or abstract entity, e.g., company, people, and event. Furthermore, users often like to organize pages into conceptual categories for better search and retrieval. In this work the web pages are categorized into conceptual categories. For experimentation ***Conf-425 dataset*** is used. The pre-processing of Conf-425 dataset includes HTML tag removal, stop-word removal, and term stemming. More importantly, the accuracy of EcT was not much worse than classification methods that used a large set of training examples.

[11] **Ahmad Pouramini,Shahram Nasiri (2015)** proposes a wrapping language supported by a visual tool to create wrappers for extracting the main content from web pages. In this language, various types of features such as syntactical, semantic, visual can be employed in the extraction rules to identify the content of interest. For experimentation the method is tested on ***websites such as Wikipedia, yahoo news, NYTimes, BlogSpot and Word press.***

[12] **Tao Jiang, Ah-Hwee Tan, Ke Wang (2007)** proposes a two-step procedure to mine generalized Associations of semantic relations from textual web content. First, RDF (Resource Description Framework) metadata representing semantic relations are extracted from raw text using natural language processing techniques. Then, a novel generalized association pattern mining algorithm (GP-Close) is applied to discover the underlying association patterns on RDF metadata. The experiments were performed on a ***desktop PC running Windows XP with a P4-2.6G CPU and 1 G RAM.*** The GPClose algorithm was implemented using Java (JDK 1.4.2). Two variants of GP-Close with different sizes of ***tidset buffer*** were used in the experiments, namely, GP-Close-0 with a tidset buffer of 0 KB and GP-Close-50000 with a tidset buffer of 50,000 KB. The experimental result shows that the GP-Close algorithm substantially reduce the pattern redundancy and perform much better than the original generalized association rule mining algorithm in terms of time efficiency.

[13] **S.Yasodha, S.S.Dhenakaran** (2014) presents an Ontology-Based framework for Semantic Web Content Mining. Framework has been implemented for three major domains ***Education, Medicine and Tourism.*** The algorithm has been implemented in JAVA with RDF at the back end for storing ontology's. The performance of the framework is evaluated by three metrics: Precision, Average Precision and Relevance Score. The efficiency of the framework is measured in terms of precision, average precision and relevance score.

[14] **G.S. Tomar, Shekhar Verma, Ashish Jha** (2006) introduces the concept of a classification tool for web pages called Web Classify, which uses modified naïve Bayesian (NB) algorithm with multinomial model to classify pages into various categories. The tool starts the classification from downloading training web text from internet, preparing the hypertext for mining, and then storing web data in a local database. Web pages were taken from web directories, which are pre-classified into various categories and those pages were pre-processed before the words can be sent to training data set. The system has been tested on a very small set of test documents and the vocabulary size of the corpus is also low. The modified approach is measured with a threshold value of 0.4. So NB has a classification accuracy of 42.5% while the approach has a classification accuracy of 55%. This clearly shows the enhancement in performance.

[15] **Majid Javid Moayed, A. Hamid Sabery, A. Hamid Sabery** (2008) Investigates usage of a swarm intelligence algorithm in the field of the web page classification. Focusing on Persian web pages Ant Miner II is the used algorithm. It also proposes a simple text preprocessing technique to reduce the large numbers of attributes associated with web content mining. The web pages of news are the most suitable choices for the experimentation because of the solidarity of their contents and being classified. In order to experiment the model, the ***web pages of Irna news*** are used. The result shows Ant Miner II and proposed preprocessing technique is efficient in the field of the web page classification.

[16] **Vladimír Bartík** (2009) describes association based classification for relational data, which can be used for the data extraction from web pages. The method is tested on two ***Relational Databases NURSERY and ADULT which are taken from the UCI Machine Learning Repository***. Next, the experiments with various data is presented, with emphasis on data obtained by extraction and segmentation of web pages. The accuracy showed to be about 80%.

