IMPROVED SEMANTIC GRAPH-BASED PLAGIARISM DETECTION

AHMED HAMZA OSMAN AHMED

A thesis submitted in fulfilment of the requirements for the award of the degree of Doctor of Philosophy (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

MARCH 2013
To my beloved parents, brothers, sisters, wife, son (Albraa), and daughter (Braa’h)
ACKNOWLEDGEMENT

First of all, I would like “Praise be to Allah, the cherisher and the sustainers of the world”, “praise be to him he who taught by the pen, taught man, that which he did not know”

First of all, I would like to thank my supervisor professor Dr. Naomie Binti Salim for allowing me to carry out this thesis, and the wonderful experience and all the good times she gave to me in addition to, continuous motivation and encouragement. I also owe my great deal of thank to UTM and Faculty of Computer Science and Information System for providing the good environment to carry out the research, and to our professors and staff for their friendly support and help throughout the study and special thank to Prof. Dr. Naomie for supporting me in this study.

And most importantly, I would like to thank to my family for their love and support, especially my mother, my father and my wife for their continuous support and supplication.

This thesis would not see the light and would not have been possible without the continuous financial support of my university “International University of Africa” and for providing me the opportunity to study in Malaysia.

Finally, I would like to thank all my friends and colleagues for their support and help.
ABSTRACT

Plagiarism detection occurs when the content of a text is copied without permission or citation. Nowadays, many text documents on the internet are easily copied and accessed. This study proposed improved methods to handle plagiarism. The proposed plagiarism detection methods are developed using graph-based representation and semantic role labeling which are improved using fuzzy logic technique and chi-squared automatic interaction detection. The graph-based method does not only represent the content of a text document as a graph, but also captures the underlying semantic meaning in terms of the relationships among its concepts. Semantic role labeling is superior in generating semantic arguments for each sentence. This semantic role labeling plays an important part in plagiarism detection as it segments the role of concepts in documents to labels which are compared and used to detect plagiarism. Scoring for each argument generated by the fuzzy logic method to select important arguments is also another feature of this study. Chi-squared Automatic Interaction Detection technique was applied to enforce the results obtained from the fuzzy logic and semantic role labeling by selecting important arguments from the sentences. It is concluded that not all arguments in the text are useful in the plagiarism detection process. Therefore, only the most important arguments were selected by the fuzzy logic and Chi-squared automatic interaction detection, and the results were used in the similarity calculation process. Experiments were tested on the PAN-PC-2009 for standard artificial simulation corpus and the Short Answers Questions (CS11) for human simulation corpus in plagiarism detection. The proposed methods detected many types of plagiarisms, such as copy paste plagiarism, rewording or synonym replacement, changing of word structure in the sentences, modifying the sentence from passive voice to active voice and vice-versa. Results from the experiments using the proposed methods in comparison to other plagiarism detection techniques (Fuzzy Semantic-Based String Similarity and Longest Common Subsequence) achieved better performance in terms of recall (93%), precision (90%) and f-measure (91%).
ABSTRAK

# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DECLARATION</td>
<td></td>
<td>ii</td>
</tr>
<tr>
<td>DEDICATION</td>
<td></td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td></td>
<td>iv</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td></td>
<td>v</td>
</tr>
<tr>
<td>ABSTRAK</td>
<td></td>
<td>vi</td>
</tr>
<tr>
<td>CONTENTS</td>
<td></td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td></td>
<td>xiv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td></td>
<td>xvi</td>
</tr>
<tr>
<td>LIST OF ABBREVIATION</td>
<td></td>
<td>xix</td>
</tr>
<tr>
<td>LIST OF APPENDICES</td>
<td></td>
<td>xxi</td>
</tr>
</tbody>
</table>

1  INTRODUCTION 1

1.1 Introduction 1
1.2 Background of Research 3
1.3 Problem Statement 11
1.4 Objectives 11
1.5 Research Scope 12
1.6 Significance of Research 13
1.7 Expected Contributions 13
1.8 Thesis Organization 14

2  LITERATURE REVIEW 16

2.1 Introduction 16
2.2 Text Plagiarism Detection 17
2.3 Plagiarism Detection Process 19
2.3.1 Plagiarism detection process stages 19
2.4 Plagiarism Detection Methodology  
2.4.1 Character-Based Methods  
2.4.2 Structure-Based Methods  
2.4.3 Classification and Cluster-Based Methods  
2.4.4 Syntax-Based Methods  
2.4.5 Cross language-Based Methods  
2.4.6 Semantic-Based Methods  
2.4.7 Citation-Based Methods  
2.5 Related work  
2.5.1 Graph-based Representation  
2.5.2 Graph Matching (Similarity) and Subgraph Detection  
2.5.3 Semantic Role Labelling (SRL)  
2.5.3.1 SENNA Toolkit  
2.5.3.2 The Proposition Bank (PropBank)  
2.6 Introduction  
2.6.1 Fuzzy Logic Systems  
2.6.2 Chi-squared Automatic Interaction Detection  
2.7 Evaluation Measure and PAN-PC Data Set  
2.7.1 Plagiarism Detection Corpus  
2.7.2 Plagiarism Detection Performance  
2.8 Discussion  
2.9 Summary  

3 METHODOLOGY  
3.1 Introduction  
3.2 Research design  
3.3 Operational framework  
3.3.1 Phase 1: Planning and literature review  
3.3.2 Phase 2: Plagiarism detection using graph-based representation  
3.3.3 Phase 3: An improved plagiarism detection scheme based on semantic role labelling  
3.3.4 Phase 4: An improved plagiarism detection
technique based on fuzzy semantic role labelling

