CHAPTER 9

CONCLUSIONS AND FUTURE WORK

In this thesis, machine learning techniques have been applied for the problem of web page classification. URL based first level filter has been proposed to address the challenges in web page classification. The ways of deriving informative features from URLs have been studied and we have proposed simple feature extraction technique to derive URL features. Our method of using URL features does not need the contents to be downloaded for extracting the features and hence it is fast and saves bandwidth compared to other content based classifiers. We have conducted experimental study on two benchmark datasets viz., ODP dataset with 1.2 million URLs and small WebKB dataset. It is found that, the feature combination of tokens with n-grams (n=3 to 8) of concatenated URLs is better than all the other methods. The feature combination of tokens with n-grams of concatenated URLs demands feature selection methods to perform URL classification.

For multi-class classification of URLs, statistics based supervised feature selection methods have been proposed for selecting the relevant and contributing URL features. By constructing category-specific statistical term dictionaries with the relevant features, URLs are classified using Naive Bayes classifier. We have shown that statistically chosen relevant terms play significant role in classifying the URLs. We have also shown that embedding the term-goodness measure of CHIR based supervised feature selection method with Naive Bayes algorithm outperforms other state-of-art classifiers. For individual URL based classifiers, we have obtained an $F_1$ measure of 0.561 on ODP dataset for multi-class classification which is an improvement of 52% over the existing methods.
On WebKB dataset also, we have achieved an improvement of 16% over the existing approaches in multi-class classification. The feature selection process is done off-line, so it helps to take quick decisions while classifying the URLs.

For binary one-against-all classification of URLs, linear SVM weight based feature selection method has been proposed for finding relevant and discriminating features. By employing this method, discriminating features of type viz., Tokens, 3-grams, 4-grams, 5-grams, 6-grams, 7-grams and 8-grams have been selected for every category. By performing feature fusion, all the discriminating feature types were combined and universal dictionary was constructed for each class using ODP dataset. To show that the learnt dictionary is applicable on other datasets, experiments have been conducted on Google dataset. Using this ODP dictionary approach, the feature vector dimensionality is made independent of training set size. The macro-average precision of 91% has been achieved by the proposed ODP dictionary approach, which is an improvement of 4% over the existing method.

To realize the URL based classifier as a first level filter, suitable rejection frameworks have been proposed for both binary classifier and multi-class classifiers. Based on the confidence score of the classifier, either the URL is accepted or rejected. The posterior probability output of Naive Bayes classifier is used to determine the confidence score of the classifier in multiclass scenario, whereas the distance output of SVM classifier is converted into a calibrated probability measure and then used as a confidence score. The proposed approach combined with a rejection framework can be used as an initial stage classifier for any real world applications like information filtering and focussed crawling by tuning the rejection threshold. From the trade-off analysis on the training data of ODP dataset, we have shown that multi-class accuracy of 70% can be achieved by rejecting 30% of URLs. We have also verified this on the test set. As a first level classifier using URL based features, multi-class accuracy of 0.979 can be achieved by choosing the rejection threshold of 0.9. From the trade-off analysis
on the Google dataset, we have shown that, the binary classification accuracy of 90% can be achieved by rejecting 10% of URLs.

Even though, we have not developed the URL classifier for any particular application, the operational settings for two scenarios viz., information filtering and focussed crawling have been included.

The scope of this dissertation is limited only to the first level filter design combined with a rejection framework by deriving simple URL features alone without downloading the web page. This work can be extended in the future by designing the successive levels of classifiers by extracting more costlier features after downloading the corresponding web page.