[17] **Jiao Lijuan, Feng Liping** (2010) presents a method to improve classification accuracy of Web pages by using the hyperlink factor. The Web pages are classified by using KNN classifier. Three hundreds of documents are selected to this experiment, 210 of which are taken as training ***corpus including sixty on financial, fifty on sports, sixty on culture, forty on military and ninety of which are testing ones***. There are 8617 feature items which are extracted. Classification results are evaluated by precision and recall rate which are accepted internationally. The classification accuracy would be increased by 10% or more if hyperlink factor is inducted for web pages by the experiment. Introduction of hyperlink elements of web pages can improve the classification accuracy

in feature selection method based on mutual information and correlation by experiment. So the improvement is effective in web page classification.

[18] **Vladimír Bartík** (2010) describes text-based Web page classification that uses both textual and visual information to find a suitable representation of web page content, based on term frequency (TF) or inverse document frequency (TF-IDF) weighting . The model is experimented on ***WebKB corpus*** of web pages to verify the functionality. The first dataset contains 4518 web pages from the computer science department websites. The second dataset was manually created Using web pages taken from several English written news websites (CNN.com, Reuters.com, nytimes.com, boston.com and usatoday.com). The accuracy achieved is approximately 80%.

[19] **Lay-Ki Soon, Sang Ho Lee** (2010) describes classifying web pages using information extraction patterns – preliminary results and findings. The model uses natural language processing (NLP) techniques such as Naïve Bayesian classifiers, Support Vector Machine (SVM) and association rule mining (ARM). For experimentation the model uses ***Sundance Sentence Understanding and Concept Extraction (Sundance)*** to obtain the IE patterns. The experimental results indicate that the existence of a word in different contexts has different impact to the classification task. Thus, the extraction patterns used to represent each document are more semantically meaningful and give better insight to web classification in comparison with keywords.

[20] **Hakan Ayral, Sırma Yavuz** (2011) describes an automated domain specific stop word generation method for natural language text classification. The model implements a Bayesian Natural language classifier working on web pages. The model is experimented on ***PASCAL dataset***, the results shows that document coverage rank and topic coverage rank of words belonging to natural language corpora follows Zipf's law.

[21] **Hu Mingsheng, Jia Zhijuan, Zhang Xiangyu** (2012) employees an approach for text extraction from Web news page. The model Uses tree structure of Document Object Model (DOM) when analyzing web page. The model is experimented on some randomly selected websites such as ***News.qq.com, News.sohu.com, News.163.com, News.tom.com, Cn.msn.com,*** **www.china.com.cn** ***and news.cn.yahoo.com***. The results obtained shows that the method is both versatile and a highly accurate. If the web Page is some fairly standard news web pages, the accuracy Rate can reach to 98%. In practice, they have carried out extraction on pages from 150 websites; the accuracy of the sampling rate is 94%.

[22] **Wang Zhixing, Chen Shaohong** (2011) represents web page classification based on semi-supervised Naive Bayesian EM algorithm. The model uses Hierarchical Clustering EM framework to train Naive Bayesian Classifier iteratively. The model is tested on ***Look Smart database*** and most of the selected Web pages contain mainly text, including 2000 pages and 6 chief categories namely: Entertainment, Work, Shopping, Sport, Travel and Society. The result of the experiment proved that the method introduced in the model shows good effect of Web classification.

[23] **Dongjin Choi, Byeongkyu Ko, Eunji Lee, Myunggwon Hwang, Pankoo Kim** (2012) employed automatic evaluation of document classification using n-gram statistics, which have a great possibility to find similarities between given documents. The proposed method is compared with traditional method suggested by Keselj. The model is tested on ***bioinformatics data base, computer vision, fuzzy system, mobile and NLP***. Each category contains 100 research documents collected from IEEE digital library. The performance using this method is better than the Keselj approach.

[24] **Hammad Haleem,Pankaj Kumar Sharma,MM Sufyan Beg** (2014) describes a novel frequent sequential patterns based probabilistic model for effective classification of web documents. The classification model proposed utilizes the features of naïve bayes classifiers and the Pattern discovery model (PTM). The model is experimented on a collection of documents related to predefined categories. The documents were taken from two different sources, firstly they used the ***Reuters Corpus version 1 (RCV1)*** and secondly on ***crawled dataset***. After testing the novel approach on RCV1 dataset, 88% accuracy is obtained.

