International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-3 Issue-11 September-2013

Multi-Document Text Summarization using Mutual Reinforcement and Relevance Propagation Models Added with Query and Features Profile

Poonam P. Bari[1] , Shanta Sondur[2]

Abstract

Text summarization is the process of abridging the larger text into a shorter version preserving its information content and meaning. Using text summarizer user gets sense of the full-text, or able to know its information content without reading all sentences within the larger text. Text summarization reduces the text by removing less useful data which helps user to find the required information quickly without wasting time in reading the whole text. Lot of work has been done for automatic text summarization.

In this approach the technique based on the multi-document summarization using mutual reinforcement and relevance propagation models is modified by adding features profile to it. This enables us to group the document set into several topic themes and then these are clustered according to the query. Further proposed work identify the salient sentences from each cluster by applying feature profile and then these are ranked according to their weights of importance. Finally, the mutual reinforcement between the ranking of sentence set and ranking of the cluster set is found out using reinforcement after relevance propagation (RARP) algorithm.

Keywords

Feature profile, relevance propagation, mutual reinforcement, query focused, multi-document summarization.

1. Introduction

Today's world is full of information, that too online. The World Wide Web contains billions of documents and is growing at an exponential pace, it has become increasingly important to provide improved mechanisms to find and present textual information effectively.

Poonam P. Bari, M.E. Department of Information Technology, VESIT Chembur, Mumbai, India.

Shanta Sondur, Department of Information Technology VESIT Chembur, Mumbai, India.

Text summarization (TS) is the process to reduce (long) textual information to its most essential points; to distill the most important information from a source or sources to produce an abridged version of it (Endres-Niggemeyer, 1998; Mani and Maybury, 1999; Sparck-Jones, 1999). TS systems are designed to take a single article, a cluster of news articles, a broadcast news show, or an email thread as input, and produce a concise and fluent summary of the most important information.

Before discussing TS, first we should know what a summary is. Summary is a document produced from one or more documents, that tell important information from original text and it is shorter than it. Text summarization falls into two categories extractive and abstractive text summarization. Extractive summarization consists of selecting important sentences, paragraphs etc. from the original document and presenting them into shorter form. Extractive summaries are framed by extracting key text segments (sentences or passages) from the text. Abstractive summarization attempts to develop an understanding of the main concepts in a document and then express those concepts in clear natural language.

It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document.

This paper presents a new approach with query and features based multi-document summarization using mutual reinforcement and relevance propagation models. The main contribution of this work is addition of the features profile of sentences with existing query and RARP based summarization. And also increase the compatibility of system by accepting all types of text file formats of documents. The query based summarization can make available brief information conforming to given queries. RARP is used for getting perfect ranking to the sentences; thus users get the most important sentences in summary. This formed the motivation, if system thinks like human i.e. if we combine the features profile of sentences with query based and RARP. Xiaoyan cai et al. [1] demonstrated RARP with query based summarization had better

performance in comparison with only manifold ranking based multi-document summarization and basic like only query based summarization. This motivated us to combine features profile with RARP and query based summarization.

The rest of paper is organized as follows: Section 2 concisely reviews existing work. Section 3 presents the proposed system overview. The methodology of proposed system discussed in Section 4. Section 5 concludes the paper.

2. Related Work

Early work in summarization dealt with single document summarization where systems produced a summary of one document. Now a day everyone refers many documents to get more information in less time. Thus study of single document summarization evolved a new type of summarization i.e. multi-document summarization. It was motivated by use of the web. It produced a summary of bundle of documents.

A. Extractive Summarization

Different summarization techniques have been discussed in the literature; for extractive or for abstractive summarization. Extractive summarization assigns a significance score to each sentence and extracts the sentences as it is from the original text(s), with highest scores to form the summaries. The proposed system is based on extractive techniques. Bag-of-words model is built at sentence level, with the usual weighted term frequency and inverse sentence frequency paradigm [2], where sentence-frequency is the number of sentences in the document that contain that term.

