IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

A Texonomy on Web Page Categorization

Bhavana[1], Neeraj Raheja[2]

[1]CSE Department, Maharishi Markandeshwar (Deemed to be) University, Mullana, Haryana, India
[2]CSE Department, Maharishi Markandeshwar (Deemed to be) University, Mullana, Haryana, India

*Corresponding Author: er.bhavanachoudhary@gmail.com

Available online at: www.ijcseonline.org

Accepted: 24/Jan/2019, Published: 31/Jan/2019

***Abstract* -** Web Page Categorization becomes essential due to the increase in the information on the Internet. As pages on the web are growing regularly and can cover almost all types of information. However finding accurate and useful information from these large amounts of web pages for a user is difficult, so efficient and accurate methods for categorizing this large of information is very necessary. Web page categorization is to categorized web pages into specified categories. It improves the efficiency of search on the web. This paper discusses various methods, approaches & uses of web page categorization.

***Keywords* -** Web Page Categorization, Web Mining, Web Content Mining, Naive Bayes, KNN, SVM

I. INTRODUCTION

Web Mining is used for finding important, useful and hidden information from the web. Web Mining discovers new and useful patterns from web activities and from web documents. Web Mining extracts relevant information; represent new knowledge from relevant data, information personalization and getting information about user and several other [1]. Web Mining is of three types: Web Content Mining, Web Usage Mining and Web Structure Mining.

Web Content Mining discovers relevant knowledge from the contents of the web such as the text, pictures and videos [2]. This huge volume of web contents creates problem in getting relevant information from the web for a particular user. So Web Page Categorization is used for categorizing this large amount of information. Web page Categorization improves the efficiency of web search and helps in information management and information retrieval [3]. It is also used for filtering of information. Web page categorization is to assign most appropriate label to a specified document [4]. Web Page Categorization is done by two methods [5]: The manual method and automated method of web page categorization. In manual method of web page categorization, experts assign web pages to appropriate category manually, but this task is impossible today because of huge volume of web information and it consumes lot of time and effort [6]. In Automatic method, web pages are automatically assigned to specified category by using different categorization algorithms.

A. Web Page Categorization Process

Web Page Categorization consists of following phases:

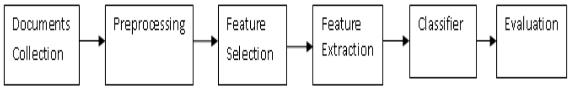

Figure 1. Web Page Categorization Process

a. *Documents Collection:* This phase collects different types of documents .html, .pdf, .doc, web content etc.
b. *Preprocessing:* Preprocessing of web pages, removes HTML tags, Stop words, rare words from the web pages and do word stemming.
c. *Features Selection:* It removes redundant features from the web pages. Document frequency, Mutual information, term- weighting are the techniques used for feature selection.
d. *Feature Extraction:* It extracts desirable features (which are much smaller than original set of features).
e. *Classifier:* Different methods and algorithm such as machine learning and classification algorithms are applied to produce the categorization such as Bayesian classifier, Decision tree, K nearest neighbor, super vector machine etc.
f. *Evaluation:* It evaluates classifiers on the basis of different factors such as recall, accuracy, precision, error etc and report experimental results.

B. Types of Web Page Categorization

Web Page Categorization is also known as Web Page Classification. On the basis of problems Web Page Categorization can be categorized into:

a. *Subject Categorization:* It categorized web pages on the basis of subject or topic such as judging whether a page belongs to Physics, Chemistry, or Computer.

b. *Functional Categorization:* It categorized the web page on the basis of the role a Web page plays such as decided whether a page is examination page, homepage, or information page.

c. *Sentiment Categorization:* It categorized the Page on the basis of the opinion that a web page have. For example attitude of an author for a particular subject or a topic.

On the basis of the classes in the problems, categorization can be of two types:

a. *Binary Categorization:* It categorized the web pages into exactly one of the two classes.

b. *Multiclass Categorization:* It categorized web page into one of more classes. It works with greater than two classes.