3.3.4.4 Fuzzy semantic role labelling 106
3.3.5 Phase 5: An Improved Plagiarism Detection Method Based on Semantic Role Labelling and Chi-squared Automatic Interaction Detection 108
3.3.6 Phase 6: Result Evaluation and Validation 111
3.4 Discussion 111
3.5 Summary 114

4 PLAGIARISM DETECTION USING GRAPH-BASED REPRESENTATION 115

4.1 Introduction 115
4.2 Proposed Graph-based Representation for Plagiarism Detection 116
4.2.1 Data Pre-processing 116
4.2.2 Text Graph-based Representation 116
4.2.3 Concept Extraction 122
4.2.4 Similarity Detection and Topic Signature 122
4.2.5 Experimental Design and Dataset 124
4.2.6 Results and Evaluation 125
4.2.7 Discussion 130
4.3 Summary 132

5 AN IMPROVED PLAGIARISM DETECTION SCHEME BASED ON SEMANTIC ROLE LABELLING 133

5.1 Introduction 133
5.2 Semantic Role Labelling (SRL) 134
5.2.1 SENNA Toolkit 136
5.3 Plagiarism Detection Using SRL 136
5.4 Experimental Design and Dataset 143
5.4.1 Similarity Detection 144
5.5 Results and Discussion 147
AN IMPROVED PLAGIARISM DETECTION
SCHEME BASED ON FUZZY SEMANTIC ROLE LABELLING

6

6.1 Introduction
6.2 Fuzzy Logic and Fuzzy Inference Systems
6.2.1 Fuzzification, Inference System and Membership Functions
6.2.2 Construct the Fuzzy IF-THEN Rules
6.2.3 Defuzzification
6.3 Experimental Design and Dataset
6.3.1 Similarity Detection
6.4 Results and Discussion
6.5 Discussion
6.6 Summary

7

AN IMPROVED PLAGIARISM DETECTION
METHOD BASED ON SEMANTIC ROLE LABELLING AND CHI-SQUARED AUTOMATIC INTERACTION DETECTION

7.1 Introduction
7.2 The CHAID Algorithm
7.2.1 Pre-processing
7.2.2 SRL Extraction
7.2.3 CHAID Algorithm Identification
7.3 Important Arguments Based on CHAID Algorithm
7.4 Binning and merging mechanism of the CHAID algorithm
7.4.1 Binning of Predictors
7.4.2 Merging Categories for Predictors (CHAID)
7.5 Experimental Design and Dataset 199
7.5.1 Similarity Detection 199
7.6 Results and Discussion 199
7.7 Discussion and Summary 203

8 CONCLUSIONS AND FUTURE WORK 204

8.1 Introduction 204
8.2 The Review of the Current Studies 205
8.3 Findings and Contributions of the Study 211
8.3.1 Findings 211
8.3.2 Contribution 212
8.4 Future Work 215
8.5 Summary 215

REFERENCES 216
Appendix A - C 229 -253
## LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>The summary of the character and string-based methods</td>
<td>31</td>
</tr>
<tr>
<td>2.2</td>
<td>The summary of the structure-based methods</td>
<td>33</td>
</tr>
<tr>
<td>2.3</td>
<td>The summary of the classification and cluster-based methods</td>
<td>35</td>
</tr>
<tr>
<td>2.4</td>
<td>The summary of the syntax-based methods</td>
<td>36</td>
</tr>
<tr>
<td>2.5</td>
<td>The summary of the cross language-based methods</td>
<td>38</td>
</tr>
<tr>
<td>2.6</td>
<td>The summary of the semantic-based methods</td>
<td>40</td>
</tr>
<tr>
<td>2.7</td>
<td>The summary of the citation-based methods</td>
<td>42</td>
</tr>
<tr>
<td>2.8</td>
<td>The summary of the structure-based methods</td>
<td>49</td>
</tr>
<tr>
<td>2.9</td>
<td>Argument types and their descriptions</td>
<td>59</td>
</tr>
<tr>
<td>2.10</td>
<td>Document statistics in the PAN-PC-2011</td>
<td>73</td>
</tr>
<tr>
<td>2.11</td>
<td>Statistics and distribution of plagiarism cases in PAN-PC-2011</td>
<td>74</td>
</tr>
<tr>
<td>4. 1</td>
<td>A sample of result cross the PAN-PC-09 data set</td>
<td>125</td>
</tr>
<tr>
<td>4. 2</td>
<td>Results and comparison using Recall, Precision, and F-measure across the set of documents from PAN-PC-09</td>
<td>126</td>
</tr>
<tr>
<td>4. 3</td>
<td>Heavy plagiarism class</td>
<td>128</td>
</tr>
<tr>
<td>4. 4</td>
<td>Light plagiarism class</td>
<td>129</td>
</tr>
<tr>
<td>4. 5</td>
<td>Cut-and-paste plagiarism class</td>
<td>129</td>
</tr>
<tr>
<td>5. 1</td>
<td>Argument types</td>
<td>148</td>
</tr>
<tr>
<td>5. 2</td>
<td>Results across the set of documents</td>
<td>149</td>
</tr>
<tr>
<td>5. 3</td>
<td>Results after similarity calculation</td>
<td>150</td>
</tr>
<tr>
<td>5. 4</td>
<td>Results after similarity calculation using CS11</td>
<td>150</td>
</tr>
<tr>
<td>5. 5</td>
<td>Behaviours of the arguments after the optimisation process</td>
<td>152</td>
</tr>
<tr>
<td>5. 6</td>
<td>Evaluation results after selection of important arguments</td>
<td>154</td>
</tr>
<tr>
<td>5. 7</td>
<td>Statistical significant testing using t-test</td>
<td>155</td>
</tr>
<tr>
<td>5. 8</td>
<td>The comparison of average Recall, Precision and F-measure cross 100 documents from PAN-PC-09 corpus with other</td>
<td></td>
</tr>
</tbody>
</table>
methods

