

**CROSS DOCUMENT RELATION IDENTIFICATION FOR  
MULTI DOCUMENT SUMMARIZATION BASED ON  
ENHANCED CASE BASED REASONING FRAMEWORK**

**YOGAN JAYA KUMAR**

**UNIVERSITI TEKNOLOGI MALAYSIA**

# UNIVERSITI TEKNOLOGI MALAYSIA

## DECLARATION OF THESIS / ~~POSTGRADUATE PROJECT REPORT AND COPYRIGHT~~

Author's full name : YOGAN JAYA KUMAR

Date of birth : 26/12/1981

Title : CROSS DOCUMENT RELATION IDENTIFICATION FOR MULTI DOCUMENT SUMMARIZATION BASED ON ENHANCED CASE BASED REASONING FRAMEWORK

Academic Session : 2013/2014 (2)

I declare that this thesis is classified as :

- CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)\*
- RESTRICTED** (Contains restricted information as specified by the organization where research was done)\*
- OPEN ACCESS** I agree that my thesis to be published as online open access (full text)

I acknowledged that Universiti Teknologi Malaysia reserves the right as follows:

1. The thesis is the property of Universiti Teknologi Malaysia.
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.



SIGNATURE

811226-10-5729

(NEW IC NO. /PASSPORT NO.)

Date: March 3, 2014

Certified by :




SIGNATURE OF SUPERVISOR

Prof. Dr. Naomie Salim  
NAME OF SUPERVISOR

Date: March 3, 2014

“I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of the degree of Doctor of Philosophy (Computer Science)”

Signature :   
Name of Supervisor : Prof. Dr. Naomie Salim  
Date : March 3, 2014

**BAHAGIAN A – Pengesahan Kerjasama\***

Adalah disahkan bahawa projek penyelidikan tesis ini telah dilaksanakan melalui kerjasama antara \_\_\_\_\_ dengan \_\_\_\_\_

Disahkan oleh:

Tandatangan : ..... Tarikh : .....

Nama : .....

Jawatan : .....  
(Cop rasmi)

*\* Jika penyediaan tesis/projek melibatkan kerjasama.*

---

---

**BAHAGIAN B – Untuk Kegunaan Pejabat Sekolah Pengajian Siswazah**

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar : **Prof. Dr. Narayanan Kulathuramaiyer**  
**Faculty of Computer Science and Information**  
**Technology,**  
**Universiti Malaysia Sarawak,**  
**94300 Kota Samarahan,**  
**Sarawak**

Nama dan Alamat Pemeriksa Dalam : **Prof. Dr. Ali bin Selamat**  
**Research Alliance K-Economy,**  
**UTM Johor Bahru.**

Nama Penyelia lain (jika ada) : -

Disahkan oleh Timbalan Pendaftar di Sekolah Pengajian Siswazah:

Tandatangan : ..... Tarikh : .....

Nama : **ZAINUL RASHID BIN ABU BAKAR**  
.....

CROSS DOCUMENT RELATION IDENTIFICATION FOR MULTI  
DOCUMENT SUMMARIZATION BASED ON ENHANCED CASE BASED  
REASONING FRAMEWORK


YOGAN JAYA KUMAR

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 2014

I hereby declare that this thesis entitled “*Cross Document Relation Identification For Multi Document Summarization Based On Enhanced Case Based Reasoning Framework*” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature :  .....

Name : Yogan Jaya Kumar

Date : March 3, 2014

*To my beloved father, mother, wife, sisters and brothers*

## ACKNOWLEDGMENTS

Firstly, I would like to express my gratitude to my supervisor, Professor Dr. Naomie Salim, for her continuous support and availability on the development of my research study, especially because she has never been too busy to keep an eye on my progress in spite of her numerous obligations. She has greatly helped me in a lot of ways throughout this study. The most important lesson I learnt from her is that whatever you work, work with dedication and sincerity. Again, I owe her my deepest thanks.

