ABSTRACT

Web page classification has become a challenging task due to the exponential growth of the World Wide Web. There is a great demand for URL based web page classification systems as the features are extracted only from URLs thereby avoiding the need for downloading web pages and increasing the speed of classification. However achieving high accuracy may not be possible as URLs contain minimal information. Nevertheless URL based classifiers along with a rejection framework can be used as a first level filter in a multistage / hierarchical classifier by extracting cheaper features derived only from URLs. The costlier features such as title of the web page, anchor text, full text, and content of the sibling pages may be extracted at later stages based on the confidence score of URL classifier.

In this dissertation, machine learning techniques are proposed to design first level URL based classifiers in both multi-class and binary class scenarios. Noisy and irrelevant features present in URLs demand prudent feature selection methods for URL classification. For multi-class classification of URLs, a supervised feature selection method is proposed by which relevant URL features are identified using statistical techniques. A new feature weighting method is also proposed for Naive Bayes classifier by embedding the term goodness obtained from the feature selection method. For taking quick decisions in multi-
stage classifiers and in hierarchical classifiers, a confidence score based rejection framework is combined to the Naive Bayes classifier.

For binary one-against-all classification of URLs, a feature selection method for finding discriminating features is proposed. Linear SVM weight based approach is employed to select the relevant discriminating features. Using ODP dataset, dictionary is constructed for each class with the learnt discriminating features. Even though dictionaries of discriminating features are learnt using a particular dataset, it is shown through experimental studies that they are applicable across the datasets. For realizing binary URL classifiers as first level filters, SVM classifiers along with appropriate rejection framework are also proposed.

Extensive experiments have been carried out for multi-class classification of URLs and the proposed approach has been evaluated on two benchmark datasets viz., ODP dataset with 1.2 million URLs belonging to 13 different categories and WebKB dataset with 4 categories. With the proposed supervised feature selection method, even as an independent URL based classifier, an improvement of 52% in terms of $F_1$ measure has been achieved for multi-class URL classification on ODP dataset. For the small dataset WebKB also, better performance with an improvement of 16% over the existing method has been achieved. As a first level filter, by combining the proposed approach with rejection framework, the multi-class accuracy can be pushed up to 0.979 for URL based web page classification. With the proposed feature selection method for binary classification, category-specific dictionaries with discriminating features are learnt
using ODP dataset and applied on Google dataset. By this approach an $F_1$ measure of 0.881 was achieved and the performances of all the binary classifiers are better than the results reported in the literature. By adding a rejection framework to the SVM binary classifier, the accuracy of the URL based classifier can be pushed up to 0.99.

The scope of this dissertation is limited only to the first level filter design with appropriate machine learning models and rejection framework by extracting simple features from URLs alone. This work can be extended in the future by designing the successive levels of classifiers with more features.