ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 5, May 2016

# WEBPAGE CATEGORIZATION: AN OVERVIEW

**P.Malarvizhi [1], N.Radhika [2]**

Research Scholar, Department of CSE , Karpagam University, Coimbatore , Tamilnadu, India[1]
Professor, Department of CSE, Amrita Vishwa Vidyapeetham, Coimbatore , Tamilnadu, India[2]

***Abstract*— The explosive growth of web content creates a need for classification of web pages to facilitate searching and retrieval of information from it. Web page classification is much more difficult due to the noisy information embedded in it. Web page categorization provides a way for managing web information, maintaining web directories , focused crawling and improves the quality of the search. This paper presents an overview of web page categorization and its importance in the field of web applications.**

***Index Terms*— classification, classifiers, features, web pages**

## I. INTRODUCTION

With the rapid growth of World Wide Web (WWW) there is an increasing need to provide automated assistance to Web users for Web page classification. Such assistance is helpful in organizing the vast amount of information returned by keyword-based search engines or in constructing catalogues that organize web documents into hierarchical collections [1]. Web page classification also known as web page categorization is the process of assigning a web page to one or more predefined category labels. The general problem of web page classification can be divided into more specific problems of subject classification, functional classification, sentiment classification and other types of classification. Subject classification is concerned about the subject or topic of a web page. For example, judging whether a page is about 'arts', 'business' or 'sports' is an instance of subject classification. Functional classification cares about the role that the web page plays. For example, deciding a page to be a 'personal homepage', 'course page' or 'admission page' is an instance of functional classification. Sentiment classification focuses on the opinion that is presented in a web page, that is the author's attitude about some particular topic. Other types of classification include genre classification, search engine spam classification and so on [2] . The first step of web page classification is the collection of web pages the data set upon which the classification is to be performed. Data set the web pages are collected from standard data set or by crawling the web using webcrawler. Webcrawler is an automated software agents called crawlers used by crawler based search engines, visit a web site, read the information on the actual site, read the meta tags of the site and also does indexing on all linked web sites by following the links that are connected by the site. The process by which the webcrawler collects pages from the web is known as crawling. Webcrawler begins with a single URL, downloads that page, retrieves the links from that page to others, and the process is repeated with each of those pages [3]. Based on the number of classes in the problem, classification can also be divided into binary classification and multi-class classification wherein binary classification categorizes instances into exactly one of two classes and multi-class classification deals with more than two classes. If a problem is multi-class, say four-class classification, it means four classes are involved, say Arts, Business, Computers, and Sports. Classes may refer to categories here, such as 2 categories in binary classification or multiple categories in multi-class classification [4]. Web page classification gives a great impact in the field of web applications and the applications of web page classification are presented here.

- Managing web information
- Maintaining web directories
- Improving quality of search
- Querying
- Web content filtering
- Focused crawling

The paper is organized as follows. Section 2 provides a brief review of the related work while Section 3 presents the web page classification. The phases employed in the web page categorization are detailed in Section 4 and Section 5 presents the various types of evaluation metrics. Section 6 ends with conclusion.

## II. RELATED WORK

Xiaoguang Qi and Brian D. Davison [2] have presented the importance of the web specific features and algorithms, described the state-of-the-art practices and tracked the underlying assumptions behind the use of information from neighboring pages. In their work, they have detailed the features and its types, various techniques to select the features to reduce the dimensionality of the feature space to improve the quality of the web search. Also, they have presented the applications of web page classification along with other issues of dataset selection and web page preprocessing.

## III. WEB PAGE CLASSIFICATION

This section presents the types and methods of features and classification.

### *A. Classification types*

Classification is of single class or multiclass classification. In single class classification, only one predefined class is assigned to an instance among two predefined classes where in multiclass classification an instance is assigned to one or more predefined classes. Single class classification may be referred as binary classification and multiclass classification is of one for all classification is shown below in Figure 1, Figure 2 and Figure 3 where D1,D2,D3, etc., represent the documents and C1,C2,C3,etc., represent the categories.

