Literature Review

S.No	Title	Authors	Year	Points
1	PEBL-Website classification without negative examples.	Hwanjo Yu, Jiawei Han, and Kevin Chen-Chuan Chang	2004	• PEBL: Positive Example Based Learning • Eliminate need for negative examples(Examples for websites that do not belong to particular category) • Uses an algorithm, called Mapping-Convergence (M-C), to achieve high classification accuracy (with positive and unlabeled data) as high as that of a traditional SVM (with positive and negative data). • M-C runs in two stages: the mapping stage and convergence stage. In the mapping stage, the algorithm uses a weak classifier that draws an initial approximation of "strong" negative data. Based on the initial approximation, the convergence stage iteratively runs an internal classifier (e.g., SVM) which maximizes margins to progressively improve the approximation of negative data. Thus, the class boundary eventually converges to the true boundary of the positive class in the feature space.
2	Data mining techniques for website classification	Gabriel Fiol-Roig, Margaret Miró-Julià, Eduardo Herraiz	2011	• Object Attribute Table(OAT) is made on the basis of statistical information of websites(such as number of videos,number of images,number of external links,number of external images,length of website...) • Among available classification methods, decision trees were selected for their simplicity and

S.No	Title	Authors	Year	Points
				intuitiveness. Decision trees were developed using WEKA (Waikato Environment for Knowledge Analysis) a collection of machine learning algorithms for data mining tasks • Decision trees were able to provide about an accuracy of 90%
3	A webpage classification algorithm concerning webpage design characterestics	Shih-Ting Yang*	2012	• Proposed model is based on analysis of tag attributes and tag-regions to search for text contained in tag-regions and extract the corresponding keywords. Based on tag attributes and specific tag-region layout of webpage, the corresponding weight values are assigned to various tag-regions. • Establishes a hyperlink webpage screening mechanism to collect the hyperlink webpage with higher correlations to slightly modify or adjust the categories of the target webpage. • Whole model include three modules 1. tag region weight assignment 2. web page category determination 3. hyperlink web page determination
4	Novel approach of website classification using feature interval	J. Alamelu Mangai, Dipti D. Kothari and V. Santhosh Kumar	2012	• An algorithm called weighted voting of feature intervals has been proposed. The classifier based on this algorithm will first discretizes the web page features using a supervised disctretization algorithm which identifies the number of intervals each

S.No	Title	Authors	Year	Points
				feature has to be discretized automatically. Each feature is then made to predict the class of the corresponding feature in the test web page using the class distribution of its intervals. The final class of the test web page is predicted by aggregating the weighted vote of each feature. • Experiments done on a benchmarking data set called WebKB has shown good classification accuracy when compared with many of the existing classifiers.
5	Automated classification of websites using Naive Bayes Algorithm	Ajay S. Patil, B.V. Pawar	2012	• Classify websites into categories assuming that home page of website will contain brief description of website. • The NB approach, is one of the most effective and straightforward method for text document classification and has exhibited good results in previous studies conducted for data mining • Accuracy about 89.05% • Classification accuracy is proportional to number of training documents upto a limit
6	Web Page Categorization using Multilayer Perceptron with Reduced Features	Kavitha S,Vijaya MS	2013	• Demonstrates the web page categorization problem as the multi classification task and provides a suitable solution using a supervised learning technique namely multilayer perception. • The multilayer Perceptron network is the most widely used neural network classifier. It is a feed

S.No	Title	Authors	Year	Points
				forward artificial neural network model that maps sets of input data into a set of appropriate output. It is a varation of the standard linear perceptron in that it uses three or more layers of nodes with nonlinear activation functions and is more powerful than the perceptron in that it can distinguish data that is not linearly separable, or separable using hyperplane. MLP networks are universal, flexible and nonlinear models consisting of a number of units organized into multiple layers. • Classification accuracy is about 5% more than Naive Bayes classifier. How ever it take too much time as compared to Naive Bayes classification method.
7	Fast webpage classification using URL features	Min-Yen Kan,Hoang Oanh Nguyen Thi	Not specified	• Demonstrates usefulness of url alone in website classification • URLs are often meant to be easily recalled by humans, and websites that follow good design techniques will encode useful words that describe their resource in the website's domain name as advocated by best practice guidelines • Much faster compared with other method as there is no need for fetching whole webpage content. • This approach splits the URL into meaningful chunks and adds component, sequential and orthographic features to model unique patters. • Sometimes this performs better than full-text

S.No	Title	Authors	Year	Points
				methods. • URL features can be used to predict prestige of a webpage too. • Two step machine learning approach is used 1. A URL is first segmented into meaningful tokens using information-theoretic measures. This is necessary as some components of a URL are not delimited by spaces (especially domain names). These tokens are then fed into an analysis module that derives useful composite features for classification. These features model sequential dependencies between tokens, their orthographic patterns, length, and originating URI component. 2. machine learning is used to induce a multi-class or regression model from labeled training URLs that have been processed by the above pipeline. New, unseen test URLs can then be classified by processing them first to extract features, and then applying the derived model to obtain a final classification. A key result is that the combination of quality URL segmentation and feature extraction results in a significant improvement in classification accuracy over baseline approaches
8	Web Page Classification with an Ant Colony Algorithm	Nicholas Holden	Not specified	• Make use of ant-miner algorithm.(An ant colony algorithm for classification) • A data structure called wordnet is used which is

S.No	Title	Authors	Year	Points
				an electronic lexicon that contains relationships between words in a tree like structure. It is an attempt to map human understanding of words into an electronic database • Concentrate on RSS filed to which give concise, accurate and accessible information about many web pages.
9	Hierarchical Classification of Web Content	Susan Dumais,Hao Chen	Not specified	• Explores the use of hierarchical structure for classification • Make use of SVM classifiers • The use of a hierarchical decomposition of a classification problem allows for efficiencies in both learning and representation. • First divide website into a toplevel category then divide into sub levels.
10	Web Page Classification Based on k-Nearest Neighbor Approach	Oh-Woog Kwon and Jong-Hyeok Lee	Not specified	• Employs k-NN approach with a feature selection method and a term-weighting scheme using markup tags, and reform document-document similarity measure used in vector space model. • The k-NN approach can be broken down in two steps. Given a test document, the first step is to find k nearest samples among the training documents, using document-document similarity. The similarity score of each neighbor document to the test document is used as the weight of the categories pre-assigned to the training document.

S.No	Title	Authors	Year	Points
				The second step is to estimate the likelihood of each category by summing the weight of the category of the k nearest documents • Involves feature selection(Selection of words capable of deciding category) • Term weighting- Assign more weight age to terms given in powerful html tags such as <meta>,<titlle>,<head>....etc • A new similarity measurement approach is proposed.