[25] **Kolla Bhanu Prakash, M.A. Dorai Rangaswamy, Arun Raja Raman** (2013) represents attribute based content mining for regional web documents. The model outlines the use of attributes for content extraction, using basic pixel attributes, pattern matching, statistical model, and Artificial Neural Network training. This preliminary study is focused to bring out the complexities in ***regional web documents*** and how to present popular text mining techniques.

[26] **Moonis Javed, Aly Akhtar, Akif Khan Yusufzai** (2015) describes classification of web pages as evergreen or ephemeral based on content. The approach that has been used is a combination of text classification and other binary classification. The training dataset provided by ***StumbleUpon*** was a set of urls with some Meta information like the category of the page, html ratio, is news, along with some boilerplate code (like title and content) of the page. The model is tested on StumbleUpon and an overall accuracy of 88% is reported.

[27] **Jia Wu, Shirui Pan, Xingquan Zhu, Zhihua Cai** (2015) describes boosting for multi-graph classification. The model uses graph-based learning problem and multi-graph classification (MGC), which aims to learn a classifier from a set of labeled bags each containing a number of graphs inside the bag. The model is experimented on datasets such as ***DBLP Multi-Graphs***, the DBLP dataset consists of bibliography data in computer science. Experiments and comparisons on real-world multi-graph learning tasks demonstrate the algorithm performance.

[28] **Wenlong Ren ,Jianzhuo Yan** (2015) represents an improved Cerebellar Model Articulation Controller (CMAC) neural network model for web mining. The CMAC is an excellent classification technique, but when it is applied to deal with high-dimensional dataset such as the data on the Internet, the memory required increase intensively. However, this improved CMAC model requires less memory for processing high-dimensional dataset. The model is tested on datasets such as ***Syskill & Webert Web pages*** ratings. Experiments on the four topics show that the improved CMAC model performs a better predicting accuracy rate to identify user interesting Web pages than other well-known classification method.

[29] **Aanshi Bhardwaj, Veenu Mangat** (2015) represents a novel approach for content extraction from web pages. The model discusses various approaches for extracting informative content from web pages and a new approach for content extraction from web pages using word to leaf ratio and density of links.

[30] **Jian Zhu, Hanshi Wang, JinTao Mao** (2014) describes sentiment classification using genetic algorithm and conditional random fields. Conditional Random Fields (CRFs) is employed to model the emotional tendency of web pages which are divided into different types of comments, such as positive comments, negative comments and objective comments. To test the model a ***corpus of 400 online product reviews from the product: x61***. Experimental results on the product reviews and the 1998 People's Daily corpus show that the proposed algorithm works reasonable in the real calculation.

III. CHALLENGES/ISSUES

The existing web page classification methods have many issues as listed below:

a. Selection of appropriate features
b. Either the HTML tags or terms are considered as the features
c. Tag information representation.
d. It is difficult to classify the unstructured data from web pages.
e. Difficulty in finding relevant information.
f. Extracting new knowledge from the web.

IV. CONCLUSION

In this paper, the techniques for Web page classification techniques are described. The techniques employ Web-page summarization algorithm, support vector machine, Semantic Search, Natural Language Queries, genetic algorithm, Markov Clustering(MCL), Resource Description Framework(RDF),term frequency(TF) ,inverse document frequency (TF-IDF) etc which are experimented on datasets such as look smart web directory, Sports news from udndata website, Cricket domain for ICC World Twenty20,2012-13 Series are presented. According to the survey done the classification is done on features such as keywords, visual features, expressions, hyperlink factor, sentiments etc. However still there is a challenge to develop new classification techniques.

