Web Page Classification Using MDAW*k*NN

J. Alamelu Mangai
Birla Institute of Technology and Science Pilani, Dubai

V. Santhosh Kumar
Birla Institute of Technology and Science Pilani, Dubai

Karthik Ramesh
Birla Institute of Technology and Science Pilani, Dubai

INTRODUCTION

Web page classification, WPC, is quite simply the process of assigning labels (categories) to Web pages based on the kind of content they have. For e.g. News, sport, education, entertainment etc. It is slightly more challenging than text classification because of the dynamic content that Web pages have. This content ranges from text to flash, videos and picture. It can broadly be subdivided into two kinds – functional classification and subject classification. Functional classification is basically classifying the Website based on the role it plays while subject classification deals with the actual content of the Webpage.

The various applications of Web page classification include but are not restricted to constructing and expanding Web directories, improving the quality of search results, assisted Web browsing (suggesting similar content on pages such as YouTube), knowledge base construction and Web content filtering (blocking illegal content).

One of the techniques to classify Web pages is to use the content of the Web page with the traditional machine learning methods namely decision tree based methods like J48 (Mark Hall,2009) probabilistic methods like Naïve Bayes NB, instance based methods like K Nearest Neighbor (*k*NN), etc. Of these, the *k*NN algorithm is one of the most simple and easy to use methods. It is based on the principle of using distance measures to classify an unknown sample. One of the most commonly used distance measure is the Euclidean measure. Given a training set and a test data, the distance between the test sample and each of the training samples is calculated. Based on this distance, the test sample is assigned the class label of its *k* nearest neighbors using majority voting.

This is one of the most universally used algorithms with several advantages, the most important being the ease of use. It is also robust with regard to search space meaning classes need not be linearly separable. It can also be easily updated and it deals with very few parameters. However it has certain disadvantages too. It is very much computationally intensive as the Euclidean distance needs to be calculated for each training sample with the test sample. There is also no systematic approach to choosing the best value of '*k*'. Another problem deals with tie breaking, a scenario that occurs when there are an equal number of nearest neighbors that belong to different classes. *k*NN is also sensitive to noisy attributes.

In this chapter, the traditional *k*NN algorithm is improved for Web page classification. As thousands of features are used to induce a Web page classifier, the traditional *k*NN utilizes more system resources and needs more induction time, as it needs to compare the test data with every training example. Also it identifies the *k* nearest neighbors to the test data and applies simple majority voting to predict the class of the test data. In a data set with imbalanced class distribution, most of the *k* nearest neighbors may belong to

DOI: 10.4018/978-1-4666-5202-6.ch239

the majority class of the training set. Applying a simple majority voting on the class labels of such *k* nearest neighbors may wrongly predict the class of the test data and hence reduce the classification accuracy. In this chapter, the performance of *k*NN is improved through a pre-processing step, which identifies those Web pages, which are in a homogeneous neighborhood i.e, the Web page, and its neighborhood should belong to the same class. Such Web pages are made more significant in predicting the class of the test Web page using a weighting procedure, rather than simple majority voting as in traditional *k*NN.

BACKGROUND

Many approaches for automatic Web page classification have been witnessed over years in literature. With no preprocessed data there is no quality mining results. Since Web pages are of higher dimensions and have noisy information they need to be properly preprocessed which would otherwise increase the learning time and complexity of the classifiers. Feature selection is one way of solving the curse of dimensionality for content based Web page classifiers. Web page classification is improved by selecting the features through various methods as in (Indra Devi, Rajaraman, & Selvakuberan, 2008; Han, Lim, & Alhashmi, 2010; Selamat & Omata 2004 ; Chen, Ming, & Chang, 2009 ; Wakaki, Itakura, & Tamura, 2004 ; Jensen, & Shen, 2006; Peng, Ming, & Wang, 2008; Farhoodi, Yari, & Mahmoudi, 2009; Xu & Wang 2011).

By exploiting the characteristics of Chinese Web pages a new feature selection method by assigning weights to the HTML tags is proposed in (Chen, Du, Zhang & Han, 2010). The structures of the Web pages are used to classify them into information, research and personal home pages (Asirvatham & Ravi, 2001). Blocks in (Dai et al., 2006) are units that compose a Web page namely paragraphs, tables, lists and headings. The association between these blocks, Web pages and the queries are used to frame a query with content-based classification framework to classify a Web page. Visual features of a Web page like color and edge histograms, Gabor and texture features (De Boer, Someren, & Lupascu, 2010) summaries generated by human experts are used in (Shen, Chen, Yang, Zeng, Zhang, Lu, & Ma, 2004). These approaches of Web page classification cannot be applied in situations which suffer from hardware and software limitations. Further, they require lot of human expertise and are computationally complex. The various technologies that can be explored in Web information extraction have been explored in (Xhemali, Hinde, & Stone, 2007) and the authors have expressed their concern that many researchers start with the complex approaches directly rather than trying out the simpler ones first.

Machine learning methods (Tsukada & Washio, 2001; Zhang, 2001) have also been tweaked to improve performance of content-based classification in this domain. It is proved in (Xhemali et al., 2009) that Naïve Bayes, NB and C4.5 decision tree models are fast, consistent, easy to maintain and accurate in the training courses domain. NB classifier based on Independent Component Analysis (Zhongli & Zhijing, 2008), Hidden Naïve Bayes (Bo et al., 2009) with Symmetrical Uncertainty for word selection perform more satisfying in Web page categorization. It is identified in (Balamurugan, Pramala, Rajalakshmi, & Rajaram, 2010) that during the DT induction algorithms a tie appears when there are equal proportions of the target class in the leaf nodes, which leads to a situation where majority voting cannot be applied. The DT algorithm is improved to handle those exceptions.

Hence, due to the sheer volume of data on the Web manual categorization of Web pages is always incomplete. Meta tags cannot be used since there is a possibility for the Web page author to intentionally include keywords which do not reflect its content merely to increase its hit-rate. Link based and structure based approaches also