ABSTRACT

The dream of semantic Internet populated with enormous amount of textual information remains a difficult challenge. The increasing criticality of un-structured textual data have amplified the need for an efficient text mining methodologies to extract patterns and knowledge from documents. This leads to the development of knowledge model for the machine to understand the text information for proper exploitation, classification and retrieval of textual data. Text Classification or Topic Classification (TC) is a task of assigning text document to one or more pre-defined topics based on their content. TC is constructed based on the training corpora which has the correct class label for each input, hence it is viewed as a Supervised Machine Learning task. The central challenge in TC is bridging the gap between information in text corpus and their association with the structured topics.

In this thesis, the potentiality of TC is applied for efficient Information Retrieval (IR). An Intelligent Topic Classification model is proposed by learning the semantics of text documents that supports Information Retrieval system. The proposed work semantically associates documents to meaningful topics by employing supervised learning strategy, where a classification model with pre-classified documents is first built and then the model is used to predict the topic of new documents. The traditional Bag of Words(BOW) representation for IR considering only the occurrence of words as features suffer with certain inherent deficiencies that ultimately affects the accuracy of the underlying classifier. The major deficiencies are the poor quality of the training data, lack of the discriminating power of the classifier and limited semantic richness of the features to represent documents. Moreover, performing TC in a sparse and high dimensional feature space remains a difficult task. It is also observed that the computational cost of building a classifier for huge number of relevant features is too expensive. Another important issue needed to be addressed is ambiguities in IR where multiple words that represent the same concept or single word mapping multiple concepts that affects the retrieval system to perform better in finding all of the relevant documents.

This thesis explores the semantic representation of text in order to overcome the drawbacks of the BOW method with the aim to improve Topic Classification effectiveness. The proposed work automatically assigns the documents to their meaningful topics by utilizing the hidden semantic structures in the document. In this thesis, three important approaches for Topic Classification are proposed and implemented. The first one is the semantic feature extraction and representation, the second one is the significant terms (feature) selection and the third one is supervised learning model construction. The research work also investigates to what extent the semantics helps to improve classification accuracy. The experimentations are performed through the following contributions

Initially, a new approach for Web content classification is proposed by identifying the concept of individual words in the document. The feature representation with Bag-of-Words (BOW) contains only the individual terms and their frequency as an index to train the classifier. In this work, concepts are extracted as Hypernyms that serves as a semantic features in addition to the original terms for training the classifier. Hence, the feature representation for index contains both words and Hypernyms. A Hypernym word tree is created for each document and significant features are extracted through weighted Term Frequency and Inverse Document Frequency(TF-IDF) where word weight is computed based on the Hypernyms number present in the radix search trie (re-trie-val). The performance of the proposed work is evaluated with standard WebKB dataset through k-Nearest Neighbor and Naive Bayes to show the classification accuracy, precision and recall.

An enhanced Topic Classification model is proposed with additional semantics. The enhancement of the model is achieved by proposing two important approaches. The first approach is the Semantic Smoothing technique which serves as a probabilistic model for significant feature selection, along with un-supervised weighted Clustering. The second approach is a Semantic Deep Learner(SDL) which serves as latent topic modelling for IR to train the features and to predict the category (topic) of the documents from different domain. This work enhances the document representation in BOW by taking compound words instead of individual words through the corpus based Semantic Smoothing technique. By using the semantic features, the training and prediction phases are processed to evaluate the influence of the enhanced semantics on Topic Classification effectiveness. Two standard datasets such as WebKB and 20NewsGroups are used for experimentation. The proposed SDL is compared with Neural Network Classifier in terms of accuracy.

An Intelligent Topic Classification model is proposed to deal with ambiguities in the polysemous word representation of training corpus. A Decision List based Word Sense Disambiguation(DLWSD) with Semantic Smoothing technique is introduced for selecting the rich features to train SDL. The semantically hidden concepts are extracted and the same was incorporated as additional features for training the classifier. The term space and semantic space are constructed to generate document-term matrix that serves as an input to Enriched Semantic Deep Learner (ESDL) for better prediction. The semantic similarity between the features are identified through

Point-wise Mutual Information. The performance of the model is evaluated by calculating the accuracy of the topic prediction using standard WebKB and 20NewsGroups datasets. Additionally, Electrical dataset is used for experimentation. The proposed ESDL is compared with SDL and Neural Network based classifier.

Finally, the thesis discuss about employing the proposed model in Patent domain. As patents are a rich knowledge source needed to be organized efficiently and conveniently, the proposed model is deployed to classify the patents to their appropriate topic based on their abstract and claim information. The proposed approach is new to the patent domain and provides considerable improvement in the classification accuracy when compared to other state of the art classifier using International Patent Classification (IPC) dataset.

This research work explores the use of semantic information for text classification across multiple domains and text types to support IR system. This research work portrays a semantic understanding based approach to minimizes the issues in the text mining. The proposed model for Topic Classification system understands the text to provide better performance than BOW representation.