Comparative Study of Web Page Classification Approaches

Pooja Vinod Nainwani
B. Tech IT student
Department of Information Technology, CSPIT,
Charotar University of Science and Technology,
Changa

Purvi Prajapati
Assistant Professor,
Department of Information Technology, CSPIT,
Charotar University of Science and Technology,
Changa

ABSTRACT
Classification of Web pages is one of the challenging and important task as there is an increase in web pages in day to day life provided by internet. There are many ways of classifying web pages based on different approach and features. This paper explains some of the approaches and algorithms used for the classification of webpages. Web pages are allocated to pre-determined categories which is done mainly according to their content in Web page classification. The important technique for web mining is web page classification because classifying the web pages of interesting class is the initial step of data mining. The agenda of this paper is first to introduce the concepts related to web mining and then to provide a comprehensive review of different classification techniques.

Keywords
Web page classification, Web Mining, Data Mining, uniform resource locator (URL), SVM (Support Vector Machine), KNN (K-Nearest Neighbor), Naïve Bayes, Artificial Neural Network.

1. INTRODUCTION
World Wide Web (WWW) is a widespread and collaborating medium with excellent growth of amount. World Wide Web has made it essential for users to operate automated tools in finding the desired information resources [2]. The World Wide Web is the collection of text files, documents, images, and other forms of data in unstructured, semi structured and structured form [11]. The Web is the largest data source in the world [10]. Classification plays a vigorous role in many information management tasks and reclamation tasks [4].

Classification of the Web pages have been considered largely since the Internet has become a huge source of information, in terms of variance and volume [1,17]. Classification being considered a supervised learning problem in which a set of labelled data is used to train a classifier which can be applied to label upcoming instances [3]. There are mainly two types of classification, one is Binary classification and the other is multi-class classification. The issue of classification has been broadly considered in the data mining, database, and information retrieval communities [12].

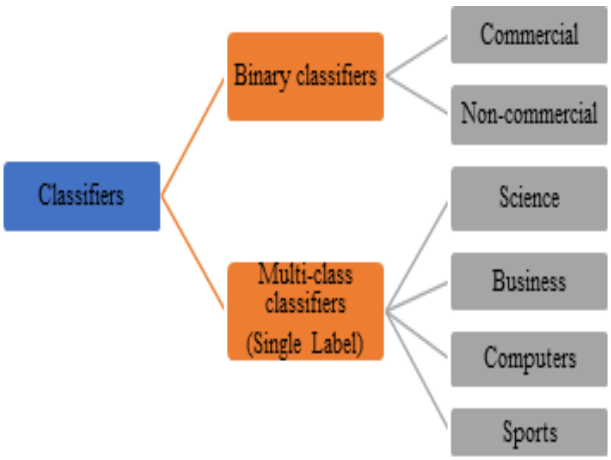

Fig 1: Types of Classification with its example.

Web mining can be generally stated as the analysis and discovery of valuable information from the World Wide Web by the mining of stimulating and convenient patterns and implied information from activity related to the World Wide Web [2].

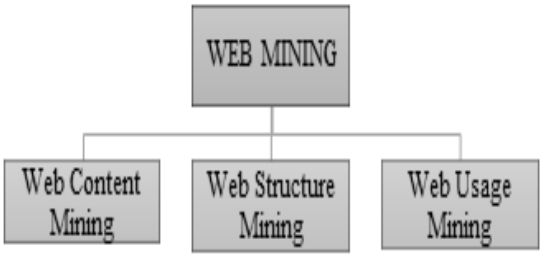

Fig 2: Types of Web Mining

Web content mining, it is the method of data mining from the content of documents or their brief descriptions, which includes Text, Audio, Images, Video and Structure Record. Web structure mining, it is the process of concluding knowledge from the World Wide Web organization and also from the links between references in the Web and it includes Hyperlink and Document structure. A Web usage mining, it is the process of extracting fascinating patterns in web access logs and it includes web server logs, Application Level logs and Application Server logs [13]. From a point of view of data mining, Web mining, has three parts of concentration: clustering (finding natural groupings of pages, users etc.), associations (which URLs tend to be demanded together), and classification (characterization of documents) [2].

Web page classification, it is the method of allocating a Web page to one or more predetermined category labels [4]. It is an

important method in terms of many information retrieval tasks like reclamation of scientific papers, e-books and digital library from the web [14]. Present webpage classification methods use a variation in information to classify a goal page: the text of the page itself, its structure of hyperlink, the link structure and anchor text from pages directing to the goal page and its place (given by its URL) [9]. The universal problematic part of web page classification can be separated into multiple sub-problems, one is the subject classification, sentiment classification, functional classification, and many other kinds of classification [4], in which subject classification mainly focusses on topic or matter of a webpage and the Functional classification is mainly concerned about the role of the web page [3]. Web page classification can be helpful in improving the quality of web search [4].