5. 9 Comparison based on the type of plagiarism
5. 10 Heavy plagiarism class
5. 11 Light plagiarism class
5. 12 Cut-and-paste plagiarism class
5. 13 Comparison between the proposed method and other techniques by time complexity

6. 1 Sample of arguments scores cross 1,000 documents
6. 2 Fuzzy linguistic variables
6. 3 Behaviours of the arguments after optimisation using FIS
6. 4 Evaluation results after we selected the important arguments.
6. 5 Statistical Significance testing using t-test
6. 6 The comparison of average Recall, Precision and F-measure scores of difference plagiarism detection methods of 100 documents

6. 7 Comparison between the proposed method and other techniques by time complexity

7. 1 Data with frequency field
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. 1</td>
<td>Four-stage Plagiarism Detection Process</td>
<td>20</td>
</tr>
<tr>
<td>2. 2</td>
<td>Plagiarism Types with Some Related Detection Principals</td>
<td>23</td>
</tr>
<tr>
<td>2. 3</td>
<td>Taxonomy of Plagiarism Methods</td>
<td>23</td>
</tr>
<tr>
<td>2. 4</td>
<td>Supergraph</td>
<td>44</td>
</tr>
<tr>
<td>2. 5</td>
<td>Subgraph</td>
<td>44</td>
</tr>
<tr>
<td>2. 6</td>
<td>Common Sub Graph of G1 and G2</td>
<td>45</td>
</tr>
<tr>
<td>2. 7</td>
<td>Maximum Common Subgraph (MCS)</td>
<td>45</td>
</tr>
<tr>
<td>2. 8</td>
<td>Minimum Common Supergraph (MCS)</td>
<td>46</td>
</tr>
<tr>
<td>2. 9</td>
<td>MMCSN Distance Measure between Two Graphs</td>
<td>47</td>
</tr>
<tr>
<td>2. 10</td>
<td>An Example Pattern Graph P and Data Graph (Ullmann, 1976)</td>
<td>51</td>
</tr>
<tr>
<td>2. 11</td>
<td>A Partial Search-Tree for Ullmann's Algorithm, Mapping Vertices from Pattern Graph P to Data Graph G.</td>
<td>52</td>
</tr>
<tr>
<td>2. 12</td>
<td>General process of SRL</td>
<td>58</td>
</tr>
<tr>
<td>2. 13</td>
<td>Triangular membership function</td>
<td>63</td>
</tr>
<tr>
<td>2. 14</td>
<td>Gaussian Membership Function</td>
<td>64</td>
</tr>
<tr>
<td>2. 15</td>
<td>Bell Membership Function</td>
<td>64</td>
</tr>
<tr>
<td>2. 16</td>
<td>Trapezoidal membership function</td>
<td>65</td>
</tr>
<tr>
<td>2. 17</td>
<td>Fuzzy inference systems for tipping problem(MathWorks, 2012)</td>
<td>69</td>
</tr>
<tr>
<td>2. 18</td>
<td>Example of a decision tree with categorical predictors (Merel van Diepen and Philip Hans Franses, 2006)</td>
<td>71</td>
</tr>
<tr>
<td>3. 1</td>
<td>Operational Framework</td>
<td>88</td>
</tr>
<tr>
<td>3. 2</td>
<td>Phase2 Plagiarism Detection Using Graph-based Representation</td>
<td>91</td>
</tr>
</tbody>
</table>
3.3 Graph-based representation
3.4 Synsets Related to the Term “Car”.
3.5 Steps of Semantic Roles Labelling
3.6 Similar Arguments Comparison
3.7 Phase 4 Plagiarism Detection Technique Based on Fuzzy Logic and SRL Method
3.8 Plagiarism Detection Method Based on CHAID Algorithm and SRL Method.

4.1 Sentence Representations for Example 1
4.2 Comparisons between Original Document and Suspected Document Using Concept Level Similarity
4.3 Graph-Based Technique Comparisons with LCS and Semantic Techniques
4.4 Comparison Results with Heavy Plagiarism Class Cross CS11 Corpus
4.5 Comparison Results with Light Plagiarism Class Cross CS11 Corpus
4.6 Comparison Results with Cut-and Paste Plagiarism Class Cross CS11 Corpus

5.1 General Process of SRL
5.2 Structure Phase of the Proposed Method
5.3 Comparisons of Similar Arguments
5.4 Analysis for the Original Sentence Using SRL
5.5 Analysis for the Suspected Sentence Using SRL
5.6 Similarity Calculations between the Suspected and Original Document
5.7 Behaviours of the Arguments after the Optimisation Process
5.8 Comparison Results with Plagiarism Detection Techniques
5.9 Comparison Results with Heavy Plagiarism Class Cross CS11 Corpus
5.10 Comparison Results with Light Plagiarism Class Cross
CS11 Corpus

5. 11 Comparison Results with Cut-and Paste Plagiarism Class Cross CS11 Corpus

6. 1 Gaussian Membership Function for our Fuzzy Model
6. 2 Sample of IF_THEN Rules with “AND” Operator
6. 3 Comparison Results with Other Plagiarism Detection Techniques