V. COMPARATIVE STUDY

The comparative study of all approaches for Web page classification is summarized below:

TABLE I. COMPARATIVE STUDY

Author	Description of method	Datasets	Accuracy
DouShen,Qiang Yang,Zheng Chen(2007) [1]	Noise reduction through summarization for Web-page classification	2 million web pages crawled from the look smart web directory	12.0%
Rung-ching chen ,Chung-Hsun Hsieh (2005) [2]	Web page classification based on a support vector machine using a weighted vote schema	Sports news from udndata website	The weighted vote support vector machine yields a better accuracy even with small data set.
A.J.Shaikh, V.L.Kolhe (2013) [3]	Framework for Web Content Mining Using Semantic Search and Natural Language Queries	Cricket domain for ICC World Twenty20,2012-13 Series.	SPAQRL query search gives more precise results compared to keyword based search.
Ali Ahmadi, Mehran Fotouhi ,Mahmoud Khaleghi(2010) [4]	Intelligent classification of web pages using contextual and visual features	1295 web pages as training set which includes 700 porn pages (which includes text, image or both) in both English and Persian and also includes 595 non-porn pages which again includes pages with medical, health, sports.	Here it attains 95% accuracy by using test dataset with 290 web-pages
Chih-Ming Chen,Hahn-Ming Lee,Yu-Jung Chang(2007) [5]	Two novel feature selection approaches for web page classification	China-Times Web site (http://news.chinatimes.com/) and Reuter-21578.	DPM can reduce both redundancy and noise features to set up a better classifier.

Inma Hernandez, Carlos R.Rivero, David Ruiz, Rafael Corchuelo (2013) [6]	CALA: An unsupervised URL-based web page classification system	40 websites from Alexa which is written in English .	And have achieved an average precision of 98% and average recall of 90%.
Selma Ayse Ozel (2010) [7]	A Web page classification system based on a genetic algorithm using tagged-terms as features	The course and the student datasets obtained from WebKB project website (http://www.cs.cmu.edu/-webkb)	When there is enough number of negative documents in the training dataset, the classifier reaches 95% accuracy.
R.Etemadi, N.Moghaddam (2010)[8]	An Approach in Web Content Mining for Clustering Web Pages	Pages with five subjects of software engineering , computerized networks , architecture of computer, parallel processing and operating system	The results obtained from simulation show high efficiency of the algorithm proposed in separating web pages and their clustering.
Tarique Anwar,Muhammad Abulaish (2012)[9]	An MCL-Based Text Mining Approach for Namesake Disambiguation on the Web	Bekkerman and McCallum.	It is found that the computational complexity of the proposed method is quite satisfactory in comparison to other state-of-art techniques.
Aixin Sun,Ying Liu,Ee-Peng Lim (2011)[10]	Web classification of conceptual entities using co-training	Conf-425 dataset. The pre-processing of Conf-425 dataset included HTML tag removal ,stop-word removal, and term stemming	The accuracy of EcT was not much worse than classification methods that used a large set of training examples.
Ahmad Pouramini,Shahram Nasiri (2015) [11]	Web content extraction using contextual rules	Websites such as Wikipedia, yahoo news, NYTimes, BlogSpot, Word press	Not reported
Tao Jiang, Ah-Hwee Tan, Ke (2007)[12]	Mining Generalized Associations of Semantic Relations from Textual Web Content	Desktop PC running Windows XP with a P4-2.6G CPU and 1 G RAM, Two variants of GP-Close with different sizes of tidset buffer, GP-Close-0 with a tidset buffer of 0 KB and GP-Close-50000 with a tidset buffer of 50,000 KB.	GP-Close algorithm substantially reduce the pattern redundancy.
S.Yasodha, S.S.Dhenakaran (2014)[13]	An Ontology-Based Framework for Semantic Web Content Mining	The framework has been implemented for three major domains, Education, Medicine and Tourism.	The efficiency of the framework is measured in terms of precision, average precision and relevance score.
G.S. Tomar, Shekhar Verma, Ashish Jha (2006) [14]	Web Page Classification using Modified Naïve Bayesian Approach	Web directories, The proposed approach has been tested on a very small set of test documents and the vocabulary size of the corpus is also low	The modified approach is measured with a threshold value of 0.4. So NB has a classification accuracy of 42.5% while our approach has a classification accuracy of 55%.
Majid Javid Moayed, A. Hamid Sabery, A. Hamid Sabery (2008)[15]	Ant Colony Algorithm for Web Page Classification	Web pages of news, the web pages of Irna news are used	The result shows Ant Miner II and proposed preprocessing technique are efficiency in the field of the web page classification.
Vladimír Bartík (2009) [16]	Association Based Classification for Relational Data and Its Use in Web Mining	For Experiments with Relational Databases NURSERY and ADULT datasets taken from the UCI Machine Learning Repository	The accuracy showed to be about 80%.

