# An Examination of Feature Selection Frameworks in Text Categorization

Bong Chih How, Wong Ting Kiong

Faculty of Computer Science and Information Technology
94300 Kota Samarahan,
Sarawak, Malaysia

{chbong, wongtingkiong}@gmail.com

**Abstract.** Feature selection, an important task in text categorization, is used for the purpose of dimensionality reduction. Feature selection basically can be performed locally and globally. For local selection, distinct feature sets are derived from different classes. The number of feature set is thus depended on the number of class. In contrary, only one universal feature set will be used in global feature selection. It is assumed that the feature set should preserve the characteristic of all classes. Furthermore, feature selection can also be carried out based on relevant feature set only (local dictionary) or both relevant and irrelevant feature set (universal dictionary). In this paper, we explored the different frameworks of feature selection to the task of text categorization on the Reuters(10) and Reuters(115) datasets (variants of Reuters-21578 corpus). We then investigate the efficiency of 7 different local or global feature selections corresponds the use of local and universal dictionary. Our experiments have shown that local feature selection with local dictionary yields optimal categorization results.

## 1 Introduction

Features selection is used for the purpose of dimensionality reduction by selecting significant terms from text. It can be performed basically in two ways: local and global feature selection. Furthermore, feature set can be seen in another perspective: local dictionary and universal dictionary. In this paper, we intend to answer the following questions with empirical evidence on the 7 feature selections:

- Which feature selection framework is optimal in text categorization?
- To what extend the performance of local feature selection compared to global one?
- Is combining positive and negative feature yields better classifier reading?

Feature selection in text categorization has enjoyed rich literatures in the past 2 decades, especially on local and global feature selection, plus small number of works on local and universal dictionary. However, there is no work reporting the correspondence effect of local and global feature selection on the different dictionary.