Website Classification

Submitted in the partial fulfillment of requirements for the award of degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER ENGG.

Submitted by
Akshay Kumar(10-CSS-06)
Niyas C(10-CSS-44)

Department of Computer Engg.
Jamia Millia Islamia
New Delhi,110025
Year-2013

Introduction

World Wide Web(www) is a large repository of information which contain a variety of information in the form of web documents. Information stored in web is increasing in a very rapid way and peoples rely more and more on internet for acquiring information. Internet World Stats reveals world internet usage has increased by 480% within the period 2000-2011. This exponential growth of the web has made it difficult to find and organize data. If we categorized data on internet it would be more easy to find relevant piece of information from internet. There are some popular web directories such as yahoo directory and mozilla directory in which web sites are organized according to their category. The proper classification has made these directories popular among web users. However these web directories make use of human effort for classifying web sites. Rapid growth of web has made it increasingly difficult to classify web sites manually. Web site classification using machine learning algorithms has become a major research topic in these days. A number of algorithms has been proposed for the classification of web sites by analyzing it's features.

In our project we would like to develop a desktop application that will predict the category of a web site (such as arts, sports, politics, fashion, technology, ...etc) by taking it's URL as input. Desktop application will also contain features to find it's accuracy in prediction.

Technical Details

1. **Web Site categorization**

   Most important and challenging task of this project is classification of web site according to it's textual content. Classification of web site is different in some aspects as compared with text classification. It is different because of uncontrolled nature of web content. The web content will be semi structured and will contain formatting information in the form of html tags. Also web sites will be containing hyper links to other web pages. This interconnected nature also make it difficult to classify web pages. In our project we will use Naive Bayesian algorithm for web site classification. We use the content of home page of a web site to predict the category with the assumption that web developer might have given a brief description regarding web site on the home page. HTML tags that provide extra power to words such as ,<bold>,<title>,<head> will be treated with special consideration. All stop words will be removed as they contribute nothing to category of web site.

2. **Implementation Details**

   - whole coding will be done using python(An interpreted language which provide several built-in text processing functions)
   - Training set will be stored in some database softwares like marianDB.
   - GUI of app will be designed using pygtk library.

Area of Application

Proper classification of web sites has numerous number of application in day to day life.

- Classification of web sites allow to make web directory projects (services that allow us to browse through different categories) automatic and to make the process faster and cheaper. This will help search engines to provide more relevant piece of results in response to user searches.

- Web site/page classification allow machine learning systems to collect necessary information from internet in a more easy way. For example if a system want to collect all news regarding stock exchange,then it can make use web page classification system.

- There will be software that controls proxy servers(For example cyberoam in Jamia). Automatic web site classifiers will allow these software to decide the category of a web site before user see it, and facilitate optional blocking. In big organizations like universities , this will help to prevent students/workers from wasting time and bandwidth.

- Younger generation may loss the rhythm/order of their life because of technology on finger tip. Classification of web sites will allow to prevent youth/children from misuse of technology. Some of the morally progressed nations like modern Turkey impart strict restrictions on internet usage to save youth from moral crazes.

Future Scope

We have already discussed about applications of web site classification on different areas. These areas have ever increasing importance and will remain as major application of web site classification.

The biggest limitation of our project is that it classify web sites according to it's textual content only. But modern web pages are designed to have several types of contents such as audio,picture,video,flash animation...etc. Another issue is it is not necessary that all of web sites may not fall into categories that we have chosen. For example, if we classify web sites according their content it is not possible to include google.com in a particular category. So it would be better if we classify web sites functionally also along with content classification. Algorithms that bring these things into consideration for the classification will have a great value and it will be a challenging task in front of computer science engineers.

Reference/Bibliography

1. Automated Classification of Web Sites using Naive Bayesian Algorithm-2012
   Ajay S. Patil, B.V. Pawar

2. Data mining techniques for web site classification-2011
   Gabriel Fiol-Roig, Margaret Miró-Julià, Eduardo Herraiz

3. Recent Researches in web page classification-a review-2012
   Alamelu Mangai J, Santhosh Kumar V,Sugumaran V

4. Fast webpage classification using URL features
   Min-Yen Kan,Hoang Oanh Nguyen Thi

5. Web Page Classification: Features and Algorithms-2007
   Xiaoguang Qi and Brian D. Davison