# Naive Bayes Approach for Website Classification

R. Rajalakshmi and C. Aravindan

Department of Computer Science and Engineering
SSN College of Engineering, Chennai, India
{rajalaxmi,aravindanc}@ssn.edu.in

**Abstract.** World Wide Web has become the largest repository of information because of its connectivity and scalability. With the increase in number of web users and the websites, the need for website classification gains attraction. The website classification based on URLs alone plays an important role, since the contents of web pages need not be fetched for classification. In this paper, a soft computing approach is proposed for classification of websites based on features extracted from URLs alone. The Open Directory Project dataset was considered and the proposed system classified the websites into various categories using Naive Bayes approach. The performance of the system was evaluated and Precision, Recall and F-measure values of 0.7, 0.88 and 0.76 were achieved by this approach.

**Keywords:** Website classification, URL, Bayes classifier.

## 1 Introduction

Web page classification is the process of assigning a web page to one or more predefined category labels. The classification of web pages is necessary and it is normally performed based on the content of the website. If the web sites are classified based on URLs alone, the web pages need not be fetched and analyzed. The URL based website classification can be used to identify the abnormal traffic generated from an organization by observing packets. So systems are designed to automate the classification process based on URLs.

In the proposed system, the classification of websites was done based on their URLs alone using Naïve Bayes approach. The websites were classified into one of the categories viz., Arts, Business, Computers, Games, Health, Home, News, Recreation and Reference. The performance of the system was evaluated using ODP dataset and Precision, Recall and F-measure of 0.7, 0.88 and 0.76 were achieved.

## 2 Related Works

Qi, X. and Davison [1] surveyed the existing automated classification systems and compared the approaches in terms of features and classifiers used. But, all these systems require the content be downloaded before classification. This method of classification after downloading the web page content may not be useful if the objective is to

V.V. Das, G. Thomas, and F. Lumban Gaol (Eds.): AIM 2011, CCIS 147, pp. 323–326, 2011.

block objectionable content or when speed is of crucial importance. Min Yen Khan [2] proposed an approach for categorizing web pages without web content using the ILP98 WebKB dataset. For this SVM based classification, the features viz., URL text, anchor text, title text and page text were considered and they reported an average F-measure of 0.43 for the following 4 categories: Course, Faculty, Project and Student. In the approach proposed by Min-Yen [3], the URLs were segmented into meaningful chunks and features such as URL component length, content, orthography, token sequence and precedence were considered. Then Maximum-entropy learning was applied to classify the web sites and an F-measure of 0.62 was achieved. Lim Wern Han et al.[4] proposed a framework for classifying the websites considering the features from URL, web page title and metadata information from web pages. They reported an average F-measure of 0.78 for 6 categories viz., Business, Economy, Entertainment, Government, Health, News and Sports. But they used additional features that are not part of the URL, and considered only 6 categories among which 3 are subcategories. In the proposed system, Naïve Bayes approach for classification of websites was explored using the features extracted only from the URLs.

## 3 URL Classification

The system was designed to classify the given URL into one of the 9 categories, viz., Arts, Business, Computers, Games, Health, Home, News, Recreation and Reference. A total of 8,55,939 URLs of 9 categories were extracted from the ODP dataset [5]. Among this 7,70,309 URLs were chosen randomly for forming the dictionary. The URLs from each category were parsed into tokens. The common words such as "www", "http", "html" etc. were removed from the list of tokens and only the unique tokens were used for constructing the dictionary. For 9 categories, a total of 9 dictionaries were used for training the Bayes classifier.

By applying Bayes theorem, the probability of given URL to belong to Category $C_i$ can be computed and the URLs can be classified using Naïve Bayes Classifier. To obtain the most probable hypothesis, given the training data, Maximum a posteriori hypothesis $h_{MAP}$ was used.

$$h_{MAP}=\arg\max P(D|h) * P(h) \quad (1)$$

For classification, an URL was separated into tokens and each token was checked with all the dictionaries in the training set. In each dictionary, the occurrence of these tokens was checked and the number of equal matches and the partial matches were counted. The value of each token was computed as shown in Equation (2). The value of the URL depends on its individual tokens, and the dictionary size of the corresponding category.