7. 1 Sample Result of our Proposed Method using CHAID Algorithm Tree.
7.2 Important Arguments Selected By the CHAID Algorithm
### LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>Function Labelling The Nodes</td>
</tr>
<tr>
<td>$B$</td>
<td>Function Labelling The Edges</td>
</tr>
<tr>
<td>A_F</td>
<td>Arguments features</td>
</tr>
<tr>
<td>ADV</td>
<td>General purpose</td>
</tr>
<tr>
<td>ALG</td>
<td>Argument Label Group</td>
</tr>
<tr>
<td>Arg0</td>
<td>Agent</td>
</tr>
<tr>
<td>Arg1</td>
<td>Direct Object/Theme/Patient</td>
</tr>
<tr>
<td>Arg2–5</td>
<td>Arguments not fixed</td>
</tr>
<tr>
<td>Bellmf</td>
<td>Bell Membership Function</td>
</tr>
<tr>
<td>BOW</td>
<td>Bag of Word</td>
</tr>
<tr>
<td>CART</td>
<td>Classification and Regression Trees</td>
</tr>
<tr>
<td>CHAID</td>
<td>Chi-squared Automatic Interaction Detection</td>
</tr>
<tr>
<td>CHK</td>
<td>Chunking</td>
</tr>
<tr>
<td>CoNLL</td>
<td>Conference on Natural Language Learning</td>
</tr>
<tr>
<td>COPS</td>
<td>Copy Protection System</td>
</tr>
<tr>
<td>CS11</td>
<td>Clough and Stevenson Corpus</td>
</tr>
<tr>
<td>DIR</td>
<td>Direction</td>
</tr>
<tr>
<td>DIS</td>
<td>Discourse Connectives</td>
</tr>
<tr>
<td>E</td>
<td>Set of edges connecting with nodes</td>
</tr>
<tr>
<td>E</td>
<td>Edge</td>
</tr>
<tr>
<td>EXT</td>
<td>Extent</td>
</tr>
<tr>
<td>FLIS</td>
<td>Fuzzy Logic Inference System</td>
</tr>
<tr>
<td>FLS</td>
<td>Fuzzy Inference System</td>
</tr>
<tr>
<td>FM</td>
<td>Frequency Model</td>
</tr>
<tr>
<td>G</td>
<td>Graph, Data Graph</td>
</tr>
<tr>
<td>Gaussmf</td>
<td>Gaussian Membership Function</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>HITS</td>
<td>Hyperlink Induced Topic Search</td>
</tr>
<tr>
<td>ID</td>
<td>Identification Number</td>
</tr>
<tr>
<td>IR</td>
<td>Information Retrieval</td>
</tr>
<tr>
<td>LCS</td>
<td>Longest common subsequence</td>
</tr>
<tr>
<td>LOC</td>
<td>Location</td>
</tr>
<tr>
<td>LSI</td>
<td>Latent Semantic Indexing</td>
</tr>
<tr>
<td>LZ</td>
<td>Lempel-Ziv</td>
</tr>
<tr>
<td>MCS</td>
<td>Minimum common supergraph</td>
</tr>
<tr>
<td>Mcs</td>
<td>Maximum common subgraph</td>
</tr>
<tr>
<td>ML-SOM</td>
<td>Multi-Layer Self-Organizing Maps model</td>
</tr>
<tr>
<td>MMCSN</td>
<td>Maximum Minimum Common Subgraph/ Supergraph Normalized Measure</td>
</tr>
<tr>
<td>MNR</td>
<td>Manner</td>
</tr>
<tr>
<td>MOD</td>
<td>Modal verb</td>
</tr>
<tr>
<td>NEG</td>
<td>Negation marker</td>
</tr>
<tr>
<td>NER</td>
<td>Name Entity Recognition</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>O</td>
<td>Adjective</td>
</tr>
<tr>
<td>P</td>
<td>Pattern Graph</td>
</tr>
<tr>
<td>PAN-PC</td>
<td>PAN Plagiarism Corpus</td>
</tr>
<tr>
<td>PNC</td>
<td>Purpose</td>
</tr>
<tr>
<td>POS</td>
<td>Part-of-Speech</td>
</tr>
<tr>
<td>Propbank</td>
<td>Proposition Bank</td>
</tr>
<tr>
<td>P-value</td>
<td>Probability Value</td>
</tr>
<tr>
<td>RFM</td>
<td>Relative Frequency Model</td>
</tr>
<tr>
<td>SC</td>
<td>Structural Characteristic</td>
</tr>
<tr>
<td>SCAM</td>
<td>Stanford Copy Analysis Method</td>
</tr>
<tr>
<td>SIM</td>
<td>Total similarity score values between the original and suspected documents</td>
</tr>
<tr>
<td>SPT</td>
<td>Stanford Parser Tree</td>
</tr>
<tr>
<td>SRL</td>
<td>Semantic Role Labelling</td>
</tr>
<tr>
<td>STA</td>
<td>Semantic Term Annotation</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>Term Frequency–Inverse Document Frequency</td>
</tr>
</tbody>
</table>
TMP - Time
Trap - Trapezoidal membership function
Trimf - Triangular Membership Function
TSK - Takagi-Sugeno-Kang
TTS - Chinese to Taiwanese System
UK - United Kingdom
USA - United States of America
V - Set of nodes (vertices)
V - Node or Vertex
V - Verb
VGj - Vertex Mapping of Data Graph
Vpi - Vertex Mapping of Pattern Graph
VSM - Vector Space Model
# LIST OF APPENDICES

<table>
<thead>
<tr>
<th>APPENDIX</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>An example of verb-specific “afford”</td>
<td>229</td>
</tr>
<tr>
<td>B</td>
<td>CHAID algorithm rules</td>
<td>231</td>
</tr>
<tr>
<td>C</td>
<td>Related publications</td>
<td>235</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Introduction

Plagiarism is defined as the “unacknowledged copying of documents or programs” (Joy and Luck, 1999). It can occur in many areas. For example, companies may look for a competitive advantage in the market, and academicians may need to advance their careers by way of quick publishing. Many empirical studies and analyses have been undertaken by the academic community to deal with student plagiarism. The correct selection of text features is a key aspect in the task of discriminating between plagiarized documents and non-plagiarized documents.