These sentence vectors are then scored by similarity to the query and the highest scoring sentences are picked to be part of the summary. Tf-Idf techniques and clustering of documents is used together to increase the performance. Bundle of documents is related with many topics. They are typically fragmented up into sections. This organization applies even to summaries of documents. Multi-document summarization is also required clustering of documents related to the topics. The theme is represented by words with top ranking term frequency, inverse document frequency (TF-IDF) scores in that cluster [3]. Once themes have been known, a representative passage in each theme is selected and included in the summary. Later Dragomir R. Radev et al. [4] developed a multi-document summarizer, MEAD, which generates summaries using cluster centroid. It summarizes clusters of news articles automatically grouped by a topic detection system. MEAD uses Term Frequency-Inverse Document Frequency (TF-IDF) to calculate the weight for the word / term is used to select salient sentences. A. P. Siva Kumar et al. [5] have discussed query based summarization. The use of TS allows a user to get a sense of the content of full-text, or to know its information content without reading all sentences within the full-text. The sentences containing the query phrases are given higher scores than the ones containing single query words. Then, the sentences with highest scores are incorporated into the output summary. Portions of text may be extracted from different sections or subsections. The resulting summary is the union of such extracts. The number of extracted sentences depends on the compression rate given by users. A. Kogilavani et al. [6] have considered the feature profile for sentences. Feature profile is generated by considering word weight, sentence position, sentence length, sentence centrality, proper noun in the sentence and numerical data in sentence.

Xiaoyan Cai et al. [1] have discussed reinforcement after relevance propagation (RARP) i.e. manifold ranking based relevance propagation with mutual reinforcement between sentences and clusters. They ranked a sentence higher if it is contained in the theme cluster which is more relevant to the given query while a theme cluster ranked higher if it contains many sentences which are more relevant to the given query. Different from the traditional query-focused summarization approaches, which were either the simple extensions of generic summarizers and did not uniformly fuse the information in the query and the documents, or based on semi-supervised learning methods and/or supervised learning methods, Wang et al. [7] proposed a manifold-ranking-based approach to make uniform use of sentence-to-sentence and sentence-to-query relationships.

B. Manifold Ranking for Multi-Document Summarization

Xiaojun Wang et al. [7] proposed the manifold-ranking process to compute the manifold-raking score for each sentence that denotes the information richness of the sentence, and then uses the greedy algorithm to penalize the sentences highly overlapping with other informative sentences. The summary is produced by choosing the sentences with highest overall scores, which are deemed both informative and novel, and highly biased to the given topic. Manifold ranking is a semi-supervised learning

that explores the relationship among all the data points in the feature space [7], [8]. It has two versions regarding the different tasks: i) to rank the data points, or ii) to predict the labels of the unlabeled data points. For the task of ranking, the prior assumptions of it include a) nearby points are likely to have the same ranking scores; and b) points on the same structure (typically referred to as a cluster or a manifold) are likely to have the same ranking scores.

C. Mutual Reinforcement for Multi-Document Summarization

Zha [9] proposed a mutual reinforcement principle that was capable of extracting significant sentences and key phrases at the same time. In his work, a weighted bipartite document graph was constructed by linking together the sentences in a document and the terms appearing in those sentences. The mutual reinforcement was reduced to a solution for the singular vectors of the transition matrix of the bipartite graph. The relevance of each text unit to the given query was calculated by the cosine similarity and characterized by the corresponding text vertex in a three-layer text graph.

3. System Overview

Multi-document query based and feature specific summarization system as shown in Figure 1 is aimed at generating a summary document from multiple documents of similar theme. MDQFS provides functionalities such as uploading documents, grouping documents according to query, ranking the sentences with score, and generating summary.

In proposed system the performance of RARP and query based summarization is enhanced by adding feature profile. Scoring of the sentences is done using word weight, sentence position, sentence length, sentence centrality, proper noun in the sentence and cue phrase in sentence. The MDQFS allows users to add documents of all types of text format. Documents can be plain text, HTML files, .pdf files, .rtf files, Microsoft document (.doc) files etc.