In addition, I am grateful to my colleagues at my research lab, especially those in text research group, for the useful discussions, comments and knowledge sharing. Thanks also to my friends here at UTM, whom have in one way or another helped, encouraged and motivated me during the progression of my study.

I am also grateful to the members of Soft Computing Research Group as well as to all staff of the Faculty of Computing, UTM, for their kind cooperation during my study and stay here. I would also like to thank Universiti Teknikal Malaysia Melaka (UTeM) and the Ministry of Higher Education in Malaysia (MOHE) for providing the scholarship for this PhD study.

Last but not least, I like to thank my parents, my wife, my sisters and brothers for their support, guidance, love and prayers. Thank you for all the encouragement you have given me. Once again, thank you.



## ABSTRACT

Documents which are available through online search often provide readers with large collection of texts. In the context of news documents, different news sources reporting on the same event usually contain common components that make up the main story of the news. This study aims to produce high quality multi document summaries by taking into account the generic components of a news story within a specific domain. Since this study involves multiple documents, the research further investigates the automatic identification of cross-document relations from un-annotated text documents, where the case based reasoning (CBR) classification model is proposed. Cross-document relations are used to identify highly relevant sentences to be included in the summary. With the aim to improve the cross-document relation identification, genetic algorithm (GA) is integrated to enhance the CBR classifier. GA is used to scale the relevance of the data features used by the CBR classifier. Following that, this research proposes two new sentence scoring mechanism based on the identified cross-document relations. The first approach is based on a voting technique named votCombMAX which gives votes to sentences based on the relationship types between sentence pairs. The second approach investigates the benefits of fuzzy reasoning over the identified cross-document relations; since not all cross-document relation types have positive effect towards summary generation. In this study, the Document Understanding Conference (DUC) 2002 data sets are used; and as for the evaluation, the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) evaluation metrics are used. The evidence from this study showed that the proposed methods yield significant improvement over the mainstream methods.

## ABSTRAK

Dokumen-dokumen yang diperolehi melalui carian internet sering menyediakan pembaca dengan koleksi teks yang luas. Dalam konteks dokumen berita, sumber berita berbeza yang melaporkan peristiwa yang sama biasanya mengandungi komponen umum yang membentuk cerita utama berita tersebut. Kajian ini menghasilkan rumusan multi-dokumen berkualiti tinggi dengan mengambil kira komponen generik berita dalam domain spesifik. Oleh kerana kajian ini melibatkan multi-dokumen, kaedah pengenalpastian hubungan merentas-dokumen secara automatik dari dokumen teks yang tidak ditandakan telah dikaji, di mana model klasifikasi penaakulan berasas kes (CBR) telah dicadangkan. Hubungan merentas-dokumen digunakan untuk mengenal pasti ayat yang relevan untuk disertakan ke dalam rumusan. Dengan matlamat untuk memperbaiki pengenalpastian hubungan merentas-dokumen, algoritma genetik (GA) telah diintegrasikan untuk meningkatkan klasifikasi CBR. GA digunakan untuk menskalakan kerelevanan ciri-ciri data yang digunakan oleh CBR. Berikutan itu, kajian ini mencadangkan dua mekanisme baru penskoran ayat berdasarkan hubungan merentas-dokumen yang telah dikenal pasti. Pendekatan pertama adalah berdasarkan teknik undian bernama *votCombMAX* yang memberikan undi kepada ayat berdasarkan jenis hubungan antara pasangan ayat. Pendekatan kedua menyiasat manfaat penaakulan kabur ke atas hubungan merentas-dokumen yang telah dikenalpasti; kerana bukan semua jenis hubungan mempunyai kesan positif terhadap generasi rumusan. Dalam kajian ini, set data dari *Document Understanding Conference (DUC) 2002* telah digunakan; dan untuk tujuan penilaian, metrik penilaian *Recall-Oriented Understudy for Gisting Evaluation (ROUGE)* telah digunakan. Keputusan yang diperolehi melalui kajian ini menunjukkan bahawa kaedah yang dicadangkan mencapai prestasi yang lebih baik berbanding dengan kaedah-kaedah lain yang sedia ada.