2. APPROACHES

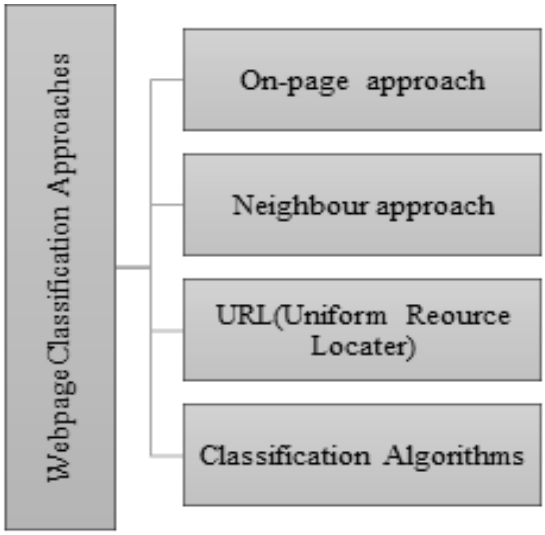

Fig 3: Different Approaches of Classification

The first approach is the on-page feature which includes written tags and content which are placed on the page itself, one of the most upfront feature that one can consider to use is the text content. One noticeable feature that does not seems in plain text documents but HTML documents is HTML tags [4]. To increase the classifier's performance, it has been confirmed that it can be done by using information which are derived from tags. Thus, by using tags we can take the benefit of the structural information surrounded by the HTML files, which is generally overlooked by methods of plain text. Nevertheless, subsequently maximum HTML tags are preoccupied with symbols rather than the semantics, web page authors might create diff erent but theoretically corresponding tag structure. Thus, using the HTML tagging information in web classification might suff er from the unreliable formation of HTML documents.

The second approach is using neighbor feature, in order to identify this problem of classifiers to make sensible decisions based on features presented on the webpage, features can be pull out from neighboring pages that are relevant in some way to the page that is to be classified to supply additional information for classification. There are many ways to originate such connections amongst pages, one of the mostly used connection is the hyperlink [4].

The uniform resource locator (URL), this approach is amounts faster than distinctive web page classification, as the webpages themselves do not have to be retrieved and analyzed. This approach divides the URL into expressive portions and adds component, sequential and orthographic features to model relevant patterns. The resultant binary features are used in supervised extreme entropy demonstrating. It is examined, the approach's usefulness in binary, multi-class and hierarchical classification. A URL is first divided into meaningful tokens using theoretic measures. This is important as few components of a URL are not surrounded by spaces (especially domain names). These tokens are then inputted into an analysis module that derives important composite features for classification. In the second step, machine learning is used to bring a multiclass or regression model from categorized training URLs that have been processed by previous module [9].

Next approach is, decision trees are mainly known for their easiness and instinctiveness. WEKA was used to develop Decision trees [6] i.e. (Waikato Environment for Knowledge Analysis) which is the collection of machine learning algorithms used for data mining errands [6]. Different classification algorithms were also considered, among them Logistic Model Tree (LMT), Best First Decision Tree (BFT), J48 Pruned Tree (J48PT) and J48 Graft Tree (J48GT). Decision tree is developed by generating the Object Attribute Table after gathering all the information which includes External links (EL), Page's text length (TL), Image (Im), Internal links (IL), Word Blog (WB), External Images (EI), Word Video (WV) Multimedia Objects (MO), Word Flash (WF), Word Image (WI), Word News (WN), Internal Images (II).

Table 1. Comparative study of Classification Algorithms

SR.NO	ALGORITHM	FEATURES	LIMITATIONS
1	K-Nearest Neighbor Algorithm	It's not mandatory that classes should be linearly separable. Sometimes its robust because of some noisy training data and can be effectively used for multimodal classes.	When finding nearest neighbor in large training dataset, it takes excessive time. It is delicate to irrelevant or noisy attributes. Number of dimensions used decides the performance of algorithm.
2	Naïve Bayes Algorithm	It has decent computational efficiency and classification rate and implementation is simple. It expects accurate outcome for the prediction problem and classification.	The accuracy of algorithm declines if the volume of dataset is less. For accurate and good results large amount of dataset is required.
3	Support Vector Machine	Accuracy is high. Works effectively even when the data is linearly or not linearly separable.	For both training and testing the requirement of speed and size is more. High difficulty and widespread memory requirements for classification.
4	Artificial Neural Network	It can be easily used by adjusting only few parameters. A neural network learns and reprogramming is not needed. Implementation is easy and applicable to large range of real life problems.	Requires high processing time is there is large neural network. Difficult to predict the number of layers and neurons needed. Learning process takes time.

3. DATA SETS

The universal data set is the internet which is a free open directory containing millions of web pages. Data sets for single label is available in this http://web.ist.utl.pt/acardoso/datasets/ and for multi-label data sets you can find here http://sci2s.ugr.es/keel/multilabel.php and there are many more data sets available on the internet.