		were used	
Jiao Lijuan, Feng Liping (2010) [17]	Improvement of Feature Extraction in Web Page Classification	300 documents are selected to this experiment, 210 of which are taken as training corpus including sixty on financial, fifty on sports, sixty on culture, forty on military and ninety of which are testing ones. There are 8617 feature items are extracted	The classification accuracy would be increased by 10% or more if hyperlink factor is inducted for web pages by the experiment
Vladimír Bartík (2010) [18]	Text-Based Web Page Classification with Use of 7Visual Information	4518 web pages from the computer science department websites. The second dataset was manually created. It contains web pages taken from several English written news websites (CNN.com, Reuters.com, nytimes.com, boston.com and usatoday.com).	The accuracy was approximately 80% for both weightings.
Lay-Ki Soon, Sang Ho Lee (2010)[19]	Classifying Web Pages using Information Extraction Patterns – Preliminary Results and Findings	Four university dataset The dataset consists of 8,282 web pages collected from computer science departments of several universities . 4,162 of the web pages are	The extraction patterns used to represent each document are more semantically meaningful and give better insight to web classification in comparison with keywords.
		from Cornel, Texas, Washington and Wisconsin University, while the other 4,120 are from other universities .	
Hakan Ayral, Sırma Yavuz(2011)[20]	An Automated Domain Specific Stop Word Generation Method for Natural Language Text Classification	PASCAL Dataset	Investigated the distribution of stop-word lists Generated by the model and compared their contents against a Generic stop-word list for English language.
Hu Mingsheng, Jia Zhijuan, Zhang Xiangyu (2012)[21]	An Approach for Text Extraction From Web News Page	Randomly selected from some websites such as News.qq.com, News.sohu.com, News.163.com, News.tom.com, Cn.msn.com, www.china.com.cn, news.cn.yahoo.com.	From the results obtained this Method is of both versatility and a high accuracy. If the web Page is some fairly standard news web pages, the accuracy Rate can reach to 98%. they have carried out extraction on pages from 150 websites, the accuracy of the sampling rate is 94% or More
Wang Zhixing, Chen Shaohong(2011)[22]	Web Page Classification based on Semi-supervised Naive Bayesian EM Algorithm	Look Smart(Entertainment, Work, Shopping, Sport, Travel, Society) .	The result of the experiment proved that the method introduced in the paper shows good effect of Web classification.
Dongjin Choi, Byeongkyu Ko, Eunji Lee, Myunggwon Hwang, Pankoo Kim(2012)[23]	Automatic Evaluation of Document Classification using N-gram Statistics	Training data sets were chosen from 'bioinformatics data base,' 'computer vision,' fuzzy system,' 'mobile', and 'NLP'.	The performance using this method is better than the Keselj approach.
Hammad Haleem,Pankaj Kumar Sharma,M	Novel Frequent Sequential Patterns based Probabilistic Model for	The documents where taken from	After testing this novel approach on RCV1 dataset,