Web pages classification classified labels

| | | | | | | |
|---|---|---|---|---|---|---|
| D1 | | | C1 | C2 | D1 | C2 |
| D2 | | D1 | 0 | 1 | D2 | C1 |
| D3 | | D2 | 1 | 0 | D3 | C1 |
| | | D3 | 1 | 0 | | |

Figure 1 Binary class single label classification

| | | | | | | |
|---|---|---|---|---|---|---|
| D1 | | | C1 | C2 | C3 | D1 | C3 |
| D2 | | D1 | - | - | x | D2 | C1 |
| D3 | | D2 | x | - | - | D3 | C2 |
| D4 | | D3 | - | x | - | D4 | C3 |
| | | D4 | - | - | x | | |

Figure 2 Multi class single label classification

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| D1 | | C1 | C2 | C3 | D1 | C1,C2 |
| D2 | D1 | x | x | | D2 | C2,C3 |
| D3 | D2 | - | x | x | D3 | C2 |
| D4 | D3 | - | x | - | D4 | C1,C2,C3 |
| | D4 | x | x | x | | |

Figure 3 Multi class multilabel classification

### *B. Dataset and dataset types*

Dataset is the collection of web pages collected from Standard dataset or from artificially generated dataset and various methods used to represented it are Bag of words, Set of words, N-gram representation etc.,

### *C. Feature Extraction*

Features the terms are extracted from the collected web pages after performing the preprocessing of stemming and stopword removal where stemming finds the root/stem word and stopword removal removes the general terms the commonly used terms.

### *D. Features and feature types*

Features play an important role in the classification process which are the terms present in the documents are obtained after performing the preprocessing of stemming and stopword removal. Features can be divided into two broad classes of on page features and features of neighbors where on page features are directly located on the page to be classified which have text representation and visual representation and the features of neighbors are found on the pages related in some way to the page to be classified [2]. Various types of features are

- Page textual content- full text, page title, headings
- Link related textual content – anchor text, extended anchor text, URL strings
- Page structural information- #words, #page outlinks, inbound outlinks (links that point to its own company), outbound outlinks (links that point to external web site) [5].

### *E. Feature Selection Methods*

Features are selected from the extracted features to reduce the dimensionality of the feature space and the methods used are

- Human judgement /use of domain lexicon
- Feature ratios and thresholding
- Frequency counting/MI [5].

### *F. Classification Methods*

Web page classification can be classified into the following broad categories of

- Supervised or so called manual categorization is useful when classes has been predefined
- Unsupervised or clustering algorithms can group web documents without any predefined framework or background information. Most clustering algorithms such as K- means need to set the number of cluster in advance.
- Meta tags based classification- using meta tag attributes for web documents classification
- Text content based categorization – a database of keywords in a category is prepared and commonly occurring words called stop words are removed from this list and the remaining words can be used for classification.
- Link and content analysis or hub-authority analysis- The link-based approach is an automatic web page categorization technique based on the fact that a web page that refers to a document must contain enough hints about its content to induce someone to read it [6].

## IV. WEB PAGE CLASSIFICATION PROCESS

Web page classification is a task of assigning a class label from predefined classes to a given document. Binary classification has two pre-defined classes whereas multi-class classification has more than two classes. It categorizes a web page into various classes, such as news, blog, shopping, adult and review, based on the contents of the page and link information such as in-link (i.e., which pages point to this page) and out-link (i.e., urls to which this page points) [7]. The process involved in web page classification is shown as a framework in figure 4.

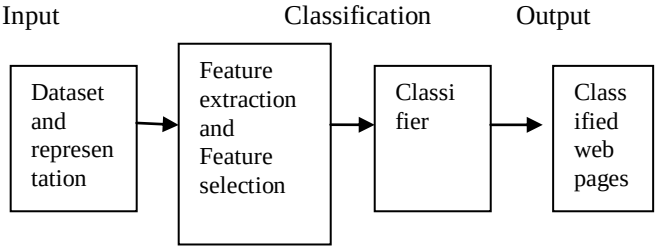


Figure 4 Web page classification

Figure 4 shows the various phases employed in web page classification process of dataset the web page collection, extraction and selection of features and then classification using classifiers are detailed in the following section.