## TABLE OF CONTENTS

<b>CHAPTER</b>	<b>TITLE</b>	<b>PAGE</b>
	<b>DECLARATION</b>	ii
	<b>DEDICATION</b>	iii
	<b>ACKNOWLEDGEMENT</b>	iv
	<b>ABSTRACT</b>	v
	<b>ABSTRAK</b>	vi
	<b>TABLE OF CONTENTS</b>	vii
	<b>LIST OF TABLES</b>	xii
	<b>LIST OF FIGURES</b>	xv
	<b>LIST OF ABBREVIATION</b>	xix
	<b>LIST OF APPENDICES</b>	xxi
<b>1</b>	<b>INTRODUCTION</b>	<b>1</b>
	1.1 Introduction	1
	1.2 Problem Background	4
	1.3 Problem Statement	8
	1.4 Objectives of Study	9
	1.5 Research Scope	10
	1.6 Significance of Study	11
	1.7 Expected Contribution of Study	12
	1.8 Thesis Organization	12
<b>2</b>	<b>LITERATURE REVIEW</b>	<b>15</b>
	2.1 Introduction	15
	2.2 Overview of Text Summarization	16
	2.3 Approaches to Sentence Extraction	19
	2.3.1 Frequency based Approach	19
	2.3.2 Feature based Approach	21
	2.3.3 Machine Learning Approach	25

2.4	Domain Specific Summarization	27
2.4.1	Medical Summarization	27
2.4.2	News Summarization	29
2.4.3	Email/Blog Summarization	30
2.5	Multi Document Summarization	32
2.5.1	Cluster based Method	34
2.5.2	Graph based Method	36
2.6	Related Works	38
2.6.1	Cross-Document Relations	39
2.6.1.1	CST for Summarization	41
2.6.1.2	Identification of CST Relations	42
2.6.2	Case Based Reasoning Approach	43
2.6.2.1	CBR for Classification	45
2.6.3	Genetic Learning Algorithm	46
2.6.4	Voting Techniques	52
2.6.5	Fuzzy Reasoning System	55
2.6.5.1	Fuzzy Membership Function	56
2.6.5.2	Fuzzy Inference System	59
2.7	Summary	61
<b>3</b>	<b>METHODOLOGY</b>	<b>62</b>
3.1	Introduction	62
3.2	Research Design	62
3.3	Research Operational Framework	64
3.3.1	Phase 1: Preliminary Study and Data Preparation	66
3.3.2	Phase 2: Cross-Document Relation Identification Using Case Based Reasoning (CBR) Approach	69
3.3.3	Phase 3: Genetic-CBR Approach for Improving Cross-Document Relation Identification	70
3.3.4	Phase 4: Multi Document Summarization based on News Components Using Voting Technique	71

3.3.5	Phase 5: Fuzzy Cross-Document Relation for Multi Document Summarization	72
3.3.6	Phase 6: Report Writing	72
3.4	Experimental Evaluation	73
3.5	Selected Benchmark Methods for Comparison	75
3.6	Summary	77
<b>4</b>	<b>CROSS-DOCUMENT RELATION IDENTIFICATION USING CASE BASED REASONING APPROACH</b>	<b>78</b>
4.1	Introduction	78
4.2	Overview of Approach	79
4.3	Feature Extraction	81
4.4	Case Based Reasoning Approach	82
4.4.1	Case Representation	83
4.4.2	Cross-Document Relation Identification	85
4.5	Experimental Setting	88
4.6	Experimental Results	88
4.7	Discussion	93
4.8	Summary	95
<b>5</b>	<b>GENETIC-CBR APPROACH FOR IMPROVING CROSS-DOCUMENT IDENTIFICATION RELATION</b>	<b>96</b>
5.1	Introduction	96
5.2	Overview of Approach	97
5.3	Feature Weighting Using Genetic Based Learning Algorithm	98
5.3.1	Chromosome and Initial Population Construction	101
5.3.2	Fitness Function Design	102
5.3.3	Selection Operation	102
5.3.4	Reproduction Operations	104
5.3.5	Termination Criteria	105