Multiclass Categorization can be further categorized into single label and Multilabel Categorization.

a. *Single label Categorization:* In this each instance is assigned by only one class label.

b. *Multilabel Categorization:* In this an instance is assigned by more than one class label.

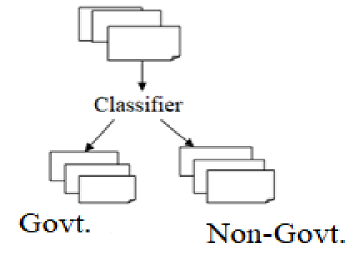

Figure 2. Binary Categorization

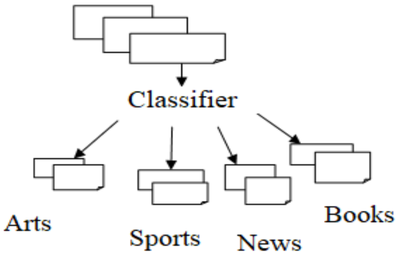

Figure 3. Multiclass Single Label Categorization

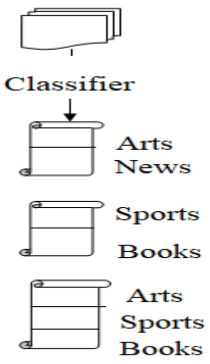

Figure 4. Multiclass Multi Label Categorization

According to the organization of classes, web page categorization is of two types.

a. *Flat categorization:* In this categories are organized in parallel, i.e., no category supersede the other category.

b. *Hierarchical categorization:* In this the categories are arranged in a hierarchical manner and each category further has subcategories.

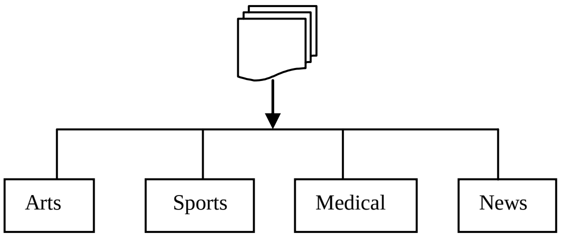

Figure 5. Flat Categorization

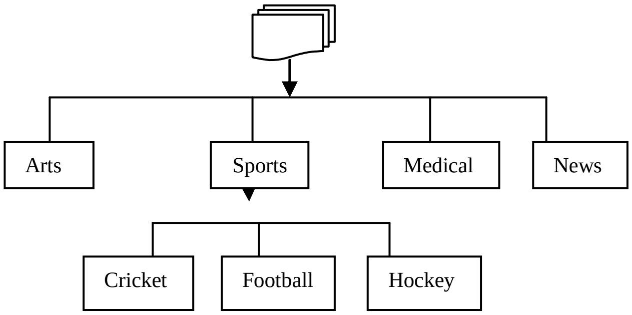

Figure 6. Hierarchical Categorization

C. *Web Page Categorization Algorithms*

Web Page Categorization is generally implemented through following algorithms:

a. *Naive Bayes Algorithm:* Due to its simplicity and efficiency, Bayesian classifier is widely used for Web Page Categorization. This algorithm applies the naive assumption and Bayes theorem. Based on the MAP rule (maximum a posteriori), categorization decision is made by it. It incorporates three distribution models into Bayesian framework. These models are Bernoulli model, Multinomial model and Poisson model that will results in Bernoulli Naive Bayes classifier, Multinomial Naive Bayes classifier and Poisson Naive Bayes classifier [7].

b. *K nearest neighbor Algorithm:* It is an instanced based learning method. It uses a common technique to represent a document. Every document is represented be a feature vector in the form of (d_1…….d_m), where d_i represents the ith feature of the document and m represents the total no of features in the document. Different weighting schemes have been introduced to determine the weights such as term frequency, TF-IDF and binary weighting. Here specific word is represented by a feature [8].

c. *Support Vector Machines (SVM):* SVM is a supervised learning method based on theory of Statistical Learning. It looks at the data and sorts into one of two categories. It is used with high dimensional feature spaces. In the linear-separable case, SVM finds the hyper plain such that margin is optimal. In the non linear separable, SVM maps the input space into a feature space based on the kernel method [9].

