INFORMATICS INSTITUTE OF TECHNOLOGY, SRI LANKA

In Collaboration with The

UNIVERSITY OF WESTMINSTER, UK

"*Crawl the News*"

(Video and Article News Recommendation using Web Crawling)

A dissertation by:

Ms. C. L. Melissa Diaz

Supervised By:

Mr. Torin Wirasingha

Submitted in partial fulfilment of the requirements for the

BSc (Hons) in Computer Science specialized in Mobile & Web Computing Degree

Department of Computing

May 2019

Abstract

The Web is an ocean of information covering a vast area of topics and is constantly updated with more and more information daily. With the advancements of the WWW and the expansion of the programmable web, more and more applications and services have begun to be increasingly data-driven. At present, the two main methods used to gather required data involve using APIs from publishers or Web Scraping. Even though they are being used widely, APIs have some limitations when gathering data.

Many people must visit multiple websites and YouTube news channels, spending their precious time, to get their daily dose of news. This research will focus on introducing a data retrieval mechanism using web crawling in the News Domain to recommend news articles with videos and visuals and help minimize time wasted by users having to visit multiple sources.

In this research a web crawling mechanism was introduced that can crawl multiple websites with different DOM structures parallelly. The crawled news is categorized based on their headlines using text classification models and then recommended to the users by mapping the news categories with the news category preferences set by the users.

Subject Descriptors:

- World Wide Web
  - Web searching and information discovery
    - Web Crawling
- Machine Learning
  - Learning Paradigms
    - Supervised learning
      - Supervised learning by classification

Keywords: Information Retrieval, Web Crawling, Machine Learning, Logistic Regression, Text Classification, News Recommendation