A Supervised Clustering Method for Text Classification

Umarani Pappuswamy, Dumisizwe Bhembe, Pamela W. Jordan and Kurt VanLehn

Learning Research and Development Center,
3939 0'Hara Street, University of Pittsburgh,
Pittsburgh, PA 15260, USA
umarani@pitt.edu

Abstract. This paper describes a supervised three-tier clustering method for classifying students' essays of qualitative physics in the Why2-Atlas tutoring system. Our main purpose of categorizing text in our tutoring system is to map the students' essay statements into principles and misconceptions of physics. A simple `bag-of-words' representation using a naïve-bayes algorithm to categorize text was unsatisfactory for our purposes of analyses as it exhibited many misclassifications because of the relatedness of the concepts themselves and its inability to handle misconceptions. Hence, we investigate the performance of the k-nearest neighborhood algorithm coupled with clusters of physics concepts on classifying students' essays. We use a three-tier tagging schemata (cluster, sub-cluster and class) for each document and found that this kind of supervised hierarchical clustering leads to a better understanding of the student's essay.

1 Introduction

Text Categorization (or Classification)[1] can be seen either as an Information Retrieval task or a Machine Learning task of automatically assigning one or more well-defined categories or classes to a set of documents. Starting with the work of Maron [1] in the early 60s, Text Classification (TC) has found a significant place in a variety of applications including: automatic indexing, document filtering, word sense disambiguation, and information extraction. Our main focus is on the machine learning aspect of TC with the goal to devise a learning algorithm capable of generating a classifier which can categorize text documents into a number of predefined concepts. This issue has been considered in several learning approaches both with a supervised learning scheme [2, 3] and with an unsupervised and semi-supervised learning scheme [4, 5, 6].

In its simplest form, the text classification problem can be formulated as follows: We are given a set of documents $D = \{d_1, d_2, d_3 \ldots d_n\}$ to be classified and $C = \{c_1, c_2, c_3, \ldots c_n\}$ a predefined set of classes and the values {0, 1} interpreted as a decision to file a document d_j under c_i where 0 means that d_j is not relevant to the class defined and 1 means that d_j is relevant to the class defined. The main objective here is to devise a learning algorithm that will be able to accurately classify unseen documents

[1] We prefer the term `Text Classification' to `Text Categorization' and hence use the same in the rest of our paper.