### *A. Dataset and representation*

The first phase of the web page classification process is the generation of dataset the web pages collected from standard datatset like WebKb, ODP, DMoz etc. or from artificially generated dataset by crawling the web using a webcrawler. The collected web pages the documents are then represented as one of the representation of bag of words, set of words and N gram representation. The process involved in generating artificial dataset is described here. The webcrawler crawls the web and collects the web pages from web till the stopping criteria is met. The documents are crawled from the web using webcrawler for classification and the web crawler WebSPHINX (Website-Specific Processors for HTML INformation eXtraction) is used to collect web pages from the web and are stored in a local

repository. Crawling begins with by feeding the web crawler a set of seed pages which are a list of uniform resource locators (URLs) will start the crawling process [8]. The crawler parses the web page to extract the hyperlinks of incoming and outgoing for further crawling and then stores the crawled web pages into the local repository. Standard URL normalization is performed on extracted hyperlinks the URLs to identify equivalent URLs which link to the same web pages and will not be included in the to-crawl list of URLs for further crawling and this process is continued till the stopping criteria of the web crawler are met which is the number of web pages downloaded or the total file size will be used as the indicators to stop crawling [9].

***B. Feature Extraction and Feature Selection phase***

The second phase of the web page classification is the feature extraction and feature selection phase. In feature extraction phase, the web pages are tokenized split into terms to extract the features. The terms the features are then extracted from the web pages after performing the preprocessing of stemming and stopword removal where stemming finds the root/stem word and stopword removal removes the commonly used terms. The process of stemming and stopword removal are presented here.

***1) Stemming***

Stemming is used to find out the root/stem of a word. For example, the words user, users, used, using all can be stemmed to the word "USE". The purpose of stemming is to remove various suffixes, to reduce number of words, to have exactly matching stems, to save memory space and time.

***2) Stop word Removal***

Stop words are language specific functional words which carry no information, occur frequently in the language of the text and may be of the following types such as pronouns, prepositions, conjunctions and not useful for classification. Stop words are commonly used words frequently filtered from text in information retrieval tasks. When removing stop words, noise are get ridded of and space is saved to store documents. For example, consider an instance ""I am a student of computer science at Wisconsin University." The stopwords "I", "am", "a", "of", and "at" are left out of the full-text index. Thus on removal of the stopwords the instance is represented by "student computer science Wisconsin University" [10].

***C) Feature Selection***

Features can be selected from the extracted features using various techniques like TF, DF, TFIDF, MI, IG etc., to reduce the dimensionality of the feature space is given below.

- Term Frequency (TF) is the number of times a term appears in a document
- Document Frequency (DF) is the number of documents in which a term occurs in a dataset.
- TFIDF is the product of the local term TF and the global term IDF
- Information Gain (IG) is the number of bits gained, for a certain category, by knowing the presence or absence of a word in the document.
- Mutual Information(MI) measures the mutual dependency between the word and the category.

For classification of web page accurately in respective class it is needed to identify the content of web pages. Once the content of web page is identified it can be decided easily that a web page belong to which category. Web page usually written in HTML form and those html tags are content of web pages they are also called feature of web page. Feature selection is important process for classification. Feature should be relevant, non redundant, noise free for accurate web page classification. There are some approaches available for feature selection from web pages are as follows filter approach, wrapper approach and swarm based optimization algorithm [11,12]. Filter approach is based on applying scoring method to evaluate effective feature from dataset of web pages for example document frequency, Chi Square and information gain etc. while wrapper approach wrap the feature around the classifier to classify data to anticipate the benefit of adding or removing certain feature from training data. Swarm based optimization algorithm has inspired from nature and are very effective in selecting the best effective features from web page [13]. All possible terms in the collection of words after pre-processing do not have the same importance in deciding the subject of a document. Normally, when the frequency of a term goes down, its influence in deciding the subject also descends respectively. Therefore, it is worth selecting only a limited amount of keywords for deciding the subject is known as feature selection. The selection criterion for the terms is based on their term frequencies while the size of the feature space (number of selected terms) is based on an experimental method.[14]. Feature selection is performed to improve the performance of the web page classification and various feature selection techniques like Term occurrence number, Term Frequency, Document Frequency, Inverse Document Frequency, Information Gain, Mutual Information etc., are used to select the features to reduce the dimensionality of the feature space and to improve the quality of the web search .

***D) Classification***

The last phase of the web page classification is the classification phase of a classifier classifies the web pages into its right category and classifiers like Naïve Baye's classifier, KNN classifier, SVM classifier, Decision tree classifier, Neural Network classifier, MapReduce classifier etc., are used for web page classification where each classifier have its own algorithm for classification of web pages.

