CLASSIFICATION OF TEXT DOCUMENTS USING DOCUMENT CONTENTS

By

Daniel Wojcik

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: COMPUTER SCIENCE

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Approved:

Mukkai Krishnamoorthy, Thesis Adviser

Rensselaer Polytechnic Institute
Troy, New York

July 2009
(For Graduation August 2009)

Proper organization of documents is an important operation in many fields. Looking specifically at text documents, many of which would be produced by OCR data of questionable accuracy, we test the viability of several different semi-supervised statistical classification methods. To achieve this, we implemented an extendable program with many adjustable parameters to allow significant testing over wide ranges of values and algorithms. Looking at four specific schemes and their associated parameters, we examine the successes and failure of their usefulness on a database of one thousand books from the late seventeenth to nineteenth centuries. While the results found were not as accurate as we had hoped, we discuss the feasibility and related issues of this approach along with possible extensions and further applications of our implementation which may improve performance.