M Sufyan Beg(2014)[24]	Effective Classification of Web Documents	two different sources, firstly they used the Reuters Corpus version 1(RCV1), secondly on their own crawled dataset.	we were able to obtain classify the test documents with 88% accuracy.
Kolla Bhanu Prakash, M.A. Dorai Rangaswamy, Arun Raja Raman(2013) [25]	Attribute based content mining for regional Web documents	Regional web documents	Not reported
Moonis Javed, Aly Akhtar, Akif Khan Yusufzai(2015)[26]	Classification of Web Pages as Evergreen or Ephemeral based on content	The training dataset provided by StumbleUpon was a set of urls with some Meta information like the category of the page, html ratio, is news, along with some boilerplate code (like title and content) of the page.	Using this we have been able to get an overall accuracy of 88%.
Jia Wu, Shirui Pan, Xingquan Zhu, Zhihua Cai(2015)[27]	Boosting for Multi-Graph Classification	DBLP Multi-Graph Dataset: The DBLP dataset consists of bibliography data in computer science.	Experiments and comparisons on real-world multi-graph learning tasks demonstrate the algorithm performance.
Wenlong Ren ,Jianzhuo Yan(2015)[28]	An Improved CMAC Neural Network Model for Web Mining	The dataset of Syskill & Webert Web pages ratings were used to test the improved CMAC model.	Experiments on the four topics show that the improved CMAC model performs a better predicting accuracy rate to identify user interesting Web pages than other well-known classification method.
+Aanshi Bhardwaj, Veenu Mangat (2014)[29]	A Novel Approach for Content Extraction from Web Pages	Not reported	Not reported
Jian Zhu, Hanshi Wang, JinTao Mao(2010)[30]	Sentiment Classification Using Genetic Algorithm and Conditional Random Fields	compiled a corpus of 400 online product reviews from this product: x61	Experimental results on both the product reviews and the 1998 People's Daily corpus show that the proposed algorithm works reasonable in the real calculation.

References

[1] Chih-Ming Che, Hahn-Ming Lee, Yu-Jung Chang," Two novel feature selection approaches for web page classification", Expert Systems with Applications 36 (2009) 260–272

[2] Rung-Ching Chen, Chung-Hsun Hsieh," Web page classification based on a support vector machine using a weighted vote schema", Expert Systems with Applications 31 (2006) 427–435

[3] A. J. Shaikh and V. L. Kolhe, "Framework for web content mining using semantic search and natural language queries," *2013 IEEE International Conference on Computational Intelligence and Computing Research*, Enathi, 2013, pp. 1-5.

[4] Ali Ahmadi, Mehran Fotouhi, Mahmoud Khaleghi "Intelligent classification of web pages using contextual and visual features", Applied Soft Computing 11 (2011) 1638–1647

[5] Chih-Ming Chen, Hahn-Ming Lee, Yu-Jung Chang "Two novel feature selection approaches for web page classification", Expert Systems with Applications 36 (2009) 260–272

[6] Inma Hernández, Carlos R. Rivero, David Ruiz, Rafael Corchuelo "CALA: An unsupervised URL-based web page classification system" , Knowledge-Based Systems 57 (2014) 168–180

[7] Selma Ays_e Ozel "A Web page classification system based on a genetic algorithm using tagged-terms as features", Expert Systems with Applications 38 (2011) 3407–3415

[8] R. Etemadi and N. Moghaddam, "An approach in web content mining for clustering web pages," *2010 Fifth International Conference on Digital Information Management (ICDIM)*, Thunder Bay, ON, 2010, pp. 279-284.

[9] T. Anwar and M. Abulaish, "An MCL-Based Text Mining Approach for Namesake Disambiguation on the Web," *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Macau, 2012, pp. 40-44.

[10] Aixin Sun, Ying Liu, Ee-Peng Lim, "Web classification of conceptual entities using co-training", Expert Systems with Applications 38 (2011) 14367–14375

[11] A. Pouramini and S. Nasiri, "Web content extraction using contextual rules," *2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, Tehran, 2015,pp.1014-1018.

[12] T. Jiang, A. h. Tan and K. Wang, "Mining Generalized Associations of Semantic Relations from Textual Web Content," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 2, pp. 164-179, Feb. 2007.

