Abstract

Web on the Internet has become an enormous repository of data. There have been many efforts to provide efficient access to relevant information in the very large data repository. One way to provide efficient access is by way of web news content extraction with primary focus to take the information in the web news.

In this Final Project implemented a method to extract key information on news web pages by using the method called Hybrid. This technique is trying to take advantage of the sequence matching techniques and tree matching. Data structure used is TSReC, a variant of tag sequences representation suitable for both sequences matching techniques and tree matching.

From analysis and test results stage shown that that Hybrid method is built proved to can get news content on news Web pages, although in some datasets, there are still noise.

Keyword: web news content extraction, sequence matching, tree matching, TSReC.