5.4	Case Retrieval Using Weighted Similarity Measure	105
5.5	Experimental Setting	108
5.6	Experimental Results	109
5.7	Discussion	113
5.8	Summary	114
<b>6</b>	<b>MULTI DOCUMENT SUMMARIZATION BASED ON NEWS COMPONENTS USING VOTING TECHNIQUE</b>	<b>116</b>
6.1	Introduction	116
6.2	Overview of Approach	117
6.3	Generic Components of News Documents	119
6.4	Component Sentence Extraction	120
6.5	Cross-Document Relation Identification	122
	6.5.1 Feature Extraction	123
	6.5.2 Genetic-CBR Model Implementation	124
6.6	Sentence Scoring Using Voting Technique	126
	6.6.1 Combining Component Sentence Score	128
6.7	Experimental Setting	129
6.8	Experimental Results	130
6.9	Discussion	136
6.10	Summary	140
<b>7</b>	<b>MULTI DOCUMENT SUMMARIZATION BASED ON FUZZY CROSS-DOCUMENT RELATION</b>	<b>141</b>
7.1	Introduction	141
7.2	Overview of Approach	142
7.3	Fuzzy Reasoning Implementation	144
	7.3.1 Knowledge Base Construction	145
	7.3.2 Rule base Construction	149
	7.3.3 Fuzzy Inference	151
7.4	Sentence Scoring and Selection	154
	7.4.1 Combining Component Sentence Score	154

7.5	Experimental Setting	155
7.6	Experimental Results	156
7.7	Discussion	162
7.8	Summary	165
<b>8</b>	<b>CONCLUSION AND FUTURE WORK</b>	<b>166</b>
8.1	Introduction	166
8.2	Proposed Methods	166
8.2.1	Cross-Document Relation Identification Using Case Based Reasoning Approach	167
8.2.2	Genetic-CBR Approach for Improving Cross-Document Relation Identification	168
8.2.3	Multi Document Summarization Based on News Components Using Voting Technique	169
8.2.4	Multi Document Summarization Based on Fuzzy Cross-Document Relation	170
8.3	Contribution of the Study	171
8.4	Future Work	174
8.5	Summary	175
	<b>REFERENCES</b>	<b>177</b>
	Appendices A-F	193-211

## LIST OF TABLES

<b>TABLE NO.</b>	<b>TITLE</b>	<b>PAGE</b>
2.1	The different dimensions of text summarization	17
2.2	Weight learning methods	24
2.3	Examples of CST relations (Zhang, Blair-Goldensohn et al. 2002).	40
2.4	Voting techniques in expert search (Macdonald and Ounis 2008).	54
3.1	Statistic of DUC 2002 Data Set (Over and Liggett 2002).	67
3.2	Population, Intervention, Comparison and Outcome (PICO).	77
4.1	Description of cross-document relations used in this study.	79
4.2	Examples of CST relationship between sentences (Zhang, Blair-Goldensohn et al. 2002).	80
4.3	An example of case representation.	85
4.4	An example of similarity computation between cases.	87
4.5	Precision, recall, and F-measure of CBR classification.	89
4.6	Precision, recall, and F-measure of SVM classification.	90
4.7	Precision, recall, and F-measure of NN classification.	91
4.8	Comparison of F-measures between SVM, NN and CBR.	92
5.1	Fitness value and selection probability.	103
5.2	Precision, recall, and F-measure of Genetic-CBR classification.	110



