A Graph Based Query Focused Multi-Document Summarization

J Balaji, Department of Computer Science and Engineering, Anna University, Chennai, India
T V Geetha, Department of Computer Science and Engineering, Anna University, Chennai, India
Ranjani Parthasarathi, Department of Information Science and Technology, Anna University, Chennai, India

ABSTRACT

An user’s information need, normally represented as a search query, can be satisfied by creating a query focused coherent and readable summary, by fusing the relevant parts of information from multiple documents. While aggregating the information from multiple documents, the quality of the summary is improved by eliminating redundant information from the document set. In this paper, we focus on removing such redundant information and identifying the essential components from multiple documents (represented as a single global semantic graph), with respect to the given query (represented as a query graph). While the redundancy elimination is carried out using various levels of graph matching which are then indicated through canonical labeling of graphs, the selection of essential components for a query focused summary is performed, through the modified spreading activation theory, where the query graph is also integrated during the spreading activation over the global graph. The proposed system shows significant improvements in generating summaries when compared to other existing summarization systems.

Keywords: Multi-Document Summarization, Query Focused Summary, Redundancy Elimination, Semantic Graphs, Spreading Activation, Universal Networking Language

INTRODUCTION

With a wide variety of documents available in the web, text summarization is one of the important tasks, which effectively compresses the information in a document(s). Multi-document summarization is a task of identifying the important common themes and/or aspects of multiple documents.

The primary tasks in multi-document summarization are the identification of similarities and differences between documents (Wan & Yang, 2008). One of the challenges of multi-document summarization is that a set of documents might contain diverse information, which is either related or unrelated to the particular topic. Therefore, effective methods are needed to analyze the information stored in different documents, and abstract the globally important information to reflect the main topic. In single-document summarization, the sentences in a document are unique and may not have redundant information.

DOI: 10.4018/ijiit.2014010102
In the case of multi-document summarization, an important challenge is that, the information stored in different documents inevitably overlaps, and hence, we need effective methods to merge and reproduce information with minimum redundancy. Moreover, multiple documents from a News corpus can convey the same event with different sentence structures, without conveying additional information. While summarizing the text from such multiple documents, the redundant information needs to be eliminated to produce a generic distinctive summary. Redundancy of information is especially an issue, when we use ranked documents obtained from a search engine, using a search query to produce a summary relevant to the given query.

In general, highly salient sentences are extracted from the document set, based on syntactic and/or statistical features (Erkan & Radev, 2004), whether the approach to summarization is rule based or machine learning based. Machine learning approaches to summarization can be categorized into unsupervised and supervised. The unsupervised method is mainly based on scoring sentences in the documents by combining a set of predefined features (Mani & Bloedorn 1998; Conroy 2006). Contrastingly, in the supervised method, summarization is treated as a classification or a sequence labeling problem and the task is formalized as identifying whether a sentence should be included in the summary or not (Shen et al., 2007). However, the method requires a large set of training examples which are not available for resource-constrained languages.

Most existing summarization approaches focus on an extractive summary, where the important sentences are extracted using the salient information from different documents. They use a shallow analysis, without paying much attention to the rich semantic features associated with words, and semantic relations expressed within and across sentences as well as documents. Yet another problem is that different users have different information needs. Thus, an ideal multi-document summarization should provide different levels of details for a specific topic, with respect to the user’s interest (Chali & Joty, 2008). This can be achieved by exploiting the lexical, syntactic, semantic, pragmatic and discourse information of multiple documents. Multi-document summarization is also carried out using graph-based approaches, such as LexRank (Erkan & Radev, 2004). LexRank has been applied a random walk in a fully connected undirected graph, to redistribute the node weights where text units (i.e. sentences) are represented as nodes, and similarities between text units are represented as edges. These approaches result in the presence of redundant information in the summaries.

One promising approach to multi-document summarization is to outline the overall structure of a set of related documents to give users an overview of a specific topic, and subsequently allow them to zoom into different areas according to their interest. This is performed by semantically analyzing multiple documents rather than a single document (Balaji et al., 2013), using the semantic relations between concepts in the sentences, between sentences, and so on. Our primary aim is to eliminate the redundant information from the semantic graphs of multiple documents and to obtain a compressed form of summary semantic graph relevant to a given query without losing any important information.

In this work, we focus on two aspects of multi-document summarization – redundancy elimination and query focused summary generation. The redundancy elimination is carried out using graph matching procedure, which is based on the structural and semantic properties of the semantic graph, representing the set of documents to be summarized. While focusing on a query oriented summary from multiple documents, relevance checking needs to be performed, using various similarities scoring measures, such as concept similarity (synonym, hyponym and hypernym concepts) and graph similarity (concept-Relation-concept (CRC) and semantic relation), between the query and the sentences, to narrow down the process of