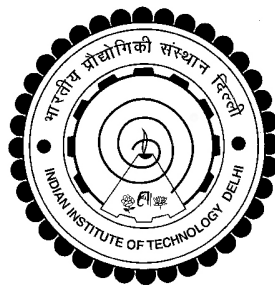


EXTRACTIVE TEXT SUMMARIZATION USING RANDOM INDEXING

PRAMOD KUMAR SAHOO



DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY DELHI
OCTOBER 2014

© Indian Institute of Technology Delhi (IITD), New Delhi, 2014

EXTRACTIVE TEXT SUMMARIZATION USING RANDOM INDEXING

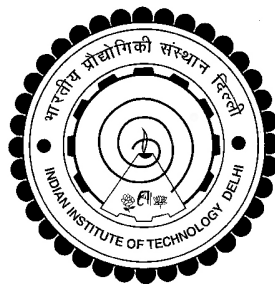
by

PRAMOD KUMAR SAHOO

Department of Mathematics

Submitted in fulfillment of the requirements of the degree of
DOCTOR OF PHILOSOPHY

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI
OCTOBER 2014

To my parents

Smt. Draupadi Sahoo

and

Sh. Krishna Chandra Sahoo

CERTIFICATE

This is to certify that the thesis entitled *Extractive Text Summarization using Random Indexing* submitted by *Mr. Pramod Kumar Sahoo* to the Indian Institute of Technology Delhi, for the award of the Degree of the Doctor of Philosophy, is a record of the original bona fide research work carried out by him under my supervision and guidance. The thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree or diploma.

Place:

Date:

Prof. Niladri Chatterjee

Supervisor

Department of Mathematics

Indian Institute of Technology Delhi

ACKNOWLEDGEMENTS

I am deeply indebted to *Mother Saraswati*; the Goddess of Knowledge. Due to *Her* divine grace I have been able to carry out this research work.

This thesis would not have been possible without the support of many people. It is impossible to acknowledge all of them here, but there are few people who deserve a special mention.

I am forever indebted to my Ph.D. supervisor Professor Niladri Chatterjee for teaching me basics of the research. I am grateful for his unfailing support throughout this work. He was great in letting me do what I wanted, and encouraged me when the things were not going fine. He has been extremely patient with my failings, and his trust in my abilities helped me overcome many difficulties.

I thank IIT Delhi for providing me necessary facilities to carry out my research work. I thank Head of Department of Mathematics, DRC Chairperson and my SRC members for their help at various stages of my research work. I am thankful to Sh. Amit Sharma, Research Scholar for his generous help during this research work.

I am thankful to Director, Institute for Systems Studies and Analysis (ISSA), Defence Research and Development Organisation (DRDO) for granting permission to join Ph.D. in IIT Delhi. I am especially thankful to Sh. Debasis Dutta and Smt. Deepakshi Shah for their constant support during my research work. Also, I am thankful to Sh. Avdhesh Kumar Gautam for his immense help in solving some software related issues.

Most importantly I would like to thank many people in my personal life for giving their unwavering support throughout, especially during this research work. No word can be ever be enough to thank my wife Smt. Lopamudra Sahoo for her wholehearted support for carrying out research work without bothering for the day-to-day problems at the family front. I am also very much thankful to my little daughter Ms. Snigdha Shravani for being kind to me during this research work. Finally, I thank my parents and in-laws for their blessings for successfully carrying out this research work.

Place:

Date:

Pramod Kumar Sahoo

ABSTRACT

Extractive text summarization aims at selecting the important sentences from the original text comprising single or multiple documents. Application of Random Indexing (RI) to extractive single-document text summarization has already been proposed in literature. This thesis studies the RI technique in detail, and proposes several improvements in the context of extractive text summarization. As a consequence, three improved versions of the RI-based text summarization algorithm, viz. RISUM, RISUM+ and MRISUM are obtained. Further, experiments have been conducted using both Cosine dissimilarity and Euclidean distance as proximity measures. A comparative study on their performance shows that while RISUM and RISUM+ perform almost at par, MRISUM outperforms both RISUM and RISUM+ significantly. MRISUM also outperforms computationally complex Latent Semantic Analysis (LSA) based summarization techniques.

This research work also focuses on studying in detail the near-orthogonality property of random index vectors, and its effect on extractive text summarization. A probabilistic definition of near-orthogonality of RI-based Word Space is presented, and a thorough discussion on the subject is conducted. Our experiments show that while quality of summaries produced by RISUM with Euclidean distance measure is almost invariant to near-orthogonality of the underlying Word Space; the quality of summaries produced by RISUM with Cosine dissimilarity measure is strongly affected by near-orthogonality. But it has been observed that unlike RISUM, the performance MRISUM follow no particular trend with respect to the near-orthogonality of the underlying Word Space.

RISUM and MRISUM algorithms developed for single-document generic summarization are modified appropriately for query-focused multi-document summarization task. The corresponding schemes have been termed as RISUM-QMD and MRISUM-QMD, respectively. While the performance of RISUM-QMD is not very encouraging, the performance of MRISUM-QMD is quite impressive. We feel that this scheme should be investigated in more detail as a comprehensive technique for multi-document text summarization.

Table of Contents

CERTIFICATE	i
ACKNOWLEDGEMENTS	iii
ABSTRACT.....	v
Table of Contents.....	vii
List of Figures	xiii
List of Tables	xv
Chapter 1 Introduction	1
1.1 Genres of Summary.....	2
1.2 Motivation	3
1.3 Objectives of the Thesis	5
1.4 Organisation of the Thesis.....	6
Chapter 2 Related Work.....	11
2.1 Graph Based Models for Text Summarization.....	16
2.1.1 Text Relationship Map Based Text Summarization	17
2.1.2 PageRank Based Text Summarization.....	18