There are many types of plagiarism mentioned by Hermann, Frank, and Bilal, (2006), such as copy and paste, redrafting or paraphrasing of the text, plagiarism of ideas, and plagiarism through translation from one language to another.

Nowadays, many documents are available on the internet and are easy to access. Due to this wide availability, users can easily create a new document by copying and pasting. Sometimes users can reword the plagiarized part by replacing words with their synonyms. This kind of plagiarism is difficult to be detected by using traditional plagiarism detection systems such as copy protection system
(COPS) (Sergey et al., 1995a), Stanford Copy Analysis Method (SCAM) (Sergey et al., 1995b) or CHECK (Antonio et al., 1997).

The biggest challenge in plagiarism detection is to provide plagiarism checking with appropriate algorithms in order to improve the odds of finding instances of plagiarism and to decrease the time spent checking. According to the literature reviewed by Alzahrani, Salim and Abraham (2011), and Bao (2003), current plagiarism detection systems were found to be too slow due to the matching techniques such as string and character matching, and the matching algorithms were dependent on the text’s lexical structure rather than its semantic structure, making it difficult to detect any text that had been paraphrased. The important question for the plagiarism detection problems examined in this study is whether application of new techniques such as Graph-based, Semantic Role Labelling (SRL), Fuzzy Logic and Chi-squared Automatic Interaction Detection (CHAID) algorithms can improve plagiarism detection of texts.

In this study, we propose a new plagiarism detection methods based on Graph-based and Semantic Role Labelling (SRL). We later improved these methods using Fuzzy Logic and CHAID algorithm. The proposed method can detect copy and paste plagiarism, rewording or synonym replacement, changing of word structure in the sentences, as well as modification of the sentence from passive voice to active voice and vice versa.

The organization of this chapter is described as follows: Section 1.2 reviews the background of the problem while section 1.3 will present the problem statement. Then, Section 1.4 will discuss the objective of this study. Section 1.5, 1.6, and 1.7 will focus on the scope of study, significance of study and expected contribution, respectively. Finally, Section 1.8 will describe the organization of this thesis.
1.2 Background of Research

Plagiarism is a form of academic misconduct. It has increased rapidly because it is now quick and easy to reach data and information through electronic documents and the Internet. Plagiarism means using the text written by others which may be re-adjusted by adding or deleting text but without any citation or reference to the original author.

There are many types of plagiarism, such as copy and paste, redrafting or paraphrasing of the text, plagiarism of ideas, and plagiarism through translation from one language to another. According to Adeva, Carroll and Calvo, (2006), at least 10% of student’s work is likely to be plagiarized in USA, Australia and UK universities (Lyon et al., 2006). Another current research project found that 70% of students confess to small instances of plagiarism, and about half of the students studied were guilty of a cheating offence on a written assignment. Additionally, 40% of students confess to using the cut and paste method when completing their assignments (McCabe, 2005).

Differentiating between the plagiarized documents and non-plagiarized documents in an effective and efficient way is one main issue in the field of plagiarism detection. Current methods of plagiarism detection are based on character matching, n-gram, chunks or terms (Alzahrani, et al., 2011; Hermann, et al., 2006; Potthast et al., 2010).

Research institutions and universities require new technology for detecting plagiarism. This is crucial in controlling and marking researchers and students’ essays, homework, papers and reports. Many plagiarism detection tools use character matching or string matching method to detect the plagiarized text (Hermann, et al., 2006; Maurer et al., 2006). However, most of the current software and techniques are not effective in detecting plagiarized text because these tools tend to compare a suspected text with original text using characters matching, some by chunks while others by words. This leads to an exhaustive, time consuming search
(Mozgovoy, 2007). As noted above, the purpose of this study is to propose new techniques for plagiarism detection based on Graph-based Representation, Semantic Role Labelling, Fuzzy Logic technique and Chi-squared Automatic Interaction Detection method.

Text Graph-Based Representation does not only represent the content of a text document as a graph, but it also captures the underlying semantic meaning in terms of the relationships among its concepts. Semantic Role Labelling (SRL) is superior in generating arguments for each sentence semantically (Martha Palmer et al., 2005). WordNet Thesaurus (Fellbaum, 1998) is an excellent tool for extracting the concepts or synonyms for each word in the sentences. Fuzzy Logic is a common expert system application which has proven successful in several predictions and control systems (Wong and Hamouda, 2000). Chi-squared Automatic Interaction Detection algorithm (CHAID) is highly visual and simple to understand and interpret. It uses multi-way splits by default; it needs rather large sample sizes to work effectively, since with small sample sizes the respondent groups can become smaller for reliable analysis (Diepen and Franses, 2006). We employed CHAID to detect an interactions between variables in the plagiarism detection corpus. Using this method it is possible to establish relationships between a ‘dependent variable’ for example, the relationship between total similarity score and other arguments types variables, such as: subject, object, verb, among others. CHAID does this by identifying discrete groups of respondents and, by taking their responses to explanatory variables, seeks to predict what the importance and impact arguments variables will be on the total similarity score variable.

