Available online at www.sciencedirect.com

ScienceDirect

IERI Procedia 10 (2014) 245 – 251

Procedia IERI

www.elsevier.com/locate/procedia

2014 International Conference on Future Information Engineering

A Redundancy Elimination Approach towards Summary Refinement

Esther Hannah. M [a], Saswati Mukherjee [b], Sakthi Balaramar [c] *

Associate Professor, St. Joseph's College of Engg, Chennai, India
Professor, DIST, Anna University, Chennai, India
Student, St. Joseph's College of Engg, Chennai, India

Abstract

A summary generated by a machine, in contrast to human-generated summaries are produced in less time, unbiased, not time or mood dependent and reliable. However many commonly used approaches are feature based methods that look out for important sentences or phrases by observing features or cues. Such feature based methods may end up producing summaries that contain sentences which are similar in meaning, mostly which depend on sentence scoring and hence not a desirable factor. The proposed work takes a machine generated summary as rough summary and uses binomial distribution to identify importance of every sentence in the rough summary. The semantic similarity between sentences is identified and the sentences are removed thereby refining the summary. By eliminating similar sentences the summary is refined so that only informative sentences are left in the summary. The proposed redundancy elimination approach is applied on summaries obtained from an existing summarization system with the fuzzy based summarization model as a case study. Evaluation of the summary refinement approach is done on DUC2002 dataset and the results are promising.

Keywords: Text Summarization; Summary Refinement; Redundancy Elimination; Binomial Distribution.

*Esther Hannah. M . Tel.: +919841627565;
E-mail address: author@institute.xxx .

doi:10.1016/j.ieri.2014.09.084

1. Introduction

The amount of data on the Internet increases every day, and therefore the task of selecting and classifying relevant information becomes all the more difficult. Text summarization systems can automate the task of generating a summary from a large text in a considerable amount of time. Traditionally researchers looked at designing statistical models for achieving this. More recently, attention has turned to a variety of machine learning algorithms that can build models automatically.

A summary consists of the main topics in one or more documents as a short and concise readable text. Summaries are generated from a single document [1] or multiple documents. The summary formed from more than one documents is called multi-document summarization. A 'query-biased' summarization [6] provides information to user on queries. Topic summarization deals with the generation of topics along with providing the most informative sentences. The present work focuses on summary refinement. Summarization makes the document more readable by making only the information [9] content provided to the user. Human generated summaries are expensive and machine generated summaries are not up to the mark. Several efforts are made by researchers in order to generate good, informative summaries.

The rest of this paper is organized as follows. Section 2 reviews the background and related work. Section 3 provides an overview of the proposed work summary refinement model. Section 4 discusses the conducted evaluation and results. Section 5 concludes the paper.

2. Background work

One of the very first works in automatic text summarization was done by Luhn et al in 1958, demonstrates research work done in IBM, focused on technical documents [11]. Luhn proposed that the 'frequency of word' proves to be a useful measure in determining the significance factor of sentences. Many approaches are already proposed on text summarization [4], based on the model they used the results vary. Some use to assign numeric weights to index terms based on the frequency of the term occurring in the document. The automatic extracting system [2] assign numerical weights to text sentence based on the weights assigned to certain machine-recognizable characteristics or clues. Some use to assign weights based on the semantic similarity measurement.

*E*xtracting key sentences from a document to form a summary can be done by measuring the relevance of sentences using fuzzy-rough sets [15]. A text summarization system that produces extractive summaries that utilize a well-defined set of features that represent the sentences in a text was proposed.

3. A Summary refinement model

System generated summaries are prone to contain similar sentences that convey similar meaning. Such similar sentences would have been the candidate sentences of the summary due to their importance in features. This leads to redundancy in summaries and thereby increases the length of the summaries. Some researchers have proposed to refine a system-generated summary using filtering sentences or phrases before they could become part of the summary. We have taken up this challenge and suggested a way of refining extractive summaries by removing redundant sentences.The proposed approach makes use of Binomial distribution for measuring Context Based Indexing [10]. By giving the weights to the topical terms, the sentence similarity weight can be assigned and a graph can help to eliminate redundant sentences to give a refined summary. The similarity values are used to construct a graph, showing the connection between sentences. The sentences with