# Conclusion and Future Works

The naive bayes document classification algorithm discussed in this paper intelligently exploits the richness of features present in the webpages of any website for effective classification into different industry type category. The algorithm here classifies the webpages into a very broad set of categories. Naive based approach for classification of websites based on webpages for six categories considered in this paper yielded a result of more than 92% accuracy. It has been observed that the classification accuracy of the classifier is proportional to number of training documents. The results are quite encouraging. This approach could be utilized by the search engines and other online directory projects like DMOZ for effective categorization of websites to build an automated website directory based on the content available on the website and type of organization. Although in this experiment, only nonhierarchical and distinct categories are considered. The above algorithm could also be used to classify any given web pages into more specific categories (hierarchical classification) by altering the set of features e.g. a web site that is ecommerce may be further classified into electronics, clothes or a book selling website.

# References

1. <http://news.netcraft.com/archives/2013/08/09/august-2013-web-server-survey.html> . Netcraft internet survey for august 2012
2. <http://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html> Stanford NLP , naïve bayes text classification
3. <http://tartarus.org/martin/PorterStemmer/index.html> Porter Stemming algorithm, with various implementations.
4. Bird, Steven, Edward Loper and Ewan Klein (2009). Natural Language Processing with Python. O’Reilly Media Inc. Python NLTK
5. [www.matplotlib.org](http://www.matplotlib.org) Python based open source mathematical analysis toolkit.
6. Automated Classification of Web Sites. Proceedings of international conference of engineers and computer scientists 2012 Vol 1 , IMECS 2012 Hong Kong
7. DMOZ open directory project. [Online]. Available: http://dmoz.org/
8. T. M Mitchell. (1997). Machine Learning McGraw-Hill Companies, Inc.
9. Data Mining concepts and techniques Han Kamber Lee Morgan Kaufman publications. 3rd Edition 2012.
10. C. J. van Rijsbergen. (1979). Information Retrieval (2nd ed.) London: Butterworths,[Online].Available:http://www.dcs.gla.ac.uk/Keith/Preface.html