# Introduction *(Heading 1)*

THE World Wide Web (WWW) started in the year 1991 and has gained large popularity day by day today more than 3.5 billion users are using internet. In the recent time we have seen internet penetration of more than 35% all over the world. Netcraft’s January 2012 survey, estimates that more than 584 million web sites exist on the internet and out of which, nearly 175.2 million are active. The number of users using the Internet is rapidly increasing. Internet World Stats reveals that the world Internet usage growth has increased by 480.4% during 2000-2011. Today there are many varieties of tools available for an average internet user to locate and identify information on the internet. These tools are broadly classified as 1.Crawler based Search Engines (SE) e.g., Google, Bing, Yahoo,Duck Duck Go etc, 2.Meta Search engines e.g. Metacrawler, Clusty etc. and 3.Subject Directories like DMOZ (Directory Mozilla), Librarians Internet Index (LII) etc. The crawler based search engines and subject directories maintain their own data repositories, whereas meta search engines don’t maintain any such data repositories, instead they depend on indices of other Search engines and subject directories to answer user queries. The database maintained by any crawler based search engine is quite large, and have considerably larger amount of data indexed in their databases as compared with subject directories. Directory Mozilla (DMOZ) [1] i.e., dmoz.com has 93,446 editors for 1.2 million categories and has indexed 5.26 million websites, which is only 2.5% of the total active web sites available on the Internet today. These directories like DMOZ need to be manually edited and maintained manually. Subject directories are popular due to proper classification of data in several categories. A larger website directory could be quite helpful in improving the quality of search results, filtering web content, developing knowledge bases, building vertical (domain specific) search engines. Hence need for automating the process of classification of websites based on their content arisen recently. Manually classifying data is really expensive to scale as well as quite labor intensive and in this paper we present a technique to reduce manual effort significantly and hence this paper presents a cost effective way for categorizing data.

This paper presents a Naïve Bayesian (NB) probabilistic model for the automatic classification of web sites based on content of their webpages. This model, is one of the most effective and straightforward model for text document classification and has exhibited good results in previous studies conducted for data mining. The model is quite optimized and has quite effectively worked to classify websites based on their content in real-time. The presented model has offered accuracy of approximately 92% for the test data given.

**The rest of the paper is organized as follows. Section II reviews previous work on the machine learning and classification. Section III and IV discusses the classification of web pages and Naïve Bayes Theorem respectively, Section V presents our approach of classifying websites based on home pages using NB technique. Section VI discusses the results of our experiment. The last section summarizes the paper and gives some directions for future research.**

# Related Work

This section briefly reviews related work on text classification with special emphasis on classification of web pages. In the early days, classification was done manually by domain experts. But very soon, classification was also carried out in semi-automatic or automatic manner. Some of the approaches for text-categorization include statistical and machine leaning techniques like k-Nearest Neighbor approach [2], Bayesian probabilistic models [3]-[4], inductive rule learning [5], decision trees [4],[6], neural networks [7],[8] and support vector machines [9],[10]. While most of the learning methods have been applied to pure text documents, there are numerous publications dealing with classification of web pages. Pierre [11] discusses various practical issues in automated categorization of web sites. Machine and statistical learning algorithms have also been applied for classification of web pages [12]-[15]. In order to exploit the hypertext based organization of the web page several techniques like building implicit links [16], removal of noisy hyperlinks[17], fusion of heterogeneous data[18], link and context analysis[19] and web summarization[20] are used. An effort has been made to classify web content based on hierarchical structure [21].

# Classification of Web pages

Classification of web content is different in some aspects as compared with text classification. The uncontrolled nature of web content presents additional challenges to web page classification as compared to traditional text classification. The web content is semi structured and contains formatting information in form of HTML tags. A web page consists of hyperlinks to point to other pages. This interconnected nature of web pages provides features that can be of greater help in classification. First all HTML tags are removed from the web pages, including punctuation marks. The next step is to remove stop words as they are common to all documents and does not contribute much in searching. In most cases a stemming algorithm is applied to reduce words to their basic stem. One such frequently used stemmer is the Porter’s stemming algorithm [22]. Each text document obtained by application of procedures discussed above is represented as frequency vector. Machine learning algorithms are then applied on such vectors for the purpose of training the respective classifier. The classification mechanism of the algorithm is used to test an unlabeled sample document against the learnt data. In our approach we deal with home pages of organizational websites. A neatly developed home page of a web site is treated as an entry point for the entire web site. It represents the summary of the rest of the web site. Many URLs link to the second level pages telling more about the nature of the organization. The information contained the title, Meta keyword and Meta description and in the labels of the A HREF (anchor) tags are very important source of rich features. In order to rank high in search engine results, site promoters pump in many relevant keywords. This additional information can also be exploited. Most of the homepages are designed to fit in a single screen. The factors discussed above contribute to the expression power of the home page to identify the nature of the organization.

[http://www.dmoz.org](http://www.dmoz.org/)[1]

[http://www.internetworldstats.com/stats.htm [1](http://www.internetworldstats.com/stats.htm%20%5b1)]

[http://en.wikipedia.org/wiki/World\_Wide\_Web[2](http://en.wikipedia.org/wiki/World_Wide_Web%5b2)]