[13]S. Yasodha and S. S. Dhenakaran, "An ontology-based framework for Semantic Web Content Mining," *2014 International Conference on Computer Communication and Informatics*, Coimbatore, 2014, pp. 1-6.

[14] G. S. Tomar, S. Verma and A. Jha, "Web Page Classification using Modified NaÃ¯ve Bayesian Approach," *TENCON 2006 - 2006 IEEE Region 10 Conference*, Hong Kong, 2006, pp. 1-4.

[15] M. J. Moayed, A. H. Sabery and A. Khanteymoory, "Ant colony algorithm for web page classification," *2008 International Symposium on Information Technology*, Kuala Lumpur, 2008, pp.1-8.

[16] V. Bartik, "Association based classification for relational data and its use in web mining," *2009 IEEE Symposium on Computational Intelligence and Data Mining*, Nashville, TN, 2009,pp.252-258.

[17] L. Jiao and L. Feng, "Improvement of Feature Extraction in Web Page Classification," *2010 2nd International Conference on E-business and Information System Security*, Wuhan, 2010, pp. 1-3.

[18] V. Bartík, "Text-Based Web Page Classification with Use of Visual Information," *2010 International Conference on Advances in Social Networks Analysis and Mining*, Odense, 2010, pp.416-420.

[19] L. K. Soon and S. H. Lee, "Classifying Web Pages Using Information Extraction Patterns Preliminary Results and Findings," *2010 Sixth International Conference on Signal-Image Technology and Internet Based Systems*, Kuala Lumpur, 2010, pp. 195-202.

[20] H. Ayral and S. Yavuz, "An automated domain specific stop word generation method for natural language text classification," *2011 International Symposium on Innovations in Intelligent Systems and Applications*, Istanbul, 2011, pp. 500-503.

[21] H. Mingsheng, J. Zhijuan and Z. Xiangyu, "An approach for text extraction from web news page," *2012 IEEE Symposium on Robotics and Applications (ISRA)*, Kuala Lumpur, 2012, pp. 562-565.

[22] C. Shaohong and W. Zhixing, "Web page classification based on Semi-supervised Naïve Bayesian EM algorithm," *2011 IEEE 3rd International Conference on Communication Software and Networks*, Xi'an, 2011, pp. 242-245.

[23] D. Choi, B. Ko, E. Lee, M. Hwang and P. Kim, "Automatic Evaluation of Document Classification Using N-Gram Statistics," *2012 15th International Conference on Network-Based Information Systems*, Melbourne, VIC, 2012, pp. 739-742.

[24] H. Haleem, P. K. Sharma and M. M. Sufyan Beg, "Novel frequent sequential patterns based probabilistic model for effective classification of web documents," *2014 International Conference on Computer and Communication Technology (ICCCT)*, Allahabad, 2014, pp. 361-371.

[25] K. B. Prakash, M. A. D. Rangaswamy and A. R. Raman, "Attribute based content mining for regional web documents," *IET Chennai Fourth International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2013)*, Chennai, 2013, pp. 368-373.

[26] M. Javed, A. Akhtar and A. K. Yusufzai, "Classification of Web Pages as Evergreen Or Ephemeral Based on Content," *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, Jabalpur, 2015, pp. 1381-1385.

[27] J. Wu, S. Pan, X. Zhu and Z. Cai, "Boosting for Multi-Graph Classification," in *IEEE Transactions on Cybernetics*, vol. 45, no. 3, pp. 416-429, March 2015.

[28] W. Ren and J. Yan, "An Improved CMAC Neural Network Model for Web Mining," *2015 8th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, 2015, pp. 614-618.

[29] A. Bhardwaj and V. Mangat, "A novel approach for content extraction from web pages," *2014 Recent Advances in Engineering and Computational Sciences (RAECS)*, Chandigarh, 2014, pp. 1-4.

[30] J. Zhu, H. Wang and J. Mao, "Sentiment classification using genetic algorithm and Conditional Random Fields," *2010 2nd IEEE International Conference on Information Management and Engineering*, Chengdu, 2010, pp. 193-196.