International Journal of Innovative Computing, Information and Control
Volume **5**, Number **12(A)**, December **2009**
 ISSN 1349-4198
pp. 4491–4499

THE APPLICATION OF LATENT SEMANTIC INDEXING AND ONTOLOGY IN TEXT CLASSIFICATION

Xi-Quan Yang[1], Na Sun[1,2], Tie-Li Sun[1], Xue-Ya Cao[1] and Xiao-Juan Zheng[3]

[1]School of Computer
[3]School of Software
Northeast Normal University
Changchun, Jilin, 130117, P. R. China
yangxq375@nenu.edu.cn

[2]Department of Basic Teaching
Jiangsu University of Science and Technology
Zhangjiagang, Jiangsu, 215600, P. R. China
sunna2004@126.com

Received July 2008; revised December 2008

Abstract. *The employment of ontology to improve text classification performance has been an active research topic in recent years. These researches include using domain ontology or using both WordNet general ontology and Latent Semantic Indexing (LSI) algorithm to realize text classification. However, in these methods, there are some problems such as high-dimensional and sparse space or inapplicability in professional fields of text classification. In order to solve these problems, this paper proposes a general framework for text classification, which is meant for the exploitation of full-fledged domain ontology as a knowledge base, to support the semantic-based text classification. Applying LSI algorithm to reduce the feature vector and using ontology knowledge as a background to classify the text can achieve higher performances in professional fields.*
Keywords: Latent semantic indexing, Domain ontology, Text classification

1. **Introduction.** The development of ontology has been one of the motivations of semantic web since it was envisioned. Thanks to its better semantic information, the use of ontology can make up the limitations of traditional text classification methods, so many scholars have focused on putting ontology ideas into the traditional methods, which has been made a big achievement in the last few years. Reference [13] proposed an approach to put ontology into text representation as the background knowledge, and realized automatic classification of XML texts. Reference [5] put forward an approach to classify web document by using domain ontologies, which were constructed from the concepts extracted from web information, and realized mapping between the concepts in ontology and the glossary in web documents.

These methods could get good classification results, but they did not take dimension reduction into account. Traditional document representation like bag-of-words encodes documents as feature vectors, which usually leads to sparse feature spaces with large dimensionality, thus making it hard to achieve higher classification accuracy and better time efficiency. The time and computational complexity of classifying increases rapidly with the augmentation of feature vector. Noise data and irrelevant features may even play a counteractive role in text classification.

Reducing dimension can enhance the computing efficiency and improve the precision of classification. So far, there are many reducing dimension methods, the most classical