CHAPTER 8

RESULTS AND DISCUSSION

Experimental results of the proposed algorithms on WebKB dataset and the custom real time dataset have been presented and analyzed in the earlier chapters. As the performances of the algorithms may vary with the different test cases used in experiments, the impact of the proposed methodologies for segmentation and summarization is analyzed further, with domain specific datasets.

This thesis analyses the performance of the system on three domain specific datasets extracted and created from real time web sites which shows that the system is domain independent and is applicable to any web document.

8.1 DATASETS COLLECTION AND CREATION FOR CONDUCT OF EXPERIMENTS

Web documents (Meng Wang et al 2009) related to ‘Science and Technology’, ‘Banking services’ and ‘Literacy’ have been extracted through Google search results and are considered as datasets D1, D2 and D3 respectively. This thesis intends to generate text summaries and hence, text intensive top 350 documents from each genre have been identified and extracted using automatic web page extraction tools.
These input collections were processed to remove non-textual contents like image, audio, video, etc. prior to experimentation. As a resultant of this step, the average sizes of documents in these collections are 24.5KB, 18.0KB and 30.5KB respectively as only textual contents are considered for further processing.

Subsequently, the proposed segmentation approaches using cosine similarity and the proposed Frequent Terms set and Semantic Relevance (FTSR) based measures have been applied on them to obtain the segments of these documents in offline.

Query based summaries were generated for 100 different queries pertaining to each domain in the collection followed by comparative summary generation for the same set of queries.

The segmentation approaches, summarization using pre-processed segments as well as comparative summarization approaches have been evaluated and the experimental results are analyzed by means of various measures in the following subsection.

8.2 PERFORMANCE ANALYSIS AND DISCUSSION

8.2.1 Performance Evaluation of the Segmentation Approaches

The proposed segmentation methodologies have been evaluated through the processing time and the space complexity. Average processing times taken by these approaches for the documents of these three domains are given in Figure 8.1.
Figure 8.1 shows that the segmentation using frequent terms and relevance based measure performs 29.93% (average) faster than the cosine similarity based approach with respect to datasets D1, D2 and D3 as the vector construction takes more time in the earlier approach.

Space complexity of the cosine similarity based approach is greater than that of the FTSR based approach as the vector model of topic blocks in cosine similarity based method contains frequency of union of terms present in both topic blocks under consideration whereas, the later method constructs the vector model of each topic block independently of other topic blocks. Table 8.1 shows the space complexity in terms of the dimensions of the constructed vectors in average to the documents present in these datasets.

From Table 8.1, we can infer that the space complexity of cosine similarity based approach is 32.95% above that of the second approach with respect to the three datasets and is outperforming in view of both processing time and space requirements.
Table 8.1  Space complexity of the proposed segmentation approaches (in average)

<table>
<thead>
<tr>
<th>Domain</th>
<th>No. of nodes</th>
<th>Number of topic blocks identified</th>
<th>Cosine similarity based method</th>
<th>FTSR based method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Average no. of union of terms in any pair of topic blocks</td>
<td>Space Complexity (Dimensions of vector)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Average no. of terms in topic blocks</td>
</tr>
<tr>
<td>Science and Technology (D1)</td>
<td>85</td>
<td>29</td>
<td>24</td>
<td>1392</td>
</tr>
<tr>
<td>Banking Services(D2)</td>
<td>76</td>
<td>24</td>
<td>20</td>
<td>960</td>
</tr>
<tr>
<td>Literacy(D3)</td>
<td>115</td>
<td>35</td>
<td>31</td>
<td>2170</td>
</tr>
</tbody>
</table>

8.2.2 Performance Evaluation of the Summarization Approach

The quality of the summary generated by the system for 100 random queries posed against each dataset is analyzed based on co-selection and content based measures. For these queries the reference summaries have been obtained with the help of human annotators against which the system generated summaries are compared for measuring their quality and usefulness.

Co-selection based evaluation

Co-selection based evaluation is done using sentence level information retrieval parameters like precision, recall and F-measure parameters. Precision represents the average probability of relevant retrieval and recall represents the average probability of complete retrieval and F-measure is the harmonic average of precision and recall. These scores are computed using Equations (8.1) to (8.3).

\[
\text{Precision} \quad P = \frac{TP}{TP + FP} \tag{8.1}
\]
Recall \[ R = \frac{TP}{TP + FN} \]  \hspace{1cm} (8.2)

F-measure \[ F = \frac{2PR}{P + R} \]  \hspace{1cm} (8.3)

where, TP- True Positive, TN – True Negative,
FP- False Positive, FN – False Negative.