Graph Based Representation relies on the different levels of processing in the text. Zhang (2009) divided graph representation into three levels: the document level, the sentence level and the term level. The representation of these levels within a graph defines the graph node and graph edge. The node can hold either a document or a sentence or a term and the edge is the weight between these levels. The Document level looks at the multi documents in the graph. Here each document in the corpus or web is represented as a node and each relationship or link between two documents is demonstrated as an edge. Prominent examples for document
graph-based representation include a network citation and World Wide Web network. In the network citation the authors can cite from the others’ work by referring to their work. Each author’s paper contains many references while every reference represents a paper or an article or a document. The relationship among these references and author’s paper is that the topics covered in the author’s paper are similar to those in the references. Due to that, each reference is represented as a node and is linked with the main paper that includes it as references by the edge.

Semantic structures or case frames were introduced by Minsky (1974) where common frames were used for common roles and themes such as FrameNet proposed by Baker, Fillmore and Lowe (1998) and PropBank proposed by (Martha Palmer, et al., 2005). A statistical system is trained on the data from the FrameNet project to automatically assign semantic roles (Daniel Gildea and Daniel Jurafsky, 2002). Surdeanu et al., (2003); Pradhan et al., (2004); and Xue, Nianwen and Martha Palmer, (2004) (Xue, et al., 2004) followed this approach by improving sets of features and machine learning methods. Semantic Role Labelling (SRL) by Johansson and Nugues (2008) achieved the best result in terms of F-measure for the corpus evaluation. Barnickel et al.,(2009) introduced a large scale application of neural network based on Semantic Role Labelling for automated relation extraction from biomedical texts. This method mainly used SENNA software (Collobert and Weston, 2007). SENNA can extract the arguments of the sentences and semantic role for the terms based on neural network algorithms where the users can adopt this software to extract the semantic relations between terms in text documents.

Semantic Role Labelling is a process used to identify and label arguments in a text. The underlying idea is that the sentence level semantic analysis of text determines the object and the subject of a text. It can be extended to the characterization of events such as determination of “who” did “what” to “whom,” “where,” “when,” and “how.” The predicate of a clause (usually a verb) establishes “what” took place, and other parts of the sentence express the other arguments of the sentence (such as “who” and “when”). The primary task of Semantic Role Labelling is to identify what semantic relation holds among a predicate and its associate participants or properties, with these relations drawn from a pre-defined list of
possible semantic roles for that predicate or class of predicates. The typical labels used in SRL are an Agent, Patient and Location for the entities participating in an event. Those labels can be extended to more specific arguments such as Time and Place in some text.

Currently, there are several tools using Semantic Role Labelling including the Proposition Bank or Propbank (Martha Palmer, et al., 2005), FrameNet (Baker, et al., 1998) and VerbNet (Karin Kipper et al., 2000). PropBank has thus far obtained the attention of the researchers in SRL technique.

Most matching algorithms are dependent on the text’s lexical structure rather than its semantic structure (Maurer, et al., 2006). Therefore, it is difficult to detect texts that are paraphrased semantically. Based on Maurer, Kappe, and Zaka (2006), the relation between sentences and their semantic content in plagiarism detection based on text semantic analysis method can be solved by the Semantic Role Labelling (SRL). A semantic role is the underlining relationship that a participant has with the main verb in the clause (Payne, 1997); also known as thematic role, semantic case, the theta role (generate grammar), and deep case (case grammar). SRL is a new method in plagiarism detection and it is superior for generating arguments for each sentence semantically.

An improvement to current plagiarism detection methods proposed by this study can be illustrated by using the Graph Based Representation and Semantic Role Labelling for text documents and then selecting the important arguments that can have an effect on plagiarism detection using Fuzzy Logic method and the CHAID algorithm.

Fuzzy Logic (Zadeh, 1965) is a common expert system applications which has proven successful in several predictions and control systems (Wong and Hamouda, 2000). It is usually used to represent ambiguous and vague information. It is an appropriate method for determining the relationship between inputs and the desired outputs of a system. It has the ability to control decision-making based
Fuzzy Logic was used to control a simple laboratory steam engine (Mamdani, 1974). It is a mathematical assumption of ambiguous reasoning that allows it to obtain results and decision-making model in linguistic terms. Fuzzy Logic has become one of the main successful technologies in many applications and sophisticated control systems. For example, Munakata and Jani (1994) mentioned that over a thousand industrial commercial fuzzy applications have been successfully developed in recent years.

Fuzzy Logic resulted in the development of the theory of fuzzy sets. Classical Logic is limited and can only deal with two values – true or false. However, there is a need for a system that can handle partial truths (neither completely true nor completely false). Fuzzy logic, therefore, is an extension classical logic as it generalizes classical logic inference rules which have the ability to deal with approximate reasoning (Klir and Yuan, 1995).

