TABLE OF CONTENTS

CHAPTER	CONTENT		PAGE
	TITLE		**i**
	DECLARATION		**ii**
	DEDICATION		**iii**
	AKNOWLEDGEMENT		**iv**
	ABSTRACT		**v**
	ABSTRAK		**vi**
	TABLE OF CONTENTS		**vii**
	LIST OF TABLES		**x**
	LIST OF FIGURES		**xii**
	LIST OF APPENDICES		**xv**
1	**INTRODUCTION**		**1**
	1.1	Introduction	1
	1.2	Problem Background	4
	1.3	Problem Statement	5
	1.4	Project Objectives	5
	1.5	Project Scope	6
	1.6	Project Justification	6
	1.7	Report Organization	7
2	**LITERATURE REVIEW**		**8**
	2.1	Introduction	8

2.2 Document Plagiarism 9

2.3 Plagiarism Detection Methods 10

2.3.1 Detection based on Stylometry Analysis 11

2.3.2 Detection based on Documents Comparison 12

2.3.2.1 Semantic-Based Detection 12

2.3.2.2 Syntactic-Based Detection 14

2.4 Existing Web-Based Plagiarism Detection Tools 16

2.5 Semantic Networks 20

2.6 Document Preprocessing 23

2.6.1 Tokenization 24

2.6.2 Stop-word Removal 24

2.6.3 Stemming 24

2.6.4 Document Chunking 25

2.7 Document Representations & Similarity Measures 28

2.7.1 Semantic Based-Representation 29

2.7.2 Syntactic Based-Representation 34

2.7.2.1 Fingerprinting 34

2.7.2.2 Term Weighting Schemes 37

2.7.2.3 N-Grams 38

2.8 Algorithms for Approximate Similarity 40

2.8.1 Signature Scheme Algorithms 40

2.8.2 Inverted Index-Based Algorithms 47

2.9 Discussion and Summary 51

3 METHODOLOGY 53

3.1 Introduction 53

3.2 Operational Framework 54

3.2.1 Initial Study and Literature Review 55

3.2.2 Corpus Preparation 55

3.2.3 Document Preprocessing 57

3.2.4 Applying Plagiarism Detection Techniques 58

3.2.4.1 Semantic Relatedness Approach 58

3.2.4.2 N-grams Approach 66

3.2.5 Web Document Retrieval 70

3.2.6 Implementation 73

3.2.7 Findings Evaluation 75

4 EXPERIMENTAL RESULTS 78

4.1 Introduction 78

4.2 Information about the Corpus 79

4.3 Sentence-to-Sentence similarity 82

4.3.1 N-grams Approach 82

4.3.2 Semantic Relatedness Approach 85

4.4 Results and Comparisons 97

4.4.1 Results of Corpus Sentence Retrieval 97

4.4.2 Results of Web Document Retrieval 102

4.4.3 Comparison with Existing Tools 111

4.5 Discussion and Summary 115

5 CONCLUSION 117

5.1 Introduction 117

5.2 Achievements and Constraints 118

5.3 Future work 119

5.4 Summary 121

REFERENCES 122

APPENDICES A-E 128-150

LIST OF TABLES

TABLE NO.	TITLE	PAGE
2.1	Properties of some existing plagiarism detection tools based on [59]	19
2.2	Some of the relations between concepts in WordNet (N=noun, V=Verb, Adj=adjective, Adv=adverb)	21
2.3	Statistics about WordNet 2.1	22
2.4	Common similarity measures between binary vectors	39
2.5	Common similarity measures between sets	39
2.6	Common factors that influence the performance of Inverted index algorithms	48
2.7	Signature-based versus Inverted index-based algorithms	52
3.1	Integrated libraries in the project and their roles	73
4.1	Number of plagiarized sentences in documents pairs (query-vertical)/(source -horizontal)	80
4.2	Statistics about the corpus and query documents	81
4.3	Statistics about part-of-speech tagging	81
4.4	Part-of-speech tagging of $s1$ and $s2$	86

4.5 Shortest path between word pairs in the joint set and *T2* (“-1” no path exists,”=” equals, “?” not of the same part of speech) 91

