Chapter 7

Epilogue

7.1 Preamble

The problems of XML URL dataset creation, XML URLs Classification, RSS XML Web page segmentation, Flash URL dataset creation and Flash Web segmentation are addressed in this thesis. Novel methods of XML URL extractor, XML URL classification and RSS XML Web page segmentation are devised. In addition, we also created Flash URL dataset and proposed a method for Flash Web page segmentation by using image processing techniques. Besides, an algorithm is proposed to identify the noisy blocks after segmenting the Flash Web pages.

In this Chapter, the summary of all the proposed methods are presented. In sequel, the major contributions of the research work are listed. Subsequently, scope for further research based on the work presented in this thesis is envisaged.

7.2 Summary

Our Research work opens up a gateway to the new trends in browsing on small screen devices and Web mining applications on XML and Flash Web pages. It is necessary to develop a system to segment, remove noise and make these kind of new technology(s) Web pages to appear on small screen devices and to improve the Web mining applications such as search engine optimization, search result, search speed, Web page clustering, Classification and so on.
The work presented in this thesis takes us to several conclusions. We presented the importance of XML and Flash Web pages in Web mining research area and the problems we encountered in real life, described its different phases and methods to overcome the current issues. A brief overview of browsing on small screen devices: its challenges and approaches is presented. A detailed survey on start-of-art technique of browsing on small screen devices and Web mining applications which motivated us to take up the problem is presented as introduction in this work.

In Chapter 2, we have presented the approach followed to create a large collection of database which contains XML URLs. We have also mentioned the problems encountered while acquiring these URLs. In dataset analysis, we have described the different types of XML URLs extracted based on their semantic orientation by analyzing them manually. Also, we have described the performance analysis of proposed crawler with existing crawlers along with their pros and cons. Further by this we have, described about Flash dataset creation method. In our research, the data set created is utilized for experimentation purpose of our proposed preprocessing and recognition algorithms which is a very challenging task.

After creating the dataset as described in Chapter 3, to know the feasibility of existing systems, here based on the results obtained by the each existing systems are compared with respect to segmentation ratio. After choosing the best performance algorithms such as VIPS and Boilerpipe, they are used to test the feasibility analysis on our Data sets (Flash and XML). We observed that both the predominantly leading algorithms fail to segment and retrieve contents from our Data sets. Result tables clearly show that the existing systems are not viable for XML and Flash kind of Web pages. Therefore, to propose the system to segment and retrieve the contents, we performed the semantic structure analysis on both the Data sets. After analyzing the semantic structure, Translation system has
been proposed to customize the Flash contents into HTML structure to make them viable on SSDs. Here, the results obtained are tested on various hand held devices such as Sony Xperia X10, Sony Ericsson WT13i and LG Optimus Net, based on various specification details. After achieving better response time, resultant vectors are compared with the response time of respective conventional Web pages and Kappa statistical analysis also has been performed to analyze the content coverage on SSDs.

To achieve the classification system, we proposed four successive algorithms to create knowledge base. After performing all the four consecutive algorithms on testing data set elements, matching has been done. Based on the highest matching score Web pages are classified into their respective classes. After proposing an algorithm, we have conducted extensive experimentation on various ratio of data set and compared the obtained F-Measure and Accuracy score with each other. Overall, we have achieved average accuracy of 96.99%. We have proposed the segmentation Algorithm to segment the various types of XML RSS Web pages by introducing the HTML structure on XML RSS Web page semantic structure orientation. After achieving the segmentation, experiments are conducted to compute the proposed system performance and accuracy level, our proposed method achieve better performance as well as accuracy measures levels (96.88%, 89.71% respectively). Interesting future directions of research are, to introduce similar segmentation Algorithm for other XML Web pages types and to identify the noise blocks by analyzing the feature vectors of segmented blocks are discussed in Chapter 4.