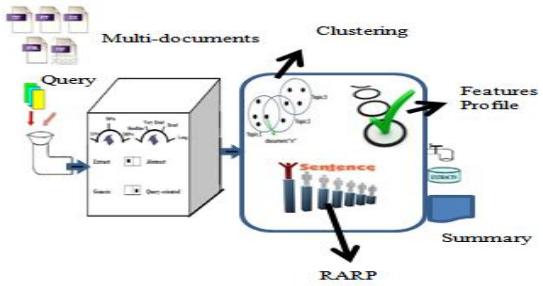

Figure 1: The proposed system

4. Methodology

The proposed system includes three modules shown in Figure 2 preprocessing module, sentence score calculation modules and sentence Ranking and sentence extraction module.

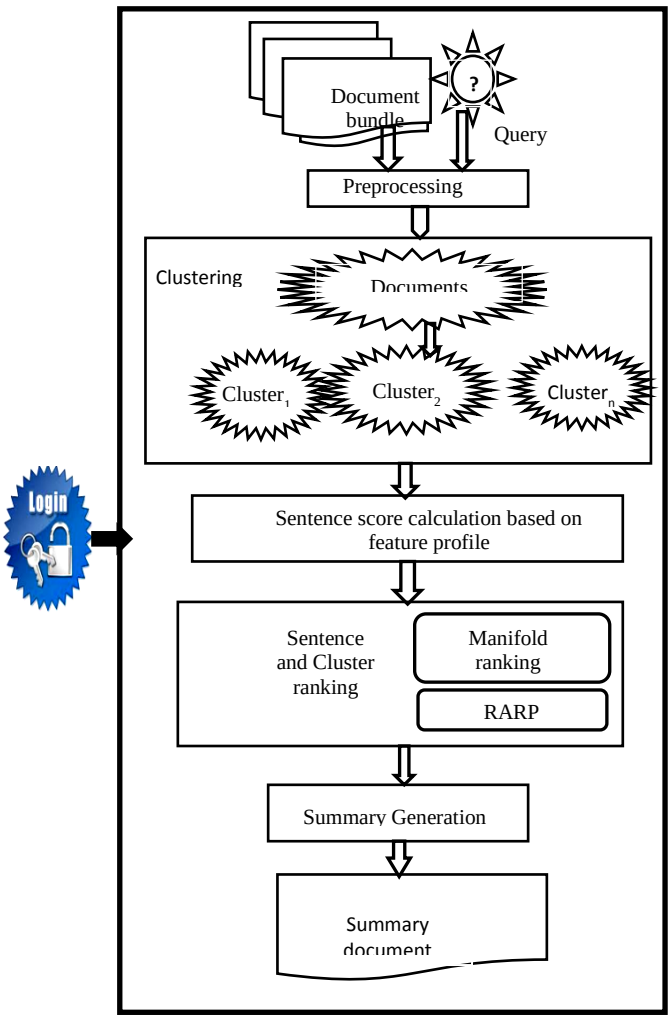

Figure 2: Working of proposed system

A. Pre-processing

The system takes all types of text documents i.e. .txt, .pdf, .rtf, .doc, .html etc. and query as input. Firstly it converts all documents in .txt files. Then it tokenizes the text documents in order to find the individual terms. Then filtering of the text is done by removing the stop words and remaining words are stemmed using Porter Stemmer algorithm.

The term weight is calculated as follows,
Term Weight = tf * idf
where, tf – term frequency
idf - inverse document frequency
After this the documents are grouped according to entered query. After grouping the documents, the next job is of scoring of sentences based on feature profile [6].

B. Sentence Score Calculation Based On Feature Profile

Feature profile is generated to capture the values of sentence-specific features of all sentences. The proposed work proceeds with features like term feature, sentence position, sentence length, sentence centrality, number of proper nouns in the sentence and number of numerical data in the sentence to generate feature profile.

Term Feature
Term Feature (TF) is defined as

$$TF=\sum Term\ Weight\ (term) * f(t, si)$$

where, f (t, si,) is the frequency of each term t in sentence si.

Position Feature
Always the first sentence of the document is most important. The position feature is defined by considering maximum positions of 3. For example, the first sentence in a document has a score value of 3/3, the second sentence has a score 2/3 and third sentence has a score value of 1/3. Position Feature (PF) defined as

$$PF\ (si) = [(M + 1) - Position(si)]/M$$

where, M – maximum positions or number of sentences in document.

