CHAPTER 3

PROBLEM STATEMENT AND RESEARCH METHODOLOGY

This research has been carried out to explore the mathematical procedures for generating comparative summary of a set of selected URLs from search results using statistical and linguistic measures. Comparative summary has been generated by aggregating the summary of individual web pages. Efficiency of the summary generation algorithm is improved by reducing the size of the textual units to be processed at query time. The performances of the algorithms against various test cases are analyzed. This chapter gives an overview of the problem definition, objectives and research methodology adopted in this research.

3.1 PROBLEM DEFINITION

Locating and retrieving the required information contents in the huge World Wide Web has become a tedious task today. The result set provided by search engines with millions of URLs for each query are helpful only to some extent. Lots of time and effort by the user is required to further read the content of these URLs to correctly locate the expected information.

Web page summarization techniques focus on providing a small digest representing the actual contents of the web page. These short extracts give a clue and guide the user in further reading the entire content and also help the user to understand the overall content of the web document without
reading the entire content. Query based summaries add user interest components to these generic summaries.

Query related summary is generated at run time according to the given query string. Entire document contents are to be processed in view of the query to identify the related contents. Earlier researches have considered the entire web page to generate the summary dynamically. Eventually, this process poses a challenge to the processing capacity of information servers that support millions of web users. One of the prime issues is to reduce the textual unit size that need to be considered for summarization at query time to reduce the time complexity.

Query sensitive significant sentence are extracted to compose the summary. Statistical, positional, HTML tag based measures and query term based scores have been used in different compositions in summarization systems to identify important sentences from the content. These measures considered the surface features, presentation features and similarity based features to score the significance of a sentence in view of the query. This would leave dissimilar but relevant contents unnoticed. For example, ‘academics’ and ‘education’ are related but dissimilar terms. Linguistic relevance between the query and the contents would be able to capture these dissimilar but relevant contents.

In this work, a mathematical approach is devised based on statistical measures like term frequency, term weight and linguistic similarity measure using WordNet semantic distance between terms for segmenting the web page. The focus of segmentation method is to increase the intra-segment similarity and to decrease the inter-segment similarity. Query related segments are identified and processed further to obtain the short extract at run time. Text surface measures like position, length, query overlap, average
distance between query terms and also linguistic measures are considered to measure the significance of a sentence. This thesis also analyzes the impact of usage of pre-processed segments for summary construction and the impact of different set of sentence scoring parameters during summarization.

3.2 OBJECTIVES

In this thesis, performance of query sensitive summarization system has been improved using statistical and linguistic techniques based web content mining methodologies. The objectives of this research constitute the following:

3.2.1 Main Objectives

- To generate query related comparative summary of set of URLs selected from search result, using improved summary generation techniques and pre-processed segments.

- To reduce the run time overhead of query based summarizers which in turn will improve response time of summarizer, by reducing the size of textual units considered for summary generation.

- To segment the web documents using statistical and linguistic parameters to facilitate for reducing textual unit size considered for summarization by selecting query relevant segments alone.

- To improve the quality of generated summaries using statistical, visual and linguistic parameters for measuring the significance of the sentences.
3.2.2 Specific Objectives

In order to achieve the above mentioned main objectives, the following specific objectives have been formulated:

- To design a framework to generate a comparative summary of contents of a set of URLs selected from search result.

- To develop an architectural design and an algorithm to facilitate segmentation of web documents in server’s repository.

- To compare and analyze the performance of proposed textual unit similarity measure for segmentation process with existing text similarity measures.

- To devise a methodology to improve the performance as well as response time of summarization system by reducing query time processing overhead using pre-processed segments.

- To formulate an algorithm to identify query relevant segments of the selected URLs and extract significant sentences from these segments considering sentence surface measures, presentation parameters and semantic relevance based measures.

- To study the impact of semantic relevance parameter in measuring the significance of contents in view of the given query.

- To study the impact of usage of pre-processed segments on summary construction process in view of quality of the summary, response time and also processing efficiency of the automated system.
3.3 RESEARCH METHODOLOGY

The research commenced with an exhaustive survey on various segmentation and summarization approaches. The various parameters considered for sentence scoring and sentence similarity measures were identified and analyzed. Space complexity and time complexity of these methodologies were studied and their shortcomings were identified.

Both algorithmic and experimental researches were carried out. As part of the algorithmic approach, algorithms for segmenting the web pages and summary generation using these segments were devised. Following an experimental research methodology, the devised algorithms were implemented and tested with real time datasets (WebKB dataset, real time web documents collection related to “Banking services”, “Computational algorithms”, “Environmental science”, “Science and technology”, “Communication technology”, “Air pollution”, “Image processing techniques”, “Infrastructure in academic institutions”, “Insurance policy”, etc.) containing web pages related to various domains collected through search engines.

A thorough analysis has been made by carrying out several experiments with different sets of test cases and sentence scoring parameters. To substantiate the results, statistical analysis using ROUGE Evaluation Tool Kit was also conducted, subsequent to the experimentation.

This chapter has brought out the problem definition that is being attempted in this research. The objectives of the research have been highlighted along with the research methodology. The general framework and basic building blocks of the proposed approaches for query sensitive comparative summarization system is presented in the next chapter.