CHAPTER 9

CONCLUSIONS AND FUTURE ENHANCEMENTS

9.1 CONCLUSIONS

This research work focused on the generation of query sensitive comparative summary of a set of URLs selected from search result to aid the web users in content analysis and quick decision making. Query related extractive summarization techniques generate a short summary of the web document by excerpting a short span of text that is closer to user query dynamically. This research addressed the issue posed to query based summarizer, to process the entire document during query time that warranted for huge processing capacity, by eliminating irrelevant portions of the document from being processed during summary generation. This reduction in textual unit size has been achieved using the proposed web document segmentation approaches.

Research in the above mentioned aspects has led to new ideas and innovations. The following are the research outcomes:

- Design and implementation of web document segmentation through two different approaches namely, cosine similarity based method and frequent terms set and semantic relevance based methods.

- Analysis of the proposed segmentation approaches by means of processing time and space complexity.
• Design and implementation of query based web document summarization system using pre-processed segments to reduce the run time overhead of summarizers.

• Analysis of the proposed summarization approach using both intrinsic as well as extrinsic summary evaluation approaches, and also the impact of usage of pre-processed segments in summary construction.

• Performance analysis on the impact of various sentence scoring parameters considered to measure the significance of contents in respect of user query and comparison with existing summarization systems.

• Design and implementation of comparative summary generation system.

• Performance evaluation using extrinsic summary evaluation method.

From the results obtained, it is shown that the processing time and space complexity of segmentation using Frequent Terms Set and Relevance based measures is 34.55% and 32.95% less than that of cosine similarity based segmentation. Usage of pre-processed segments during summarization reduces run time overhead of the summarizer and help to improve the processing efficiency by 33.5% which in turn improves the response time by 67.08% since only few query related segments have been processed to obtain the short extract. Average USI score for the summarizer reaches a maximum of 7.4 in 10.0 point scale when the relevance score to measure sentence significance is given higher priority. The precision, recall and F-measure scores of proposed system is 27.22%, 19.17% and 23.43% greater than that of MEAD-Lead based and 31.55%, 21.95% and 27.8% than the scores of
MEAD-Random systems and also the quality of the summary is not affected due to the proposed methodology of using only few query relevant segments for summary extraction.

Extrinsic evaluation results of the comparative summarization shows that the acceptability and usefulness of summary generated is in most promising level (90.51%) and the comparative summary is multipurpose and productive for web users for content analysis and faster decision making.

Both intrinsic and extrinsic summary evaluations conducted on various domain specific datasets prove that the performances of the proposed algorithms are consistent across different domains.

9.2 FUTURE ENHANCEMENTS

This research has brought out an innovative idea of utilizing pre-processed segments to reduce the processing overhead of summarizers using web content mining techniques. This research could further be extended in the following aspects:

- The proposed algorithms may be extended to support various types of web documents like .pdf, .doc, etc.
- Partial or full materialization of frequently accessed segments would also be considered to improve the response time of summarizers further.
- Associated terms can be considered for scoring the sentences in accordance with the given query.
**APPENDIX 1**

**COMPARISON AMONG WEB PAGE SEGMENTATION APPROACHES AND COMPARISON AMONG SIMILARITY MEASURES**

Comparison among Web Page Segmentation Approaches

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basic Approach</strong></td>
<td>DOM nodes are annotated as relevant or not relevant and also geometric parameters like nearby positions and compatibility with neighbouring nodes are considered.</td>
<td>Simulates human understanding of the web page layout structure. Used Degree of Coherence (DoC) measure to merge adjacent contents. Visual information is obtained from Web browsers.</td>
<td>From bottom of the DOM tree, nodes are coalesced to obtain segments of desirable size.</td>
<td>Layout and UI components are recursively segmented.</td>
</tr>
<tr>
<td><strong>Drawbacks</strong></td>
<td>Annotating DOM tree at query time is computationally intensive. Entire document need to be processed at run time.</td>
<td>DoCs across adjacent sentences are considered but semantic similarity among contents from various part of the document is not considered. Highly computationally intensive. System need to be trained in advance.</td>
<td>Content related to particular query term may occur in different parts of the document. (for example research activities carried out at different departments in the organization or institution)</td>
<td>Informative sites do not normally contain more UI components. Textual contents and their relevance are not considered.</td>
</tr>
<tr>
<td><strong>Merits</strong></td>
<td>Since DOM tree is used, time required for document graph construction is eliminated.</td>
<td>Visual cues are used which would simulate human understanding.</td>
<td>Computationally faster compared to other approaches.</td>
<td>Performs better for UI intensive web pages.</td>
</tr>
</tbody>
</table>
## Comparison among the similarity measures

<table>
<thead>
<tr>
<th>Basic Approach</th>
<th>Document vectors contain the terms frequencies of both documents</th>
<th>A binary vector contains 0s and 1s to represent the presence and absence of a word in the document</th>
<th>HITS of co-occurrence of two words from two documents are considered</th>
<th>Information drawn from semantic networks (WordNet)</th>
<th>n-gram which is a subsequence of n words from a given sentence</th>
<th>Context of words vectors, terms weight and also the semantic similarity are considered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Representation of documents</td>
<td>Documents are represented as vectors having same length</td>
<td>Documents are represented as vectors of binary attributes</td>
<td>Word-by-word the whole document is considered</td>
<td>Each word of the document is considered individually.</td>
<td>Each sentence of the document is split into n-grams</td>
<td>Vector contains the terms and term frequencies</td>
</tr>
<tr>
<td>Similarity Measurement</td>
<td><em>Dot product</em> of two vectors of attributes of A and B</td>
<td>Size of the intersection divided by the size of the union of the sets</td>
<td>Word co-occurrence using counts collected over very large corpora (WWW) using <em>HITS count</em></td>
<td>Quantify the degree to which two words are semantically related using <em>length of the shortest path</em> between two concepts using node-counting, and maximum <em>depth of the taxonomy</em>.</td>
<td>Similarity is computed using the number of co-occurrence of n-grams</td>
<td>Context is used to identify similar terms which could capture semantically relevant terms.</td>
</tr>
<tr>
<td>Drawbacks</td>
<td>Sparseness and high dimensionality, relevant but dissimilar words will not be considered</td>
<td>Sparseness and high dimensionality, relevant but dissimilar words will not be considered</td>
<td>Computational complexity is high since HITS count to be identified for each pair of words Term weight is not considered</td>
<td>Does not consider the weight of the term with in document and larger documents would get high score.</td>
<td>Computationally intensive since even for 2-5 grams all 2-5 gram word sequences need to be considered</td>
<td>System need to be trained with different word contexts. Study and analysis of word contexts are computationally intensive</td>
</tr>
</tbody>
</table>

### Corpus-based Measures
- (Rada Mihalcea et al 2006, Li et al 2006)

### Knowledge-based Measures
- (Xiao-Ying Liu et al 2008, Nitish Aggarwal et al 2011)

### n-gram method
- (Islam and Inkpen 2008)

### Latent Semantic Analysis (LSA)
- (Landauer et al 1998)

### Frequent Terms set and their semantic relevance based measure (Proposed metric)