5.3	Comparison of F-measures between SVM, NN, CBR and Genetic-CBR.	111
6.1	Example of sentence pairs relations for a component cluster with size N.	126
6.2	Description of summarization methods evaluated in this study.	130
6.3	Summarization results comparison based on average recall, precision and F-measure using ROUGE-1.	131
6.4	Summarization results comparison based on average recall, precision and F-measure using ROUGE-2.	131
6.5	Summarization results comparison based on average recall, precision and F-measure using ROUGE-S.	132
6.6	Summarization results comparison based on average recall, precision and F-measure using ROUGE-SU.	132
6.7	Paired Samples Test between Method V1 and V2.	135
6.8	Paired Samples Test between Method V1 and G1.	135
6.9	Paired Samples Test between Method V1 and C1.	136
6.10	Paired Samples Test between Method G1 and G2.	136
6.11	Examples of CST relationship between actual sentences (S1 and S2).	138
7.1	Input and output linguistic variables definitions.	147
7.2	Part of the constructed fuzzy rules.	150
7.3	Description of summarization methods evaluated in this study.	156
7.4	Summarization results comparison based on average recall, precision and F-measure using ROUGE-1.	156
7.5	Summarization results comparison based on average recall, precision and F-measure using ROUGE-2.	157
7.6	Summarization results comparison based on average recall, precision and F-measure using ROUGE-S.	157
7.7	Summarization results comparison based on average recall, precision and F-measure using ROUGE-SU.	158
7.8	Summarization results comparison for methods F1 and V1 based on average F-measure using ROUGE.	160
7.9	Paired Samples Test between Method F1 and F2.	161
7.10	Paired Samples Test between Method F1 and G1.	162

7.11	Paired Samples Test between Method F1 and C1.	162
7.12	Examples of CST relationship between actual sentences (S1 and S2).	164

## LIST OF FIGURES

<b>FIGURE NO.</b>	<b>TITLE</b>	<b>PAGE</b>
2.1	Sample document.	16
2.2	Sample summary.	17
2.3	A generalized architecture of a feature based summarizer.	23
2.4	Feed-forward network model after training (Khosrow, 2004).	26
2.5	Comment-oriented blog summarization (Hu, Sun et al. 2007).	31
2.6	Example of news article collection.	33
2.7	Example of extracted summary.	33
2.8	A generalized architecture for cluster based summarization.	34
2.9	Example graph as depicted in (Erkan and Radev 2004). Each node represents a sentence.	37
2.10	CST general schema (Radev 2000).	39
2.11	Sentence links between different document sources.	40
2.12	The CBR cycle, depicted from (Aamodt and Plaza 1994).	44
2.13	Example of chromosome representation.	47
2.14	Single point crossover.	48
2.15	Two point crossover.	49
2.16	Scattered crossover.	49
2.17	Flip bit mutation.	50
2.18	Interchange mutation.	51
2.19	Gaussian mutation.	51
2.20	General genetic algorithm procedure.	52
2.21	Example from expert search, depicted from (Macdonald and Ounis 2008)	53

2.22	Range of logical values in Boolean and fuzzy logic.	55
2.23	Typical membership functions of linguistic variable “Age”.	56
2.24	Triangular membership functions.	57
2.25	Trapezoidal membership functions.	57
2.26	Generalized bell-shape membership functions.	58
2.27	Gaussian membership functions.	58
2.28	Fuzzy inference system general architecture.	60
3.1	Research operational framework.	65
3.2	Overview of research study.	66
4.1	Overview of the proposed approach for cross-document relation identification.	80
4.2	The general phases in a case base reasoning process.	83
4.3	The process of generating cases from sentence pairs.	84
4.4	CBR approach for cross-document relation identification.	86
4.5	Performance of CBR classification.	89
4.6	Performance of SVM classification.	90
4.7	Performance of NN classification.	91
4.8	Comparison of F-measures between SVM, NN and CBR.	92
4.9	Performance comparison between SVM, NN and CBR.	93
4.10	Accuracy comparison between SVM, NN and CBR.	93
5.1	Overview of the proposed Genetic-CBR approach for cross-document relation identification.	97
5.2	Genetic algorithm procedure.	99
5.3	Structure of chromosome.	101
5.4	Chromosome population.	101
5.5	Mapping of individuals into segments.	103
5.6	Stochastic universal sampling.	104
5.7	Reproduction based on Scattered crossover.	105
5.8	Gaussian mutation of parent to form a child.	105
5.9	Genetic-CBR approach for cross-document relation identification.	107
5.10	Optimal feature weights after feature weighting.	110