Other Web page Categorization algorithms [10] are **Decision Tree method (DT), Centroid Classification method [11], Multi-layer perceptron, Back-Propagation Network** etc. The classifier that uses neural network gives more promising results [12].

D. *Pros of Web Page Categorization*

a. *Improves Efficiency of Search Results:* It provides efficient result for a user query. Results of a search are presented to the user in a ranking order. Providing classified or clustering result to the user is more useful.

b. *Constructing and expending Web directories:* Web Directories gives an efficient way for searching the information within specified categories.

c. *Question Answering System:* This system uses categorization approaches to provide the meaningful answers. These approaches use the topical information from the web pages to increase the quality and accuracy of the web search.

d. *Web Content Filtering:* The text categorization is used for filtering mails, categorized web pages and organized the web pages provided by the browsers.

e. *Web Information Management:* Web Page Categorization is used for proper management of information. It provides accurate and relevant information to user.

f. *Web Information Retrieval:* When a user makes a query, Information Retrieval is used for indexing, storing, managing and accessing relevant or useful information [13].

E. *Research Issues*

Web Page Categorization is used to resolve these issues which are arise while extracting information from web content:

a. *Irrelevant information:* Difficulty in finding new, hidden and relevant information from huge database of web.

b. *Processing Capability:* There are some fields in which timely and fast processing of information are necessary such as business and artificial intelligent. So we need to process huge volume of data in short period of time.

c. *Heterogeneity:* Can be used to handle heterogeneous types of data available over the web

d. *High rate of growth of information***:** Used to handle large amount of data available over the web.

e. *Unpredictability:* Sources for web are changes continuously. In application of real world there is requirement for maintaining these changes. Thus to handle these unpredictable web sources is a big issue.

f. *Noise:* Noise means content which does not contain useful information. Removing of the noise from the web pages is necessary for providing efficient web data.

Rest of the paper is organized as follows. Section II Contains the literature survey, Section III contains conclusion and future work.

II. LITERATURE REVIEW

D. Kimet et al. [14] presented multiple co-training approach for increasing the efficiency and accuracy of web page categorization using various document representations. Three different document representation approaches (TF–IDF, LDA, and Doc2Vec) are used for converting an unstructured document into a structural numerical vector.

A. Diaz et al. [15] presented the automatic document classifier that finds the documents which are similar, in which every document contains a set of keyword. Genetic algorithm enhances the index of a group and gives optimum features.

H. Jeong et al. [16] presented an approach that consists of techniques for text summarization and classification (1) a feature weighting method are used for summarization, which combines the features that are distributed in each category, (2) an improved text classification technique which is based on summarized information (3) a framework that combines both techniques.

Qi Luo [17] presented approach for the submission of papers management system based on document categorization. These will sort papers into classes. This management system automatically categorized the submission papers and published papers into their right subject.

J. Moore et al. [18] Presented clustering techniques which clusters the documents. Clustering techniques is useful for the retrieval of documents, document filtering and categorizing documents on the Web. And it fetches the desirable features from the documents.

S. Roy et al. [19] presented a method for classifying the videos into different categories. A rough and fuzzy method is used which works on the irregular or distorted shapes of edge components of scene videos and then classify these components into different clusters. Then a feature matrix is constructed by extracting the features from inter and intra planes of different groups. Feature matrix is given to neural network for its categorization.

H. S. Gowda et al. [20] presented a method for categorizing text documents on the basis of learning algorithm. Unlabeled text documents are labeled using this method. K-means algorithm partition the data into labeled and unlabeled one. The algorithm is then applied to every partition until a higher level partition is produced which means a single class contains related labeled documents. After that desirable clusters are formed, and then nearest neighboring rule are applied for categorizing the unknown text document.

A. Qazi et al. [21] Presented a method that is used for Web Document Categorization which is based on ontology, weighting scheme and classifier. Ontology and term weighting that is used to extract the features which increase the performance of web document categorization. Classifier computes the accuracy of the term weighting approach.

