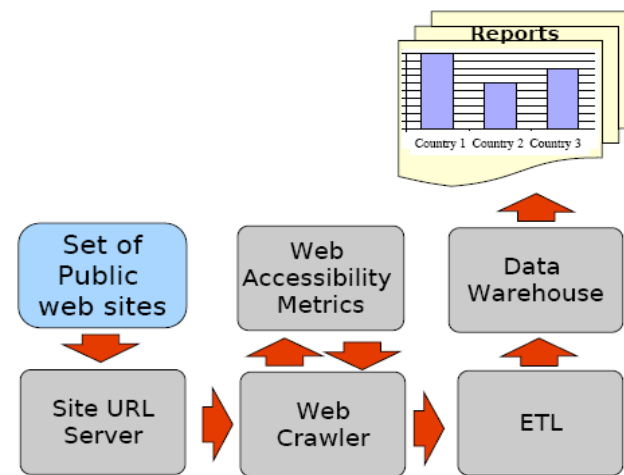


Introduction

Web mining technologies have attracted a great deal of attention in recent years. The research was included to provide effective access to the specific web content. Web sites classification could be viewed as extending the work performed by the basic search engines. There have been a lot of studies and research on the effective web information retrieval techniques, that includes data mining, clustering, and classification etc.

The purpose of this project, is to develop a application that performs automatic web sites classification for EIAO machinery. It should be performed when the web sites were downloaded in the components of "Site URL server". EIAO is European Internet Accessibility Observatory, it is established for large-scale assessment of web sites accessibility. In order to provide quality background material, Classification of Economic Activities (NACE) and The Nomenclature of Territorial Units for Statistics (NUTS) become two important specific subjects to describe specific web sites in on-line report tool of EIAO as shown in the following figure.



Proposed approach

Our solution is to remove stop words and skip html tags; use mutual information for feature selection; finally, use classifiers implemented in Naive Bayes and Decision Tree algorithm to perform the classification task. After classification, each document is assigned a class label from a set of predefined categories, which is based on a pool of pre-classified sample documents.

Remove stopword and skip html tags

Stopwords are set of words that are non-informative terms such as "a, the, of" and so on. Skip html tags is used to skip all the words in "<>" which is useful for tokenizing (X)HTML files. These methods will save vectors space and improve efficiency and accuracy of classification.

Mutual Information

It measures the associativity between terms and categories. It is used to reduce the features space without sacrificing classification accuracy and provide vectors for classification.

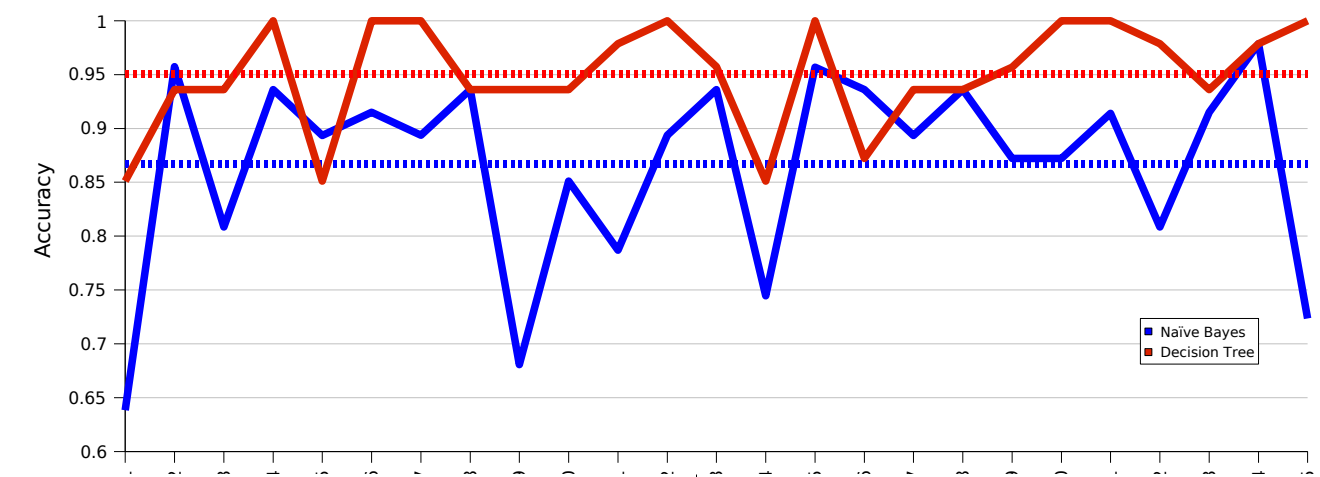
Naive Bayes and Decision Tree

Naive Bayes is typical statistical classifiers based on applying Bayes theorem in the assumption of class independence. It searches for the most probable hypothesis under consideration.

Decision Tree is a classifier that uses the given data to generate a graph or model of decision, then to determine the class label of unknown data in using the model.

Results

The following figure shows the accuracy per test run in order to



show the performance of the proposed solution in NACE code. The accuracy of Naive Bayes is in blue colour while the red colour represent the Decision Tree.

It is obvious that Decision tree has higher average accuracy than Naive Bayes as shown in the figure. The average accuracy of Decision Tree is up to 95%, while Naive Bayes has an average accuracy about 86%.

Conclusion

The result shows both naive Bayes and Decision Tree had shown good performance in automatic web sites classification. Especially Decision Tree provides great accuracy and efficiency performance during the classification task. The proposed solution offers a accurate, reliable classification tool for EIAO to evaluate the assessment of Web sites.