5.11	Performance of Genetic-CBR classification.	111
5.12	Comparison of F-measures between SVM, NN, CBR and Genetic-CBR.	112
5.13	Performance comparison between SVM, NN, CBR and Genetic-CBR.	112
5.14	Accuracy comparison between SVM, NN, CBR and Genetic-CBR.	113
6.1	General architecture of the proposed approach.	118
6.2	Component sentence extraction.	121
6.3	Snippet of JAPE grammar.	122
6.4	Sentence links between different document sources.	123
6.5	Genetic-CBR approach for cross-document relation identification from component clusters.	125
6.6	An example of initial score computation.	127
6.7	Summarization results comparison based on average recall, precision and F-measure using ROUGE-1.	133
6.8	Summarization results comparison based on average recall, precision and F-measure using ROUGE-2.	133
6.9	Summarization results comparison based on average recall, precision and F-measure using ROUGE-S.	134
6.10	Summarization results comparison based on average recall, precision and F-measure using ROUGE-SU.	134
6.11	Example of human produced summary.	139
6.12	Example of summary generated by the proposed model (V1).	139
7.1	General architecture of the proposed approach.	143
7.2	Fuzzy reasoning system.	145
7.3	Gaussian membership function.	146
7.4	The Gaussian membership functions for the input variable Identity.	148
7.5	The Gaussian membership functions for the input variable Subsumption.	148
7.6	The Gaussian membership functions for the input variable Description.	148
7.7	The Gaussian membership functions for the input variable Overlap.	149

7.8	The Gaussian membership functions for the output variable Sentence Score.	149
7.9	Surface view for a Input (Subsumption and Identity) output (Score), b (Overlap and Subsumption) output (Score), c (Overlap and Identity) output (Score), d (Overlap and Description) output (Score), e (Description and Identity) output (Score) and f (Description and Subsumption) output (Score).	152
7.10	Summarization results comparison based on average recall, precision and F-measure using ROUGE-1.	158
7.11	Summarization results comparison based on average recall, precision and F-measure using ROUGE-2.	159
7.12	Summarization results comparison based on average recall, precision and F-measure using ROUGE-S.	159
7.13	Summarization results comparison based on average recall, precision and F-measure using ROUGE-SU.	160
7.14	Summarization results comparison for methods F1 and V1 based on average f-measure using ROUGE.	161
7.15	Example of human produced summary.	164
7.16	Example of summary generated by the proposed model (F1).	165

## LIST OF ABBREVIATIONS

AVG-F	-	Average F-measure
AVG-P	-	Average Precision
AVG-R	-	Average Recall
C1	-	Method based on Clustering
CBR	-	Case Based Reasoning
Com	-	Component
CS	-	Cosine Similarity
CST	-	Cross-document Structure Theory
DUC	-	Document Understanding Conference
F1	-	Method based on News Component Using Fuzzy based Scoring
F2	-	Method based on Fuzzy Scoring Without News Component
G1	-	Method based on News Component Using Graph based Scoring
G2	-	Method based on Graph Based Scoring Without News Component
GA	-	Genetic Algorithm
GATE	-	General Architecture for Text Engineering
H1	-	Human With Human Benchmark
IDF	-	Inverse Document Frequency
IE	-	Information Extraction
JAPE	-	Java Annotation Patterns Engine
LT	-	Length Type
NER	-	Named Entity Recognition
NLP	-	Natural Language Processing
NN	-	Neural Network
NP	-	Noun Phrase
ROUGE	-	Recall-Oriented Understudy for Gisting Evaluation
SVM	-	Support Vector Machine
TF	-	Term Frequency
V1	-	Method based on News Component Using Voting based Scoring
V2	-	Method based on Voting Without News Component

- VP - Verb Phrase
- WO - Word Overlap