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.

	Certified by :
SIGNATURE	**SIGNATURE OF SUPERVISOR**
811226-10-5729	Prof. Dr. Naomie Salim
(NEW IC NO. /PASSPORT NO.)	**NAME OF SUPERVISOR**
Date: March 3, 2014	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 what ever 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

LIST OF TABLES

LIST OF FIGURES

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