D. Lopez Sanchez et al. [22] Presented a technique that uses deep learning for web categorization and to overcome the problem that requires the large training set and computational cost for the training of this data set transfer learning is used. The proposed method consists of different modules: The first module fetches the all relevant contents available on the web pages and then 2nd module uses the feature extractor and then final classifier is used to that forms web category label.

Table 1. Comparison of Web Page Classification Approaches

Approach	**Classifier**	**Reported Improvement**	**Document Representation**	**Feature Selection Criteria**	**Evaluation Dataset**
K. Donghwa et al [14]	Naive Bayesian, random forests	Accuracy Improved	Bag of words, Word frequencies and semantic information	TF–IDF, LDA, Doc2Vec	Reuters-21578, Newsgroup, Ohsumed
A.Diaz et al [15]	Genetic algorithm	Average of 85% (Precision)	Set of keywords	Term Relevance	Three Data sets with different data
Hyoungil Jeong et al.[16]	SVM	0.879,0.784, 0.894 (F1-measure)	Terms, Words	TF-IDF	Newsgroup, KORDIC, AbleNews
Qi Luo[17]	SVM	N/A	BOG	Max-Frequency	Submission and Published papers
Jerome Moore et al. [18]	Bayesian	N/A	Bag of words	TF-idf	98 web pages in 4 categories
S. Roy et al.[19]	K mean Clustering	N/A	Videos	Inter Plane Feature Extraction	dataset for 10 classes,
H. S. Gowda et al.[20]	KNN	N/A	Unlabeled text	Term Class Relevance Measure	20Newsgrup
A. Qazi et al.[21]	Multi Class Classifier(Knn, Naïve Bayes,Decision Tree, Nearest Centroid)	From 83 to 93% (F-Score)	Text	Ontology based Term-Weighting Scheme	web pages belonging to four distinct categories
D. L. Sanchez et al. [22]	KNN,SVM, LR,Perceptron	0.56, 0.96,0.9,0.93 (Accuracy rate)	Images	Deep Feature Extractor (DCNN)	images extracted from 75 web sites
Bo Tang et al.[7]	Naïve Bayes	Improves Accuracy (F-measure, G-mean)	Bags of words	Information Gain, Maximum Descrimination	Newsgroup-20, Reuters
M. B. Revanasiddappa et al. [10]	MCNN	0.193(Reduced Error Rate)	Text	RLPI	Reuters-21578
S. Shinde et al.[23]	SVM	Improved Accuracy	Text	TF-IDF	Dataset- n training instances & 1000 testing instances

Comparison Conclusion: This table defines different methods of Web Page Categorization given by different authors. Documents are represented by set of words, Bags of words, Phrases, Images, Videos. Various feature selection and classifiers are used that gives different results. Classifiers are evaluated by various factors such as F-measure, Precision, Error rate, G-Mean etc. Experiments are conducted on various datasets. From above comparison it is concluded that neural based classifiers gives optimized and promising result.

III. CONCLUSION

Web page Categorization provides efficient information to the users. It assign most appropriate label to the unlabelled documents. It improves the search results and helps in information management, retrieval and filtering of information. Web categorization converted the web pages that consist of characters, pictures, tags and hyperlink into feature vector. This feature removes contents that contains less useful information and fetch desirable features from the web pages. In future we will categorize web pages based on textual content and visual content.

REFERENCES

[1] Blockeel, R. k. " Web Mining Research:A survey". Vol. 2, PP. 1-15, 2000.

[2] R. Jain and Dr. G. N. Purohit," Page Ranking Algorithms for Web Mining",International Journal of Computer Applications, ISSN: 0975 – 8887, Vol. 13, No.5, pp. 22–25, 2011.

[3] Xiaoguang Qi and Brian d. Davison, "Web Page Classification: Features and Algorithms" ACM Computing Surveys, Vol. 41, No. 2, Article 12, 2009.

[4]P., R.B. Plastino, A. Zadrozny, B. and L.H. Merschmann, "Categorizing feature selection methods for multi-label classification", Artificial Intelligence Review, 49(1): 57-78, 2018.

