Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok
Abstrak


Kata kunci: Klasifikasi teks, Ontologi, Struktur, Dokumen berstruktur.
Abstract

Text classification (TC) is an important foundation of information retrieval and text mining. The main task of a TC is to predict the text’s class according to the type of tag given in advance. Most TC algorithms used terms in representing the document which does not consider the relations among the terms. These algorithms represent documents in a space where every word is assumed to be a dimension. As a result such representations generate high dimensionality which gives a negative effect on the classification performance. The objectives of this thesis are to formulate algorithms for classifying text by creating suitable feature vector and reducing the dimension of data which will enhance the classification accuracy. This research combines the ontology and text representation for classification by developing five algorithms. The first and second algorithms namely Concept Feature Vector (CFV) and Structure Feature Vector (SFV), create feature vector to represent the document. The third algorithm is the Ontology Based Text Classification (OBTC) and is designed to reduce the dimensionality of training sets. The fourth and fifth algorithms, Concept Feature Vector_Text Classification (CFV_TC) and Structure Feature Vector_Text Classification (SFV_TC) classify the document to its related set of classes. These proposed algorithms were tested on five different scientific paper datasets downloaded from different digital libraries and repositories. Experimental obtained from the proposed algorithm, CFV_TC and SFV_TC shown better average results in terms of precision, recall, f-measure and accuracy compared against SVM and RSS approaches. The work in this study contributes to exploring the related document in information retrieval and text mining research by using ontology in TC.

Keywords: Text classification, ontology, structural, structured documents.
Acknowledgement

It gives me great pleasure to express my gratefulness to everyone who contributed in completing this thesis. It was my pleasure to study under Associate Professor Dr. Azman Yasin’s supervision. I'm so grateful for his support during the last five years. I am so grateful for his all assistants that he gave me through these years. There are no words to express my gratitude for his guidance in helping me to achieve my goal. Without his valuable support, my thesis would not have been possible. I would like to tell him that thank you so much for everything you have been done for me to reach my goal. I would like to thank my co-supervisor Dr. Nor Idayu Mahat for her progressive thinking and her open mind. Her continuous advice and significant comments helped develop my work successfully.

To my father, whose surname I proudly carry – I am forever appreciative. I want to tell him thanks for all things you supported me and make me strong to across this stage of my life. To my mother, who gave me life and prayed for me all the time, may Allah continuously bless her with good health. To my sisters Sahar and Rafah, I would like to tell them thanks for your feelings and supporting. To my dear brothers Ali and Hassanin, thanks for their love and support. To my Husband Ghassan, who gave me power and patience during the last five years of study, I thank his from the bottom of my heart. I would also like to thank my two young babies Mohammed and Zainab, without whom my goal would not have been achieved. I dedicate this work to my family. I'm so glad to study at Universiti Utara Malaysia (UUM). During my time in UUM, I have gained a lot of friends, and studying there was like being in my hometown. My sincere gratitude to all of them for all the encouragement during my study. I want to tell all of them thank you so much for everything you help me.
# Table of Contents

Permission to Use........................................................................................................... ii
Abstrak ............................................................................................................................ iii
Abstract ............................................................................................................................ iii
Acknowledgement ........................................................................................................... v
Table of Contents ............................................................................................................ vi
List of Tables ...................................................................................................................... ix
List of Figures ..................................................................................................................... x
List of Appendices ........................................................................................................... xii

## CHAPTER ONE INTRODUCTION ............................................................................. 1
1.1 Background ................................................................................................................... 1
1.2 Problem Statement ....................................................................................................... 5
1.3 Research Objectives ................................................................................................. 8
1.4 Significant of the Study ............................................................................................. 8
1.5 Scope and Limitation ............................................................................................... 9
1.7 Thesis Organization ............................................................................................... 11

## CHAPTER TWO LITERATURE REVIEW ................................................................. 13
2.1 Introduction ................................................................................................................ 13
2.2 Text Classification .................................................................................................... 14
   2.2.1 Text Classification Algorithm ........................................................................... 14
       2.2.1.1 Support Vector Machines .......................................................................... 14
       2.2.1.2 Nearest Neighbor .................................................................................. 16
       2.2.1.3 Decision Trees ...................................................................................... 17
       2.2.1.4 Naïve Bayes Algorithm .......................................................................... 19
       2.2.1.5 Neural Network .................................................................................... 20
       2.2.1.6 Rocchio’ Algorithm .............................................................................. 23
   2.2.2 Approaches to Create Feature Vector for Text Classification ....................... 23
       2.2.2.1 Part of Speech .................................................................................... 24
       2.2.2.2 N-gram .............................................................................................. 26
### Chapter Two

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.2.3</td>
<td>Term Frequency Inverse Document Frequency</td>
<td>28</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Feature Selection Method to Reduce Dimension</td>
<td>46</td>
</tr>
<tr>
<td>2.2.3.1</td>
<td>Information Gain (IG)</td>
<td>47</td>
</tr>
<tr>
<td>2.2.3.2</td>
<td>Chi2-Test (CHI)</td>
<td>52</td>
</tr>
<tr>
<td>2.2.3.3</td>
<td>Document Frequency Thresholding (DF)</td>
<td>56</td>
</tr>
<tr>
<td>2.2.3.4</td>
<td>Mutual Information (MI)</td>
<td>60</td>
</tr>
<tr>
<td>2.2.3.5</td>
<td>Ontology to Reduce the Dimension</td>
<td>64</td>
</tr>
</tbody>
</table>

