Semantic Characterisation: Knowledge Discovery for Training Set

Tan Ping Ping, Narayanan Kulathuramaiyer, and Azlina Ahmadi Julaihi

***Abstract*—This paper has proposed the use Latent Semantic Indexing (LSI) to extract semantic information to make the best use of the existing knowledge contained in training sets: Semantic Characterisation (SemC). SemC uses LSI to capture the implicit semantic structure in documents by directly applying category labels imposed by experts to make semantic structure explicit. The training set filtered by SemC is tested on a supervised automated text categorisation system using Support Vector Machine as classifier. Category by category analysis has shown the ability to bring out the semantic characteristics of the datasets. Even with a reduced training set, SemC is able to overcome the generalisation problem due to its ability to reduce noise within individual categories. Our empirical results also demonstrated that SemC managed to improve categorisation results of heavily overlapping categories. Empirical results also showed that SemC is applicable to a various supervised classifiers.**

***Index Terms*—Automated text categorisation, dataset, latent semantic indexing.**

I. Introduction

Even though automated text categorisation (ATC) attempts to mimic the categorisation model of human experts, current supervised ATC systems tend to merely exploit the information captured from a set of pre-determined documents class label assignments. During the process of human assignment of class labels, a great deal of implicit knowledge is employed. This knowledge is however not made explicit and systematically captured during the process of manually classifying documents. Reasons why a document is assigned to a particular class, if captured effectively, could provide valuable information for knowledge intensive tasks such as text categorisation.

When a document is determined to belong explicitly to exactly one category, (which is typical for most classified datasets) single-class category labels are assigned despite of document content overlaps with other categories. The existence of content overlaps across classes tends to complicate learning process [1]. Efforts that have been taken to overcome this problem [2] mainly concentrate on very specific domains or involves the addition of unlabelled training sets, thus, causing overfitting.

This paper explores the discovery of intrinsic patterns and characteristics of datasets in an effort to determine the characteristics of datasets that influence the performance of classifiers. Latent Semantic Indexing (LSI) [3] has been applied as a means of extraction of latent document dependency structures [4]. The relationships between extracted information about datasets are explored through an analysis of text categorisation results.

Manuscript received September 22, 2012; revised November 13, 2012. This work was supported in part by Universiti Malaysia Sarawak.

The authors are with the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Malaysia (e-mail: pptan@fit.unimas.my, nara@fit.unimas.my, ajazlina@fit.unimas.my).

II. Semantic Characterisation: Algorithm

Semantic Characterisation algorithm:

For training set, T_r with predefined categories, C

Where category labels, C_i; i = number of categories in T_r

Let T_c = set of positive examples for C_i

Let $\hat{X}$ =Reduced singular value decomposition of the term x document matrix for T_r,

For category C_i,

Using C_i as query (query terms selection according to the validity),

LSI retrieval is performed on (by ignoring the predefined labels)

Let T_l = set selected by LSI

Let $T_s = T_l\ T_c$ which represents the positive examples in T_r and LSI

Let $T' = T_s$ for all C

Train supervised classifier h using Naïve Bayes / SVM on T'

This work differs from other works that employ LSI, whereby LSI is not used as a feature extraction method or by manipulating LSI's vector spaces like existing supervised LSI methods [5] – [8], SemC, on the other hand, makes use of the existing categorical information and LSI's retrieval method; query technique and manipulation of the retrieval results of LSI to re-model the training sets. Thus, our approach explicates the meaning of training sets and queries applied to become directly interpretable by users while uncovering valuable category knowledge used by experts when performing document classification.

The knowledge contained in the training sets can then be manipulated through the selection of query terms. Hence, this eliminates the need to perform singular value decomposition locally for each separate category. SemC as such, does not require additional knowledge to be elicited from experts, as it is able to make use of latent document-term distribution patterns as contained the training set. The application of SemC results in a significantly reduced training set derived from the intrinsic text content characteristics of the training set.

SemC has been tested on a probabilistic classifier: multinomial Naïve Bayes (MNB) and has shown promising results [9]. This paper then reports the findings of experimentations in applying SemC in conjunction with the SVM classifier.

DOI: 10.7763/IJIMT.2013.V4.357