[5] A. Osanyin, O. Oladipupo and Ibukun Afolabi, "A Review on Web Page Classification", Covenant Journal of Informatics & Communication Technology, Vol. 6, No. 2, Dec. 2018.

[6] S. Dixit, & 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.

[7] B. Tang, H. Haibo, M. Paul, " A Bayesian Classification Approach Using Class-Specific Features for Text Categorization", IEEE ,2015.

[8] W. A. Awad, "Machine Learning Algorithms in Web Page Classification", International Journal of Computer Science & Information Technology (IJCSIT), Vol. 4, No. 5, 2012.

[9]T. Joachims, "Text categorization with support vector machines: Learning with many relevant features", In: Proceedings of European Conference on Machine Learning E, CML, vol. 1398, pp. 137–142, 2000,.

[10] M. B. Revanasiddappa, B. S. Harish, S. V. A. Kumar, "Meta-cognitive Neural Network based Sequential Learning Framework for Text Categorization", ICCIDS, 2018.

[11] Liu, C. Wang, W. Tu, G. Xiang, Y. Wang, S. and L, F. "A new Centroid-Based Classification model for text categorization.", Knowledge-Based Systems, vol. 136, pp. 15-26, 2017.

[12] R., S., V., S.P. "Text categorization by backpropagation network", International Journal of Computer Applications, vol. 8, No. 6, pp. 1-5, 2010.

[13] C. Chang, M. Kayed, M. R. Girgis and K. F. Shaalan, "A Survey of Web Information Extraction Systems", in IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 10, pp. 1411-1428, Oct. 2006.

[14] K. Donghwa, S. Deokseong, S. Deokseong, C. Suhyoun, K. Pilsung, "Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec", 2018.

[15] Dıaz, A. B. Rios, J. H. Barron, T. Y. Guerrero, J. C. Elizondo, "An automatic document classifier system based on genetic algorithm and taxonomy", 2018.

[16] J. Hyoungil , K. Youngong , S. Jungyun, "How to Improve Text Summarization and Classification by Mutual Cooperation on an Integrated Framework", 2016.

[17] Qi Luo, "Research on Paper Submission Management System by Using Automatic Text Categorization", Springer International Publishing AG, 2018.

[18] J. Moorey, Eui-Hong (Sam) Han, "Web Page Categorization and Feature Selection Using Association Rule and Principal Component Clustering", 2010.

[19] S. Roy, P. Shivakumara, N. Jain, V. Khare, A. Dutta, U. P. and Tong Lu, "Rough-Fuzzy based Scene Categorization for Text Detection and Recognition in Video" Pattern Recognition", doi: 10.1016/j.patcog.2018.02.014, 2018.

[20] H. S. Gowda, M. Suhil(B), D.S. Guru, and L. N. Raju, "Semi-supervised Text Categorization Using Recursive K-means Clustering" Recent Trends in Image Processing and Pattern Recognition, Springer, 2016.

[21] A. Qaziaand R.H. Goudar, "An Ontology-based Term Weighting Technique for Web Document Categorization", Science Direct, Procedia Computer Science vol. 133, pp. 75–81, 2018.

[22] D. L. sanchez, A. G. Arrieta and J. M. Corchado, "Deep neural networks and transfer learning applied to multimedia web mining", Springer International Publishing AG, 2018.

[23] S. Shinde, J. Prasanna and S. Vanjale, "Web Document Classification using Support Vector Machine", IEEE, 2017.

Authors Profile

Mrs. Bhavana has completed her M.tech from University Institute of Engineering and Technology in 2013. Now she has pursuing Ph.d from Maharishi Markandeshwar (Deemed to be) University, Mullana. She is currently working as an Assistant Professor in Dayanand Mahila Mahavidyalaya. Her research area is Web Mining, Machine Learning and Networking. She has published 4 papers in reputed journals and conferences.

Dr. Neeraj Raheja has completed his Ph.d from Maharishi Markandeshwar (Deemed to be) University, Mullana. He is currently working as an Associate Professor in Maharishi Markandeshwar (Deemed to be) University, Mullana. His research area is Web Mining and Machine Learning. He has published 20 papers in reputed journals and conferences.