ROUGH SET BASED TECHNIQUES FOR EXTRACTIVE TEXT SUMMARIZATION

NIDHIKA YADAV

DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY DELHI
NOVEMBER 2020

ROUGH SET BASED TECHNIQUES FOR EXTRACTIVE TEXT SUMMARIZATION

by

NIDHIKA YADAV

Department of Mathematics

Submitted
in fulfillment of the requirements of the degree of Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI
NOVEMBER 2020

CERTIFICATE

This is to certify that the thesis entitled ***Rough Set based techniques for Extractive Text Summarization*** submitted by ***Ms. Nidhika Yadav*** 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 her 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:

Professor Niladri Chatterjee

Department of Mathematics

Indian Institute of Technology Delhi

ACKNOWLEDGEMENTS

I am deeply thankful to Almighty for his immense and divine grace without which I would had been unable 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 consistent support in guiding me till PhD Thesis submission. He encouraged me in dealing with the research and personal hurdles that I faced during this time. He taught me research writing, research methodologies and most importantly he helped me out in dealing with failures in experiments that came several times during this duration.

I would like to 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 would like to thank all my family members for their unwandering support in carrying out my goal for completing the Thesis. I would like to thank them for their constant motivation and complete support in all aspects during this time which has been my true motivation. I would also like to thank all the people in my life who helped me in several aspects at various stages of Ph.D. I would like to thank to all my elders for their blessings.

I am thankful to the anonymous examiners and reviewers for their valuable comments and suggestions for improvising the papers and the thesis.

Nidhika Yadav

ABSTRACT

Rough Set (Pawlak, 1982) theory is a well-developed mathematical methodology to deal with uncertainty in data. The idea of Rough Set is to partition a given data of entities represented as attribute-value pairs into three classes: *positive region*, *negative region* and *boundary region*. The positive and negative regions separate the objects that belong to a specific class or outside it, while the boundary region consists of entities for which such a clear decision cannot be made with certainty.

Rough Set-based approaches have been used in different application areas including feature selection, data classification, knowledge acquisition, data mining, decision analysis among others. However, no significant work has been found for its application in the area of Extractive Text Summarization. Text Summarization, in general, condenses one or more collection of source text(s) into a single concise text that preserves the information content while reducing the size of the input text. In our approach we consider Extractive Text Summarization as a decision-making process, where corresponding to each input sentence the system takes a decision whether or not the sentence will be included in the summary. From this perspective we consider the process of sentence selection for generation of extractive summary as a process of decision making under uncertainty. The uncertainty comes from several aspects: the user-specified size of the summary, the relative importance of the sentences, presence of redundant information, the size and content of the document under consideration, among others.

We have proposed four different techniques for Rough Sets based Text Summarization. The schemes developed in the thesis are the following:

(1) Novel concepts of Rough Set based *span* and *spanning sets* are proposed for generating a smaller representative representation of the Universe under consideration. Properties and computation of spanning sets are described. Further,

the unsupervised concept of spanning set is applied to generate extractive summary with a required number of important sentences, typically specified by the user, from the text.

(2) Sentence to sentence similarity measures based on Fuzzy Rough Set are proposed and are further applied to unsupervised Single document Text Summarization. This new measure when applied to the problem of Text Summarization has been found to perform even better than the span-based approach with respect to certain sets of documents.

(3) Suitability of learning from existing data (e.g. Document Understanding Conference 2002[1]) using supervised Rough Set based techniques for Multi-document Text Summarization (MDTS) is analysed. Firstly, a Rough Set based ranking measure called *Rank-Measure* is proposed for post-processing the results obtained by supervised MDTS. The results show a significant improvement over results computed without using Rough Set based ranking measure. Further, a Neighborhood Rough Set based supervised learning algorithm is also proposed for MDTS and is found to improve the ROUGE scores further.

(4) Rough Sets have been popularly used for feature selection. Selection of important features for Text Summarization using Rough Set based techniques is analysed. Further, two new algorithms for Rough Set based feature selection are proposed and their effect on performance for MDTS is analysed.

(5) Finally, two more new techniques are developed during this research which are novel efforts in different areas to improve on the summarization results. The techniques are sentiment-based summarization and Random Indexing for MDTS. The suitability of these techniques is analysed for Text Summarization.