The web page classification is of supervised machine learning in which the categories are predefined or unsupervised in which clustering is performed based on the similarity of the documents. In the supervised machine learning, a classifier is trained as a model with k-n fold of the data the training data and then the trained model is used to classify the remaining n fold of the data the test data and the trained model classifies the unlabelled web pages into its right category.

## V. EVALUATION METRICS

The performance of the classifier is evaluated on test data with various evaluation metrics of precision, recall, F-measure, confusion matrix, micro average, macro average etc., are given below.

$$\text{Precision} = \frac{\text{No of pages retrieved and relevant}}{\text{No of pages retrieved}}$$

$$\text{Recall} = \frac{\text{No of pages retrieved and relevant}}{\text{No of pages relevant}}$$

$$\text{F-measure} = \frac{2 * R * P}{R + P}$$

Confusion matrix

| | | Predicted y | Predicted n |
|---|---|---|---|
| Actual | y | TP | FN |
| | n | FP | TN |

Micro average and Macro average

Macro average is the average of F1 measure of all categories and equally weights all categories where Micro average calculates recall and precision of the whole dataset and then finds the F1 and equally weights all the documents.

## VI. CONCLUSION

The uncontrolled nature of the web content creates a need for web page classification to assist the web information retrieval task. Here, the applications of the web page classification are presented where the web page categorization have great impact in the field of web applications and the web page categorization process with dataset and its representation, feature extraction and its selection and then classification using classifiers are detailed.

## REFERENCES

[1] Prabhjot Kaur, "Web content classification : a survey", *IJCTT* Vol.10, No.2, Apr 2014.

[2] Xiaoguang Qi and Brian D. Davison, "Web page classification: features and algorithms", *ACM Computing Surveys*, Vol. 41, No.2 , Article 12, February 2009.

[3] Brian Pinkerton, "Webcrawler: Finding what people want", *Technical Report*, University of Washington, November 2000.

[4] Pikakshi Manchanda, Sonali Gupta, Komal Kumar Bhatia, "On The Automated Classification of Web Pages Using Artificial Neural Network", *IOSR Journal of omputer Engineering* (IOSRJCE), Volume 4, Issue 1, PP 20- 25, Sep-Oct. 2012.

[5] Wingyan Chung, Hsinchun Chen, Edna O.F.Reid, " An automatic classification approach to business stakeholder analysis on the web", *https://ai.arizona.edu/sites/ai/files/MIS510/21_stake_wingyan.ppt* January 16, 2003.

[6] Yi Cheng, Jianye ge, Jun Liang, Sheng Yu, " Comparison of web page classification algorithms", *sers.softlab.ntua.gr/.../*

[7] Soo-Min Kim, Patrick Pantel, Lei Duan, and Scott Gaffney, "Improving Web Page Classification by Label-propagation over Click Graphs", *CIKM'09*, Hong Kong, China, November 2–6, 2009.

[8] Gautam Pant, Padmini Srinivasan, Filippo Menczer, "Crawling the Web", *Web Dynamics*, pp. 153 – 178, 2004.

[9] Lay-Ki Soon, Yee-Ern Ku , "Web Crawler with URL Signature – A Performance Study", in 4th Conference on *Data Mining and Optimization (DMO)*, Langkawi, Malaysia, 02-04 September 2012.

[10] B. Leela Devi, A. Sankar, " Feature Selection for Web Page Classification Using Swarm Optimization", *International Journal of Computer, Electrical, Automation, Control and Information Engineering* Vol:9, No:1, 2015.

[11] M. Aghda, N. G. Aghaee and M. Basiri, "Text feature selection using ant colony optimization", *Expert Systems with Applications*, vol. 8, no. 22, 2008.

[12] C. Chen, H. Lee and Y. Chang, "Two novel feature selection approaches for web page classification", *Expert System with Applications*, vol. 36, pp. 260-272, 2009.

[13] Shashank Dixit and R. K. Gupta, "Layered Approach to Classify Web Pages using Firefly Feature Selection by Support Vector Machine (SVM) ",*International Journal of u- and e- Service, Science and Technology* Vol.8, No.5 , pp.355-364 , 2015.

[14] Chaaminda Manjula Wijewickrema, "Impact of an ontology for automatic text classification", *Annals of Library and Information Studies*, Vol. 61, pp.263-272, December 2014.