4. RESEARCH CHALLENGES

Classification of imbalanced data is difficult as the structure of data is not consistent so for the unstructured data efficiency decreases. More number of stop words, infrequent words and punctuation symbols may be present in the web page. Few of the web pages contain video, audio and/or image information linked with them. The nature of web pages is active and unpredictable. The web page has no exclusive or selective format [15].

If there are large number of categories then the accuracy of webpage assigned to specific page decreases and becomes irrelevant [5].

The imbalance of class, rarity and large example learning issues within a web manual dataset make smearing classification algorithms on such directories very tough [7].

5. CONCLUSIONS

Classification of web pages is the significant method for Web mining because the initial step of Web mining is categorizing the web pages of different classes. Classification approaches based on method of data mining, they are useful in the web mining area in order to construct effective trees that classify web pages accurately depending on their features. After studying web classification with references of its approaches and processes, it is defined as a type of problem known as supervised that purposes to classify web pages into a set of predetermined categories based on considered training data. This paper has surveyed some existing classification techniques and its approaches. The goal of classification results to generate more certain, precise and accurate system results.

5. ACKNOWLEDGMENTS.

Our thanks to the experts who have contributed towards development of the paper Prof. Purvi Prajapati<purviprajapati.it@charusat.ac.in> for her help and suggestions thought this work.

6. REFERENCES

[1] Yu H, Han J, Chang KC. PEBL: Web page classification without negative examples. IEEE Transactions on Knowledge and Data Engineering. 2004 Jan;16(1):70-81.

[2] Fiol-Roig G, Miró-Julià M, Herraiz E. Data mining techniques for web page classification. In Highlights in Practical Applications of Agents and Multiagent Systems 2011 (pp. 61-68). Springer, Berlin, Heidelberg.

[3] Nayak MA. A Comparative Study of Web Page Classification Techniques.

[4] Qi X and Davison B.D. (2009) Web Page Classification: Features and Algorithms. ACM Computing Surveys, Vol. 41, No. 2, Article 12.

[5] S. Markkandeyan1 · M. Indra Devi, "Efficient Machine Learning Technique for Web Page Classification".

[6] Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2016 Oct 1.

[7] Kwon OW, Lee JH. Web page classification based on k-nearest neighbour approach. In Proceedings of the fifth international workshop on Information retrieval with Asian languages 2000 Nov 1 (pp. 9-15). ACM.

[8] Patil AS, Pawar BV. Automated classification of web sites using Naive Bayesian algorithm. In Proceedings of

the international multiconference of engineers and computer scientists 2012 Mar 14 (Vol. 1, pp. 519-523).

[9] Kan MY, Thi HO. Fast webpage classification using URL features. In Proceedings of the 14th ACM international conference on Information and knowledge management 2005 Oct 31 (pp. 325-326). ACM.

[10] Herrouz A, Khentout C, Djoudi M. Overview of web content mining tools. arXiv preprint arXiv:1307.1024. 2013 Jul 2.

[11] Malarvizhi R, Saraswathi K. Web Content Mining Techniques Tools & Algorithms–A Comprehensive Study. International Journal of Computer Trends and Technology (IJCTT). 2013 Aug; 4(8):2940-5.

[12] Aggarwal CC, Zhai C. A survey of text classification algorithms. InMining text data 2012 (pp. 163-222). Springer, Boston, MA.

[13] Chen H, Fuller SS, Friedman C, Hersh W. Knowledge management, data mining, and text mining in medical informatics. InMedical Informatics 2005 (pp. 3-33). Springer US.

[14] Kavitha S, Vijaya MS. Web Page Categorization using Multilayer Perceptron with Reduced Features. International Journal of Computer Applications. 2013 Jan 1;65(1).

[15] Marath ST, Shepherd M, Milios E, Duffy J. Large-scale web page classification. InSystem Sciences (HICSS), 2014 47th Hawaii International Conference on 2014 Jan 6 (pp. 1813-1822). IEEE.

[16] Yuchang Lu. "Application of SVM in web page categorization" , 2006 IEEE International Conference on Granular Computing, 2006

[17] Rongfang Bie. "Automatic web pages categorization with ReliefF and Hidden Naive Bayes" , Proceedings of the 2007 ACM symposium on Applied computing - SAC 07 SAC 07, 2007

[18] Shaohong, Chen, and Wang Zhixing. "Web page classification based on Semi-supervised Naïve Bayesian EM algorithm", 2011 IEEE 3rd International Conference on Communication Software and Networks, 2011.

[19] AbdulHussien AA. Comparison of Machine Learning Algorithms to Classify Web Pages.

[20] Gabriel Fiol-Roig. "Data Mining Techniques for Web Page Classification", Advances in Intelligent and Soft Computing, 2011.