1. *URL: http://www-nlpir.nist.gov/projects/duc/guidelines/2001.html*

सार

रफ सेट (पावलक, 1982) डेटा में अनिश्चितता से निपटने के लिए एक अच्छी तरह से विकसित गणितीय पद्धति है। रफ सेट का विचार किसी दिए गए डेटा सेट को तीन वर्गों में विभाजित करना है: पॉजिटिव रीजन, नेगेटिव रीजन और बाउंड्री रीजन। पॉजिटिव रीजन और नेगेटिव रीजन उन डेटा पॉइंट्स को अलग करते हैं जो एक विशिष्ट वर्ग या उसके बाहर हैं, जबकि सीमा क्षेत्र में ऐसी संस्थाएं हैं जिनके लिए इस तरह का स्पष्ट निर्णय निश्चितता के साथ नहीं किया जा सकता है।

रफ सेट-आधारित दृष्टिकोणों का उपयोग विभिन्न एप्लिकेशन क्षेत्रों में किया गया है, जिसमें फीचर चयन, डेटा वर्गीकरण, ज्ञान अर्जन, डेटा माइनिंग, दूसरों के बीच निर्णय विश्लेषण शामिल हैं। हालांकि, एक्स्ट्रेक्टिव टेक्स्ट समराइजेशन के क्षेत्र में इसके आवेदन के लिए कोई महत्वपूर्ण काम नहीं मिला है। पाठ संक्षेपण, सामान्य रूप से, स्रोत पाठ के एक या अधिक संग्रह को एक संक्षिप्त पाठ में सम्मिलित करता है, जो इनपुट पाठ के आकार को कम करते हुए सूचना सामग्री को संरक्षित करता है। हमारे दृष्टिकोण में हम एक्स्ट्रेक्टिव टेक्स्ट समराइजेशन को निर्णय लेने की प्रक्रिया मानते हैं, जहाँ प्रत्येक इनपुट वाक्य के अनुसार सिस्टम निर्णय लेता है कि वाक्य को सारांश में शामिल किया जाएगा या नहीं। इस दृष्टिकोण से हम अर्क सारांश के निर्माण के लिए वाक्य चयन की प्रक्रिया को अनिश्चितता के तहत निर्णय लेने की प्रक्रिया मानते हैं। अनिश्चितता कई पहलुओं से आती है: सारांश का उपयोगकर्ता-निर्दिष्ट आकार, वाक्यों का सापेक्ष महत्व, निरर्थक जानकारी की उपस्थिति, विचाराधीन दस्तावेज़ का आकार और सामग्री, अन्य।

हमने रफ सेट्स आधारित टेक्स्ट सारांश के लिए चार अलग-अलग तकनीकों का प्रस्ताव दिया है। थीसिस में विकसित की गई योजनाएं निम्नलिखित हैं:

(१) रफ सेट आधारित स्पैन और स्पैनिंग सेट्स की नोवेल अवधारणाएं विचाराधीन यूनिवर्स के एक छोटे प्रतिनिधि प्रतिनिधित्व को बनाने के लिए प्रस्तावित हैं। स्पैनिंग सेट्स के गणितीय गुण और गणना का वर्णन किया गया है। इसके अलावा, स्पैनिंग सेट्स की अवधारणा, आमतौर पर उपयोगकर्ता द्वारा निर्दिष्ट महत्वपूर्ण वाक्यों की एक आवश्यक संख्या के साथ निष्कर्षण सारांश उत्पन्न करने के लिए लागू किया जाता है।

(2) फजी रफ़ सेट के आधार पर वाकया का वाकया से मेल का प्रस्ताव किया गया हैं और आगे इसे एकल दस्तावेज़ पाठ संक्षेप के लिए लागू किया गया है।

यह नया उपाय जब टेक्स्ट समराइजेशन की समस्या पर लागू किया गया है, तो दस्तावेजों के कुछ सेटों के संबंध में स्पैन-आधारित दृष्टिकोण से भी बेहतर प्रदर्शन करने के लिए पाया गया है।

(3) मौजूदा डेटा (जैसे डॉक्यूमेंट अंडरस्टैंडिंग कॉन्फ्रेंस 2002) से सीखने की उपयुक्तता, मल्टी-डॉक्यूमेंट टेक्स्ट सुमरीकरण (MDTS) के लिए पर्यवेक्षित रफ सेट आधारित तकनीकों का उपयोग करके विश्लेषण किया गया हैं । सबसे पहले, एक रफ सेट आधारित रैंकिंग माप, पर्यवेक्षण MDTS द्वारा प्राप्त परिणामों के बाद के प्रसंस्करण के लिए प्रस्तावित है। परिणाम रफ सेट आधारित रैंकिंग उपाय का उपयोग किए बिना गणना किए गए परिणामों पर एक महत्वपूर्ण सुधार दिखाते हैं। इसके अलावा, एक नेइबोर्हूद रफ सेट आधारित सुपरवाइज्ड लर्निंग अल्गोरिथम भी MDTS के लिए प्रस्तावित किया गया हैं और ROUGE स्कोर को और बहुत बेहतर पाया जाता है।

(4) फीचर चयन के लिए रफ सेट्स का लोकप्रिय रूप से उपयोग किया गया है। रफ सेट आधारित तकनीकों का उपयोग करके पाठ संक्षेपण के लिए महत्वपूर्ण विशेषताओं का चयन किया जाता है। इसके अलावा, रफ सेट आधारित सुविधा चयन के लिए दो नए एल्गोरिदम प्रस्तावित हैं और एमडीटीएस के प्रदर्शन पर उनके प्रभाव का विश्लेषण किया गया है।

Table of Contents

List of Figures

List of Tables