4.6 Subsumer depth between word pairs in the joint set and *T2* (-1 no depth exists,= equals, ? not of the same part of speech) 91

4.7 Word-to-word similarity between the joint set and *T*2 92

4.8 Raw semantic and order vectors for *T1* 93

4.9 Raw semantic and order vectors for *T2* 94

4.10 Information contents of word in the joint set 95

4.11 N-grams recall rate in 610 corpus documents with 0.5 cutoff threshold 97

4.12 Recall rate when increasing number of documents with 0.5 cutoff threshold 99

4.13 Precision, Recall, and Harmonic Mean (F-measure) in 110 corpus documents with 0.5 cutoff threshold 100

4.14 Recall rate across similarities in 110 corpus documents 100

4.15 Semantic-R recall rate in 110 corpus documents with Alpha=0.2, Beta=0.45 and 0.8 cutoff threshold 102

4.16 Results of using 3-grams searching with 64 results/query limit 104

4.17 Results of using weighted 3-grams with 64 results/query limit 105

4.18 Results of using selective searching with 64 results/query limit. 108

4.19 Results of using 3-grams searching with 8 results/query limit. 109

4.20 Results of using weighted 3-grams with 8 results/query limit. 110

4.21 Results of using selective searching with 8 results/query limit. 111

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
1.1	Hierarchical semantic knowledge base[57]	3
2.1	Taxonomy of plagiarism detection methods	10
2.2	An example HTML report generated by DocCop	16
2.3	A sample report with timeline returned from Plagium	17
2.4	The interface of EVE2 for Web searching	18
2.5	An example of a synset in WordNet	20
2.6	Bipolar adjective structure (→= similarity, --► =antonymy)	22
2.7	N-Unit non-overlapped chunking strategy with N=5	26
2.8	N-Unit chunking with K-overlap where N=5 and K=2	26
2.9	An example given by [55] to illustrate the different between most specific subsumers in WordNet.	32
2.10	Framework for signature-based algorithms [7]	41
2.11	Two documents represented as records	41

2.12 Prefix Filter scheme of Figure 2.5 with 80% Overlap similarity threshold. 42

2.13 Two vectors with hamming distance = 4 43

2.14 Two vectors with hamming distance ≤ k 4 must agree on one of the k + 1 partitions 44

2.15 The two vectors in Figure 2.14 with hamming distance = 8 and agree on one partition 44

2.16 Enumeration scheme for two vectors with hamming distance=3 45

2.17 Formal specification of PartEnum[7] 46

2.18 Formal specification of All-Pairs[6] 49

3.1 Operational Framework 54

3.2 The procedure used in obtaining the semantic attributes between two concepts 63

3.3 The algorithm for semantic relatedness between a pair of sentences 65

3.4 Binary vector representation of a sentence 66

3.5 An inverted index implementation for Cosine similarity[6] 67

3.6 An inverted index implementation for Jaccard Similarity 70

3.7 An inverted index implementation for Dice coefficients 70

3.8 The procedure of evaluating Web document retrieval techniques 72

3.9 Response format from querying the Google API 74

4.1 The inverted index for document *Q* 84

4.2 The hypernym trees for the words: *Idea (5 senses), Learning (2 senses), Adaptation (3 senses), Evolution (2 senses) and Biology (3 senses).* 89

4.3	The hypernym trees for the words: *Idea (5 senses), Learning (2 senses), Adaptation (3 senses), Evolution (2 senses) and Aspect (3 senses)*	90
4.4	Recall rate (y-axis) across similarities (x-axis) in 110 corpus documents	101
4.5	Recall rate in one-to-one exact copies	113
4.6	Recall rate in one-to-one plagiarized by synonym replacing	114
4.7	Recall rate in one-to-many plagiarized by synonym replacing	114

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	Stop-words and their corresponding frequencies in the Brown corpus	128
B	Information about ScienceDirect source documents	129
C	Information about Wikipedia corpus documents	130
D	The PENN TreeBank English POS tag set and their mappings	144
E	Examples of original/plagiarized sentence pairs and the corresponding similarities based on equation 3.6	145