### Chapter Three

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Research Framework</td>
<td>113</td>
</tr>
<tr>
<td>3.2</td>
<td>Dataset Development</td>
<td>115</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Dataset Creation</td>
<td>116</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Removing Stop Words</td>
<td>123</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Stemming</td>
<td>124</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Types of Terms Extracted</td>
<td>125</td>
</tr>
<tr>
<td>3.2.5</td>
<td>Ontology Construction</td>
<td>126</td>
</tr>
<tr>
<td>3.3</td>
<td>Create Set of Feature</td>
<td>126</td>
</tr>
<tr>
<td>3.4</td>
<td>Create Set of Concept from Created Ontology</td>
<td>127</td>
</tr>
<tr>
<td>3.5</td>
<td>Classify Document</td>
<td>127</td>
</tr>
<tr>
<td>3.6</td>
<td>Validation</td>
<td>128</td>
</tr>
<tr>
<td>3.7</td>
<td>Evaluation Measures</td>
<td>129</td>
</tr>
<tr>
<td>3.8</td>
<td>Summary</td>
<td>130</td>
</tr>
</tbody>
</table>
# CHAPTER FOUR ENHANCED ONTOLOGY BASED TEXT CLASSIFICATION ALGORITHM FOR SCIENTIFIC PAPER

4.1 Introduction ........................................................................................................ 132
4.2 Ontology Structure ............................................................................................ 135
4.3 Feature Vector Creation Algorithm for Text Classification ......................... 136
   4.3.1 Proposed Concept Feature Vector (CFV) ................................................. 137
   4.3.2 Proposed Structure Feature Vector (SFV) ............................................ 144
4.4 Ontology Based Text Classification Algorithm (OBTC) ............................. 150
4.5 Combine Feature Vector Creation Algorithm with Text Classification
   Algorithm ........................................................................................................... 156
   4.5.1 Proposed Concept Feature Vector for Text Classification CFV_TC
      Algorithm ........................................................................................................ 156
   4.5.2 Proposed Structure Feature Vector _Text Classification (SFV_TC)
      Algorithm ....................................................................................................... 161
4.6 Summary ........................................................................................................... 167

# CHAPTER FIVE RESULTS AND ANALYSIS

5.1 Result and Analysis ......................................................................................... 162
5.2 Summary .......................................................................................................... 198

# CHAPTER SIX CONCLUSION AND FUTURE WORK

6.1 Contributions ................................................................................................... 199
6.2 Future Works .................................................................................................. 200

REFERENCES ...................................................................................................... 202
List of Tables

