ABSTRACT

The exponential growth of World Wide Web (WWW) has made it the prime database in the digital era for all information needs. The enormous growth, diverse, dynamic, semi-structured and unstructured nature of web makes internet exceptionally difficult in searching and retrieving relevant information.

Search engines are helpful to some extent in locating the web resources. For each query given by user, search engines return thousands of URLs as search result which normally includes many irrelevant, less relevant and redundant results. This requires the user to further browse many web documents to retrieve just a few lines of content of significance to the user.

Today, there is a thrust for quick and immediate identification and retrieval of required information as expected by the users. Automatic summarizers help the users to get the gist of the web page in few seconds without going through the entire content by applying web content mining techniques.

The query based summarizers process the entire document content to identify and extract query relevant pieces of information at run time. This challenge to the processing capacity of summarizers can be reduced by limiting the size of the textual unit that need to be processed.
The need for quick decision making to handle information from business transactions, scientific data and text reports are increasing today. Deployment of decision making capability in web databases gives rise to web content mining.

The aforementioned problem has given rise to the development of web content mining techniques to provide a comparative summary of a set of web pages selected by user. This thesis focuses on this particular issue and provides a solution for faster content analysis by generating comparative summary of a set of web pages selected by the user, which would help in quick decision making and trend analysis.

This research work addresses the issue of run time processing overhead posed to the query biased summarizers through mathematical offline pre-processing techniques to segment the web documents in server’s repository into related segments. This thesis proposes two different methodologies to perform segmentation utilizing cosine similarity, and frequent terms set and semantic relevance based measures to identify related contents.

During query time, instead of the entire document, only a few query related segments of the selected document are identified and processed further to generate the short extract. Comparative summary of a set of web documents is composed by aggregating the individual summaries of documents in the set. This dissertation also analyses the impact of various content scoring parameters considered for measuring the significance of a
sentence in view of the given query, and also proposes the use of semantic relevance parameter to improve the quality of summary.

Experimentation of the proposed methodologies has been conducted using WebKB dataset and a custom real time dataset. Results obtained after segmentation show that the frequent terms set and semantic relevance based approach performs better than cosine similarity based approach in view of processing time and storage requirements. Methodology for summarization using pre-processed segments has been evaluated in two aspects considering the impact of usage of pre-processed segments and also taking into account the quality and information content of the generated summary. Co-selection based measures like precision, recall and F-measure, content based ROUGE scores, User Satisfaction Index (USI), response time and processing efficiency have been used to evaluate the proposed summarization method. Comparative summary has been analyzed with respect to the extrinsic evaluation parameter. Experimental results establish that the proposed methods give better results in performance measure than the existing methods.