After creating data set of Flash Web pages as images in Chapter 2, to segment Web pages as blocks, we proposed two different methods such as Haar Discrete Wavelet Transformation and Connected Component with Threshold concepts to segment Web pages as individual blocks in Chapter 5. We have conducted experiments based on human judgment to analyze the performance of proposed
systems. Here, we have achieved the overall performance of 80%, 86.46% respectively for the dataset considered. Also experiments are conducted on accuracy measures such as Recall, Precision and F-Measure. Our proposed methods achieve average 73.16% and 83.99% F-Measure respectively. Interesting future directions of research are to improve accuracy based on performing AND operations on matrices and to identify the noise blocks by analyzing the feature vectors of segmented blocks.

Block merging concepts have been introduced after segmenting the Web pages using fusion method in Chapter 6. Here, block merging concepts drastically improve the accuracy measures when compared with the existing connected component method. Proposed block merging concept outperforms other methods by achieving the average accuracy level of 95.51%. And also F-scores of proposed method has been compared with the F-scores of connected component method, the graph plotted clearly states that the out-performance of the accuracy level of proposed method.

In the final stage, after proposing the segmentation algorithm by using the block merging concepts and achieving the average accuracy level of segmentation 95.51% in Chapter 6, we have proposed an algorithm to identify and remove the noisy blocks and retrieve the informative contents based on block feature extraction and analysis. Here, we have extracted the various block features such as block contributions over the Web page, number of lesser blocks, number of greater blocks, spatial location of the block and number of repetition over the corresponding website Web pages. After extracting the feature, based on analysis decision has been made by considering the feature levels of noisy blocks. Our proposed method achieves average accuracy level of 82.33%. Finally, informative blocks are extracted based on the labels maintained.

The robustness and effectiveness of all the proposed models are brought out through extensive experiments on our datasets in the respective chapters.
Nevertheless, the superiorities of the proposed models in terms of classification accuracy and F-Measure are established by experiments where the proposed methods outperform in terms of segmentation and noise removal.

In brief, the following are the contributions of the research work presented in this thesis.

### 7.3 Contributions

- Creation of domain specific datasets called XML URL dataset and Flash URL dataset.
- Performed the comparative analysis of existing system results to opt the better systems to check feasibility on our dataset.
- After choosing the predominantly dealing algorithms, at very first step showed that the feasibility of existing systems fail to perform on our datasets.
- Introduction of a new method of Flash Web page customization into HTML structure is proposed.
- Proposal of novel XML URLs classifier based on semantic orientation.
- Successful exploration of a new novel method to segment RSS XML Web pages.
- Exploitation of segmentation techniques for Flash Web page segmentation based on image perception using image processing techniques.
- Introduced various block merging concepts to achieve better partition granularity.
- After achieving betterment of segmentation level, based on block features noisy blocks are identified and removed.
- Finally, by using labeling concepts informative blocks are retrieved for Web mining applications.

Finally, this thesis has addressed a new avenue of research on XML and Flash Web pages.
7.4 Future Research Avenues

The work presented in this thesis has provided a number of research opportunities. But, there are still some open problems to be addressed. Indeed to the best of our knowledge, this is the first attempt to classify, segment and to remove noise from XML, Flash Web pages. Therefore, the presented research work is expected to open up a new research avenue for exploiting the applicability of XML and Flash Web pages in Web mining applications.

The following is a list of some open problems that we plan to address in future research work:

- Our proposed crawler can be deployed on vast number of collected Web domains to improve our dataset and bringing the research avenues on XML URLs.
- Using Flash layers, segmentation algorithm can be proposed by retrieving the layers properties such as height, width and location etc. This kind of feature extraction and based on analysis proposing segmentation algorithm shall be an interesting future work.
- After analyzing the user defined tags of Pure XML URLs, proposing an generalized segmentation algorithm for Pure XML Web based on their semantic structures shall be an outstanding future for Web mining applications on XML URLs.
- As similar to Flash Web page noise removal algorithm, features can be analyzed for segmented blocks of Pure XML Web pages shall be an interesting area in the near future.
- Proposing an algorithm to adapt the extracted informative blocks without content loss after removing the noise could be an interesting extension work.