Some research has been done with fuzzy-set with plagiarism detection, for example the study carried by Alzahrani and Salim (2010); Yerra (2005). Matching fragments of text, such as terms and sentences, become ambiguous or approximate, and applies a range of similarity values from one (fully matched) to zero (totally different). In plagiarism detection field, the concept of fuzzy can be represented by considering each term in a text is associated with a fuzzy set that comprises terms with the same meaning, and there is a degree of similarity between terms in a text and the fuzzy set (Yerra and Ng, 2005). Fuzzy-set Information Retrieval (IR) for plagiarism detection is effective since it detects not only exact matches but also similar statements based on the degree of similarity between words in the statement and their fuzzy sets. In order to construct the fuzzy-set and the degree of similarity between words, term-to-term correlation matrix should be constructed before using the fuzzy set Information Retrieval (IR). It should contain the words and their
corresponding correlation factor that measure the degree of similarity (degree of membership between 0 and 1) among different word, such as “vehicle” and “transport.” The fuzzy-set IR technique obtains the degrees of similarity among sentences by computing the correlation factors between any pair of words from two different sentences in their respective documents. Therefore, fuzzy set IR is capable of not only detecting general similarities but also similar patterns between two documents.

The fuzzy set is an elaboration of the traditional set “crisp set” in which each member has a degree of membership to that set as determined by a membership function. The membership function is a function that assigns a membership degree to each member in the target set and the range of membership degree is between zero and one. The computer can translate the linguistic statement into actions based on a set of “IF-THEN” rules of the Fuzzy Logic. The fuzzy IF-THEN rules are normally created in the form of “if A then B” in which the condition is connected with actions where A and B are fuzzy sets. Fuzzy Logic has an advantage in terms of simplicity of development and modification because the rules are well understandable and easy to modify, add new rules, or remove existing rules. One of the objectives of this study is to select the best and most important arguments that can define the plagiarism process and improve the detecting similarity score. Arguments defined as unimportant using FIS will be ignored.

Chi-squared Automatic Interaction Detection or CHAID is an extremely effective predictive statistical method, developed by Kass (1980) used for segmentation. CHAID works by using combining predictor features according to a value that was derived by using statistical test criteria. CHAID then combines values that are similar to the target variables with the remaining, dissimilar values.

CHAID creates a decision tree by using the best predictor to form the first level. Each child node is created, or grown, from a group of features with similar features, like leaves springing from the branches of a tree. This process continues
until the tree has fully grown. Significant statistical tests used stops upon the measurement level of the target field.

Unlike methods that rely on binary trees, the CHAID tree grows wider because it has the ability to use any type of variable and it can use both weight variables and frequency variables.

One of the advantages of CHAID algorithms is that its results are highly visual and simple to understand and interpret (Merel van Diepen and Philip Hans Franses, 2006). Yih-Jeng, Ming-Shing and Chin-Yu, (2008) reports in their research that:

“A CHAID method has many advantages of applying in looking for patterns in complicated datasets. The level of measurement for the dependent variable and predictor variables can be nominal, ordinal, or interval. The predictor variables need not all be measured at the same level (nominal, ordinal, and interval). For the case of missing values in predictor variables, it can be treated as a "floating category" so that partial data can be used whenever possible within the tree.

(Yih-Jeng, Ming-Shing and Chin-Yu, 2008: 366)

Due to the advantages of The CHAID algorithm, we found that it is very good at improving the plagiarism detection results for our proposed methods. The main reason that the CHAID algorithm was employed in our study was that the algorithm can select important features from the data set and adopted this attribute to select important arguments from the sentences.

CHAID is both a statistical model and data mining method. It is belongs to a set of models identified by decision trees. It is used for classification and prediction purposes. In plagiarism detection, we will employ the CHAID algorithm to predicate important and unimportant arguments from the suspected and original documents.
The arguments will be extracted first from the documents based on SRL. Then, the similarity will be calculated based on the Jaccard similarity measure (Jaccard and Paul, 1901) between the arguments. The input of CHAID algorithms will be a set of arguments similarity scores between original and suspected documents and the output of CHAID will be a set of predicated important arguments. One of the advantages of CHAID algorithm is to select the important features form a set of features. In our plagiarism detection method, we adopted the CHAID algorithm to select the important arguments from a set of the extracted arguments using SRL method. The CHAID algorithm usually predicates the important features in form of decision tree, in the same case; the important arguments will be selected in the tree form.

In addition to Fuzzy Logic and The CHAID algorithm, another improvement that can be made to current plagiarism detection techniques can be found in the mechanism that compares the corresponding arguments between two texts, (Subject with Subject, Verb with Verb) which differs from the traditional comparison mechanisms (Subject with Verb, Subject with Adverb, Subject with Adjective and so on). The greatest benefit of employing this type of plagiarism detection is that it can detect copy and paste, semantic plagiarism, rewording or synonym replacement, changing of word structure in the sentences, modifying the sentence from passive voice to active voice and vice versa.

This study implements methods that detects plagiarism by representing text as graph and selecting arguments based on similarity score using important arguments and adjusting the weighting for each argument to improve the total results using Fuzzy Logic and CHAID algorithm with Semantic Role Labelling to extract key arguments of the original and suspected texts and estimating the relevance of suspected sentences by capturing the main content and the semantic content available in sentences using Graph-based and Semantic Role Labelling.
1.3 Problem Statement

One of the important algorithms in plagiarism detection is the matching algorithms. Matching algorithms focus on the text lexical structure rather than semantic structure. As a result, it is difficult to detect any text paraphrased semantically (Alzahrani, et al., 2011; Hermann, et al., 2006; Potthast, et al., 2010). Additionally, current plagiarism detection systems were also found to be too slow and takes too much time for each checking (Mozgovoy, 2007).

This research is concerned about a semantic matching algorithm to answer the following research question:

(i) Can graph-based representation of text documents be used for plagiarism detection?
(ii) Can the semantic role labelling (SRL) with the graph-based representation enhance plagiarism detection?
(iii) Can the arguments weight adjustment to select the important arguments in sentences be used to improve graph-based with SRL plagiarism detection?
(iv) How can fuzzy logic and Chi-squared Automatic Interaction Detection algorithm (CHAID) define important arguments that need to be used for plagiarism detection?
(v) Can the combination of fuzzy logic and Chi-squared Automatic Interaction Detection algorithm (CHAID) algorithm with semantic role labelling and graph-based give better plagiarism detection technique?