Table 8.2 summarizes the average of co-selection based average precision, recall and f-measur scores obtained for the summaries generated from these datasets with two different size (number of sentences) of extract requested.

Table 8.2  Co-selection based evaluation of summarization method (in average)

<table>
<thead>
<tr>
<th>Summary Length (no. of sentences)</th>
<th>Domain</th>
<th>Summarization without using Pre-processed Segments</th>
<th>Summarization using Pre-processed Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>10</td>
<td>Science and Technology (D1)</td>
<td>0.4781</td>
<td>0.4152</td>
</tr>
<tr>
<td></td>
<td>Banking Services (D2)</td>
<td>0.5811</td>
<td>0.5132</td>
</tr>
<tr>
<td></td>
<td>Literacy (D3)</td>
<td>0.4374</td>
<td>0.4011</td>
</tr>
<tr>
<td>20</td>
<td>Science and Technology (D1)</td>
<td>0.4937</td>
<td>0.4321</td>
</tr>
<tr>
<td></td>
<td>Banking Services (D2)</td>
<td>0.6427</td>
<td>0.5733</td>
</tr>
<tr>
<td></td>
<td>Literacy (D3)</td>
<td>0.5483</td>
<td>0.4953</td>
</tr>
</tbody>
</table>

Performance of the system has been studied under various summary size requirements by keeping the summary length parameter as 10 and 20 and the evaluation was done and the experimental results obtained
confirms that the performance of the system is being maintained irrespective of the size of the summary required.

Table 8.2 indicates that the proposed summarization system using pre-processed segments performs better than summarization without using segments as the earlier method incorporates semantic relevance based measure also in scoring the sentence for their significance with respect to user query. It is also indicated that as the size of the summary increases, the respective precision and recall scores also increase.

**Content based evaluation**

The summaries were also evaluated using content-based similarity measures that compute the similarity between two summaries at a more fine grained level than just sentences. Content based measures like ROUGE (Recall-Oriented Understudy for Gisting Evaluation), consider this content based feature to evaluate the system summaries. There are several metrics within the ROUGE suite (Lin 2004a), but the most widely used are ROUGE-2 and ROUGE-SU4(www.berouge.com).

Performance of proposed summarization using pre-processed segments is compared with that of summarization without using pre-processed segments as well as two baseline summarizers of MEAD – Lead based (Dragomir Radev 2004a) and MEAD-random based summarizers on the three domain specific datasets and is presented in Figures 8.2 to 8.4 respectively.
Figure 8.2 Content based evaluation results on Dataset D1

Precision, recall and F-measure scores of the proposed summarization systems is 32.55%, 26.27% and 28.51% above than that of existing summarization approaches.

Figure 8.3 Content based evaluation results on dataset D2
Contents based evaluation results obtained for dataset D2 and D3 are shown in Figures 8.3 and 8.4 which also substantiate to the earlier conclusion that the proposed system outperforms the existing approaches.

**Evaluation of summarization process**

Effectiveness of the approach is also measured by means of the response time for summary generation process. Response time is the time taken by the algorithm to generate the summary during query time. Figure 8.5 compares the response time of the proposed system with existing system and indicates that the pre-processed segments during summary generation process considerably improves the response time as only few query related segments are processed to extract the short summaries.
According to the experimental outcome on these datasets, the
response time improves by 60.61% in an average when pre-processed
segments are utilized.

8.2.3 Evaluation of Comparative Summary

Comparative summaries have been generated for the same random
set of queries on each dataset with a few URLs selected from the search
results to view comparative text. Individual summary of each document in the
set is composed which has been evaluated using various measures discussed
in the previous subsection, are aggregated to obtain the comparative
summary.

Generation of comparative summary is a new idea and is evaluated
through the extrinsic task oriented evaluation method by involving human
subjects to judge the quality and usefulness of the summary. Each participant
was given a post-task questionnaire, to assess their level of agreement or
disagreement via a binary answering scheme (1-Agree, 0-Disagree). The final scores obtained were mapped to a 5-point Likert scale (Murray et al 2009).

![Graph showing user feedback in a five point Likert scale.]

**Figure 8.6 Extrinsinc evaluation results (Acceptability and usefulness)**

Users feedback obtained in five point Likert scale shows that the acceptability and usefulness of comparative summary generated is in most promising level (90.51%) and the comparative summary is resourceful and beneficial for real time web users interested in analyzing and comparing the query related gist extracted from a set of web documents.

This chapter has analyzed the impact of application of the proposed algorithms on three domain specific datasets and indicates that the performance of the proposed methodologies are better than the existing methods in view of various information retrieval evaluation parameters, response time and also space complexity. It is also indicated that the proposed algorithms are domain independent and is able to support any HTML document.