Chapter 8

Conclusion

In the vast and nebulous research done to classify unstructured data from web pages, there was hitherto no single effort to compare multiple algorithms for a representative, multi-domain dynamic and text-rich webpage dataset. This project has fulfilled this need by examining Naive Bayes, K Nearest Neighbours, Decision Trees, Logistic Regression and Random Forest. They have been compared upon various parameters and the results have been used in deciding on a suitable algorithm for use in the project. Random Forest algorithm was chosen as it achieved 92.7% precision and outperformed all other classification algorithms. The project has used novel methods of text extraction which require least processing resources in terms of feature subsetting and extraction of data from provider-specific pages without unnecessary inflation of data size. Hitherto unused features such as social tagging from websites like Delicious have also been included to improve the quality and relevance of data. The web data of over 30,000 web pages crawled and classified using Random Forest algorithm by the system. The offline repository covers over 10,000 courses and over 60 providers all over the world.

This system is presented in an innovative application that can work offline and allow the user to bookmark and view course information in a crisp and concise manner with a fresh and intuitive front end interface. It will provide a hand reference to students on the move as well as academic institutions. It can even update based on the latest courses and display information which is dynamically generated based on training set data.

8.1 Limitations

The main limitation of this is that there is no way to assure 100% accuracy of courses labelled by the classifier. This is because the very aim of this project is to eliminate manual effort in collection and management of data. However due to inconsistent HTML code and unconventional webpage design, errors might creep in to the code of the pages crawled which the classifier has not estimated or encountered earlier. These could include common HTML design errors or larger dynamic script related bugs. The system will have to be retrained and reset for adding new providers.

8.2 Future Enhancements

The system has scope for multiple enhancements that depend on advancements in web and data mining technologies. The future enhancements that address the advancements are listed and explained below:

- Super-fast multithreaded crawlers or even a server resource with indexing capacity can be used to fetch and store a large number of web pages quickly. This will make it possible to re-train the system easily whenever required.
- Use of Twitter, Facebook and other improved feeds if available offline will help to enhance the quality of data in social tagging.
- The system can be expanded to cover other languages using Natural Language Processing concepts for different scripts and providers.
- Inclusion of multimedia in the system can enable users to view video courses offline using efficient resource utilization to fetch and classify this data

The system has been designed keeping in mind the potential to convert it to a fully functional mobile application. This can be taken further and made available to users across multiple mobile platforms such as Android, Windows Phone and iOS so that it is accessible to multiple users and will help those in remote areas who do not have fast internet connection to efficiently plan and schedule the courses they wish to take.

8.3 Summary

The eighth chapter describes how the project work has fulfilled the research gap. It has mentioned about new features such as social tagging that were used and main front end display features. The limitations of the project work were discussed that mainly holds inconsistent HTML code and inconsistent web page design responsible for inefficient crawling and classification. The last part of the chapter discussed the scope of the project that addresses the future enhancements of the project.