$$\text{valueToken}_i = \text{eqmatch} + (0.5 * \text{partmatch}) \quad (2)$$

The likelihood probability of an URL, given the Category was estimated by Eqn. (3), in which $\text{DictSize}C_i$ is the total number of words in the dictionary for the category

$C_i$, and i represents category number that varies from 1 to 9. The token count denoted by t, varies from 0 to n, where n represents the maximum number of tokens in a URL.

$$P(URL \mid C_i) = \frac{\sum_{t=0}^{n} valueToken_t}{DictSizeC_i} \tag{3}$$

To estimate the probability of the given URL to belong to Category $C_i$, prior probability and likelihood probability were used. The prior probability was assumed for every category according to ODP dataset. To obtain the prior probability of each category, total number of URLs in each category was divided by total number of URLs in the dataset. The probability of given URL to be classified under the Category $C_i$ was computed using equation (4).

$$P(C_i \mid URL) = P(URL|C_i) * P(C_i) \tag{4}$$

Here, $P(C_i \mid URL)$ is the posterior probability of given URL to be in the Category $C_i$, and $P(URL|C_i)$ is the likelihood probability that is calculated using Equation (3) and $P(C_i)$ is the prior probability of Category $C_i$. By Bayes theorem, final decision of classification was made. The category for which the URL has the maximum value was adjudged as the class of the URL. The experiments showed acceptable results for this approach of classification.

## 4 Experiments and Performance Evaluation

For experimental purposes, 9 categories of URLs from ODP dataset were used as shown in Table 1. A set of 9 experiments were performed by taking 80% for training and remaining for testing and cross validation was done. Finally 90% of training set was used and tested with remaining 10% URLs in each category. The system was implemented in Java. To evaluate the performance of the classifier, Precision, Recall and F-measure values were calculated and shown in Table 2. The average value of Precision, Recall and F-measure of 0.7, 0.88, and 0.76 were achieved.

**Table 1.** Number of URLs in Training and Testing Dataset

| Category | Number of URLs | Training Set | Testing Set |
|---|---|---|---|
| Arts | 227337 | 204597 | 22740 |
| Business | 227000 | 204300 | 22700 |
| Computers | 109201 | 98390 | 10921 |
| Games | 51207 | 46080 | 5127 |
| Health | 57515 | 51750 | 5755 |
| Home | 25368 | 22824 | 2544 |
| News | 8235 | 7407 | 828 |
| Recreation | 94565 | 85104 | 9461 |
| Reference | 55521 | 49967 | 5554 |
| **Total** | 8,55,939 | 7,70,309 | 85,630 |

**Table 2.** Performance measures

| | Arts | Business | Comp | Games | Health | Home | News | Recn | Refer |
|---|---|---|---|---|---|---|---|---|---|
| Precision | 0.80 | 0.97 | 0.65 | 0.69 | 0.76 | 0.80 | 0.53 | 0.66 | 0.44 |
| Recall | 0.82 | 0.57 | 0.92 | 0.90 | 0.97 | 0.97 | 0.93 | 0.93 | 0.88 |
| F-measure | 0.81 | 0.72 | 0.76 | 0.79 | 0.86 | 0.88 | 0.67 | 0.77 | 0.59 |

## 5 Conclusion

The proposed system classifies websites purely based on the URLs. With this approach of classification the Precision, Recall and F-measure values of 0.7, 0.88 and 0.76 were achieved. This system helps in monitoring the browsing behavior of users inside an organization and can assist the administrators to block unnecessary sites. This classifier can be further extended to identify the abnormal traffic generated from an organization and the classification accuracy can still be improved by using more features.

## References

1. Qi, X., Davison, B.D.: Web page classification: Features and algorithms. ACM Comput. Surv. 41(2), 1–31 (2009)
2. Kan, M.-Y.: Web Page Categorization without the Web Page. In: Proceedings of the 13th International World Wide Web Conference, pp. 262–263. ACM, New York (2004)
3. Kan, M.-Y., Thi, H.O.N.: Fast Webpage Classification using URL Features. In: Proceedings of CIKM 2005, Germany (2005)
4. Han, L.W., Alhashmi, S.M.: Joint Web-Feature (JFEAT): A Novel Web Page Classification framework. Communications of IBMA, Article ID 73408 (2010)
5. Open Directory Project, `http://www.dmoz.org`