1.4 Objectives

The aim of this research is to show how to combine a graph-based method, SRL, Fuzzy Logic, and The CHAID algorithm to form new techniques to detect plagiarism. To achieve this goal, the following objectives will be aimed:
1- To develop a graph-based technique that can be used to represent text documents as graph for plagiarism detection.

2- To investigate the use of Semantic Role Labelling (SRL) in the graph-based representation to enhance plagiarism detection.

3- To identify important arguments that can be used to improve detection of plagiarism using weight arguments, fuzzy logic technique, and the CHAID algorithm.

1.5 Research Scope

The preceding section mentioned the objectives of this research which focus on how to produce a good plagiarism detection algorithm. The following aspects are the scope of study for the mentioned objectives.

1- Plagiarism detection using graph-based representation.

2- An improved plagiarism detection based on semantic role labelling.

3- An improved plagiarism detection based on fuzzy semantic role labelling and CHAID algorithm.

4- Evaluation of the performance of the proposed methods using PAN-PC-09 and short answers questions plagiarism corpus and compare with other approaches such as fuzzy semantic-based string similarity, longest common subsequence (LCS), and semantic-based similarity.
1.6 Significance of Research

Much of the research done in the plagiarism detection field is based on character matching method and fingerprint method. This study will introduce plagiarism detection methods which use Graph-based Representation and SRL with Fuzzy Logic and CHAID technique. The text graph-based representation used to represent the content of a text document as a graph and used also to capture the underlying semantic meaning in terms of the relationships among its concepts. SRL was used to analyze the sentences semantically and the WordNet Thesaurus was used to extract the concepts or synonymies for each word inside the sentences. Fuzzy Logic technique focused on the fuzziness of argument terms in the sentence and used it as an optimisation technique with a CHAID algorithm to select important sentence arguments.

1.7 Study Contributions

The expected contribution of this study can be explained as follows:

1. A new plagiarism detection method using graph to represent text document with semantic meaning in terms of the relationships among its concepts. WordNet Thesaurus will be used to extract concepts or synonymies for each word inside the sentences.

2. A new plagiarism detection scheme based on Semantic Role Labelling which can compare sentences semantically.

3. An improved plagiarism detection method based on Arguments Weighting Scheme.

4. A new plagiarism detection method using Fuzzy Logic and CHAID algorithm
1.8 Thesis Organization

This thesis organized with eight chapters. These chapters are as follows:

Chapter 1: Introduction:

The introductory chapter of this thesis provides a brief overview of some of the issues that of concern to those working in the field of plagiarism detection. This chapter will also look at the goals, the scope of this study as well as examining the contributions this research can make to the field of plagiarism detection.

Chapter 2: Literature Review:

This chapter evaluates state-of-the-art approaches in plagiarism detection. Techniques such as Semantic Role Labelling, Fuzzy Logic, and the CHAID algorithm will be reviewed. In addition, a plagiarism detection evaluation measurements and datasets will be covered too.

Chapter 3: Methodology:

This chapter describes the methodology and principal experiments used to obtain the objectives of this research study. Topics will include; Text Graph-based Representation, Semantic Role Labelling, Fuzzy Logic and Chi-squared Automatic Interaction Detection.

Chapter 4: Plagiarism Detection Using Graph-Based Representation Detection:

This chapter proposes a plagiarism detection technique based on graph based representation. In this method, a text document is represented as a graph and used to captures the underlying semantic meaning in terms of the relationships among its concepts. The comparison between the documents is calculated according to the similarity between the terms and concepts of the sentences.
Chapter 5: *Plagiarism Detection Based on Semantic Role Labelling:*

This chapter introduces a plagiarism detection method using SRL. SRL is used to analyze the sentences semantically and WordNet Thesaurus is used to extract the concepts or synonymies for each word inside the sentences. This method can detect a plagiarism after terms arguments are extracted. The comparison is calculated according to the semantic position of the terms in the sentences. In addition, this chapter will also propose an improved plagiarism detection scheme based on semantic role labelling conducted by using and argument weight scheme. Arguments behaviours will be studied to select important arguments that can have an effect on plagiarism detection. This improvement reflects the important arguments that should be used in comparison process rather than all extracted arguments using SRL.

Chapter 6: *An Improved Plagiarism Detection Technique Based On Fuzzy Semantic Role Labelling:*

This chapter introduces an improvement of SRL plagiarism detection technique using Fuzzy Logic. Fuzzy Logic will be used as an optimisation technique by selecting the important arguments in a sentence. Selection for each argument generated by the Fuzzy Logic in order to select important arguments will also be discussed.

Chapter 7: *An Improved Plagiarism Detection Method Based On Semantic Role Labelling and Chi-Squared Automatic Interaction Detection:*

This chapter introduces an improvement of text similarity checking and plagiarism detection method based on a Semantic Role Labelling (SRL) and Chi-squared Automatic Interaction Detection (CHAID).

Chapter 8: *Conclusion and Future Work:*

Chapter 8 will review the conclusions of the research discussed throughout this study. This section will also put forward recommendations for future studies.
REFERENCES


Dahlmeier, D., and Ng, H.T. (2010). *Domain adaptation for semantic role labeling in the biomedical domain*. Bioinformatics, 26(8), 1098-1104.