Table 2.1 Literature summary on feature creation in text classification .................. 37
Table 2.2 Literature summary on reducing dimension .......................................... 71
Table 2.3 Literature summary on reducing dimension ontology as classifier for text classification task .............................................................. 105
Table 3.1 Query for creating dataset and its classes ............................................ 119
Table 3.2 Datasets for the proposed work ............................................................ 121
Table 5.1 Evaluation of CFV_TC ........................................................................ 171
Table 5.2 Evaluation of SFV_TC ........................................................................ 173
Table 5.3 Evaluation of RSS ............................................................................... 175
Table 5.4 The comparison between RSS, CFV_TC, and SFV_TC algorithm in terms of precision ................................................................. 178
Table 5.5 The comparison between RSS, CFV_TC, and SFV_TC algorithm in terms of recall ...................................................................................... 180
Table 5.6 The comparison between RSS, CFV_TC, and SFV_TC algorithm in terms of F_measure ........................................................................ 182
Table 5.7 The comparison between RSS, CFV_TC, and SFV_TC algorithm in terms of Accuracy. Evaluation of SFV_TC ............................................. 184
Table 5.8 Evaluation of SFV_TC ........................................................................ 186
Table 5.9 The comparison between RSS, CFV_TC, and SFV_TC algorithm in terms of feature size ................................................................. 188
Table 5.10 The comparison between SVM, CFV_TC, and SFV_TC algorithm in terms of precision ........................................................................ 190
Table 5.11 The comparison between SVM, CFV_TC, and SFV_TC algorithm in terms of recall ...................................................................................... 192
Table 5.12 The comparison between SVM, CFV_TC, and SFV_TC algorithm in terms of F_Measure ........................................................................ 194
Table 5.13 The comparison between SVM, CFV_TC, and SFV_TC algorithm in terms of Accuracy ........................................................................ 197
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Proposed Research Architecture</td>
<td>114</td>
</tr>
<tr>
<td>3.2</td>
<td>Scientific paper structures</td>
<td>117</td>
</tr>
<tr>
<td>3.3</td>
<td>Confusion Matrix for Text Classification Evaluation Classification of RSS feed news items using ontology</td>
<td>130</td>
</tr>
<tr>
<td>4.1</td>
<td>General Architecture of the proposed work</td>
<td>133</td>
</tr>
<tr>
<td>4.2</td>
<td>The proposed text classification framework</td>
<td>134</td>
</tr>
<tr>
<td>4.3</td>
<td>Ontology for computer science domain (Classification concept)</td>
<td>135</td>
</tr>
<tr>
<td>4.4</td>
<td>Ontology for classification concept from Computer Science ontology Science Domain (RDF)</td>
<td>136</td>
</tr>
<tr>
<td>4.5</td>
<td>The proposed Concept Feature Vector (CFV) Algorithm</td>
<td>139</td>
</tr>
<tr>
<td>4.6</td>
<td>The proposed Structure Feature Vector SFV algorithm</td>
<td>146</td>
</tr>
<tr>
<td>4.7</td>
<td>The proposed text classification Ontology Based Text Classification Algorithm</td>
<td>153</td>
</tr>
<tr>
<td>4.8</td>
<td>The proposed CFV_TC algorithm</td>
<td>159</td>
</tr>
<tr>
<td>4.9</td>
<td>The proposed SFV_TC algorithm</td>
<td>164</td>
</tr>
<tr>
<td>5.1</td>
<td>The evaluation of the first proposed algorithm CFV_TC</td>
<td>172</td>
</tr>
<tr>
<td>5.2</td>
<td>The evaluation of the second proposed algorithm SFV_TC</td>
<td>174</td>
</tr>
<tr>
<td>5.3</td>
<td>The evaluation of the RSS classification algorithm</td>
<td>176</td>
</tr>
<tr>
<td>5.4</td>
<td>The comparison between RSS, CFV_TC, and SF_TC algorithm in terms of precision</td>
<td>179</td>
</tr>
<tr>
<td>5.5</td>
<td>The comparison between RSS, CFV_TC, and SFV_TC algorithm in terms of recall</td>
<td>181</td>
</tr>
<tr>
<td>5.6</td>
<td>The comparison between RSS, CFV_TC, and SFV_TC algorithm in terms of F_measure</td>
<td>183</td>
</tr>
<tr>
<td>5.7</td>
<td>The comparison between RSS, CFV_TC, and SFV_TC algorithm in terms of Accuracy</td>
<td>185</td>
</tr>
<tr>
<td>5.8</td>
<td>Results of the SVM Classification</td>
<td>187</td>
</tr>
<tr>
<td>5.9</td>
<td>The comparison between SVM, CFV_TC, and SFV_TC algorithm in terms of precision</td>
<td>189</td>
</tr>
</tbody>
</table>
Figure 5.10 The comparison between SVM, CFV_TC, and SFV_TC algorithm in terms of recall .................................................................191
Figure 5.11 The comparison between SVM, CFV_TC, and SFV_TC algorithm in terms of F_Measured ............................................................193
Figure 5.12 The comparison between SVM, CFV_TC, and SFV_TC algorithm in terms of Accuracy .............................................................194
List of Appendices

Appendix A Samples of Data ................................................................. 228
Appendix B Output of Concept Feature Vector creation _text classification
(CFV_TC) .................................................................................................. 234
Appendix C Output of Structure Feature Vector creation _text classification
(SFV_TC) .................................................................................................. 243
CHAPTER ONE
INTRODUCTION

1.1 Background

Text categorization is the task of assigning predefined categories to free-text documents. It can provide conceptual views of document collections and has important applications in the real world (Kaur & Jyoti, 2013). In the recent years, TC has gained tremendous attention and rapidly developed. Today, TC is widely used in applications such as “automatic indexing” for "Boolean information retrieval" systems, "document organization", "text filtering", and "word-sense disambiguation" (Rafi, et al, 2012; Shimodaira, 2014).

According to (Calvo, Lee, & Li, 2006), TC reduces the time required to classify vast amounts of documents without the need for experts. While TC methods may vary in terms of accuracy and computation efficiency, TC methods generally save time and expense required to perform TC. Classification algorithms can be used to extract models describing important data classes.

There are several algorithms used to classify text such as "k-nearest neighbors" (KNN), "naïve Bayes" (NB), and "Support Vector Machines" (SVM) (Patra & Singh, 2013). To build a classifier in text classification there is need to define set of example as training set. These sets are labelled with pre-defined classes (Li & Liu, 2003). Often, a data set sample contains both positive and negative examples of a concept to induce a classification rule use machine learning algorithm (Aytug, Boylu, & Koehler, 2006).
The contents of the thesis is for internal user only
References


Calvier, F. c.-E., Plantié, M., Dray, G. e., & Ranwez, S. (2013). Ontology Based Machine Learning for Semantic Multiclass Classification. Author manuscript,


Han, J., Kamber, M., & Pei, J. (2013). Data Mining: Concepts and Techniques. Book, Chapter 9, Classification: Advanced Methods, University of Illinois at Urbana-Champaign & Simon Fraser University.


Categorization Technique. *Paper presented in IEEE International Conference on Web Intelligence and Intelligent Agent Technology WI-IAT '08*, 143 - 146.