Sentence Centrality (similarity with other sentences) Feature
Sentence centrality is the vocabulary overlap between this sentence and other sentences in the document. It is calculated as follows:

$$SCF(si) = \frac{\text{terms in si} \cap \text{terms in other sentences}}{\text{terms in si} \cup \text{terms in other sentences}}$$

Sentence with Proper Noun Feature
In general the sentence that contains more proper nouns is an important one and it is most probably included in the document summary. The following formula is used to calculate the inclusion of proper nouns (PNF) in the sentence.

$$PNF = \frac{\#(\text{proper nouns count in si})}{\text{Length(si)}}$$

Cue Phrase Feature
Sentences containing any cue phrase (e.g. "in conclusion", "this letter", "this report", "summary", "argue", "purpose", "develop", "attempt" etc.) are most likely to be in summaries.

Length - Feature
Too long or too short sentence is not suitable as the candidate for summary. The Length-Feature (LF) is defined as:
LF = (L(si)* total number of sentences)/L(dk)
where, L (si) - the length of sentence si.
L (dk) - be average sentence length of
document d.

C. Sentence Ranking by Reinforcement after relevance propagation (RARP) and sentence extraction

The proposed summarization model consists of relevance propagation and mutual reinforcement. Relevance propagation is achieved by using manifold ranking based algorithm [7]. A positive rank score is given to query point and zero to remaining sentences. Zha [9] has given mutual reinforcement principle in two sets of object which is used in this approach. In this approach the two sets are, one is sentence set and other is cluster set. A sentence should be ranked higher if it is contained in the theme cluster which is more relevant to the given query while a theme cluster should be ranked higher if it contains many sentences which are more relevant to the given query. After ranking of sentences the MDQFS selects the sentences using compression rate of user's choice.

5. Conclusions

This paper proposed Query and Features based Multi-Document Summarization using Mutual Reinforcement and Relevance Propagation models to enhance the performance of the existing query based multi-document summarization using RARP algorithm.

The proposed MDQFS accepts all types of file formats like: .pdf, .doc, .rtf, .html, etc whereas the existing text summarizer can handle only .html and .txt format of files.

References

[1] XiaoyanCai and Wenjie Li "Mutually Reinforced Manifold-Ranking Based Relevance propagation Model for Query-Focused Multi-Document Summarization" IEEE Transactions on audio, speech, and language processing, vol. 20, no. 5, july 2012.

[2] Rene Arnulfo Garcia-Herandez and Yulia Ledeneva,"Word Sequence Models for Single Text Summarization",IEEE,44-48, 2009.

[3] Yongzheng, Nur and Evangelos , "Narrative Text Classification for Automatic Key Phrase Extraction in Web Document Corpora", WIDM'5, Bremen Germany, 51-57, 2005.

[4] Dragomir R. Radev , Hongyan Jing and Molgorzata Stys,"Centroid-based summarization of multiple documents", International Journal of Information Processing and Management, 2004.

[5] A. P. Siva Kumar, Dr. P. Premchand and Dr. A. Govardhan "Query–based summarizer based on similarity of sentences and word frequency", International journal of Data Mining & Knowledge Management Process (IJDKP) Vol. 1No.3, May 2011.

[6] A. Kogilavani and Dr. P.Balasubramani, "Clustering and feature specific sentence extraction based summarization of multiple documents", International journal of computer science and information technology (IJCSIT) vol2. No.4 August 2010.

[7] X. J.Wang, J. W. Yang, and J. G. Xiao, "Manifold-ranking based topic focused multi-document summarization," in Proc. 18th IJCAI Conf., pp. 2903–2908, 2007.

[8] D. Y. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," in Proc. 17th NIPS Conf., pp. 321–328 , 2003.

[9] H. Y. Zha, "Generic summarization and key phrase extraction using mutual reinforcement principle and sentence clustering," in Proc. 25th SIGIR Conf., pp. 113–120 , 2002.

Prof. Shanta Sondur is currently working as Professor in Department of Information Technology, VESIT Chembur, Mumbai, India. Her research interests are Document summarization, Soft Computing, Optimization, Process Control, etc.

Poonam P. Bari pursuing Master in Engg. from Department of Information Technology, VESIT Chembur, Mumbai, India. Her research interests are Document summarization, Data Mining, Data Structures, Database Management Systems, etc.