Abstract No: SP-08 **Software Intensive Systems**

An online news crawling framework for an aggregated news site

W. A. I. K. Wagawaththa* and W. M. J. I. Wijayanayake

Department of Industrial Management, Faculty of Science,
University of Kelaniya, Sri Lanka
ishanthakamal@gmail.com

The internet has become one of the most widespread platforms for information exchange and retrieval as the number of news websites is increasing rapidly. During the last decade, most of the major newspapers have developed web sites providing news and other information. In addition, web-only newspapers have also appeared. News aggregator is a good substitute for news sites like BBC news. Because news aggregators can index not just the content of the BBC news but all other news sites, giving it a huge advantage in coverage. On the other hand, news aggregators may complement online news sites. Because news consumers incur costs (time and effort) in searching for news that are important to them and also they will compare the expected benefit from visiting a news site to the expected search cost, where that cost includes becoming aware of the existence of the site and finding how to navigate it.

There are few news aggregators like Google News, News Look Up, Fark which provide news aggregation facilities, but they are proprietary and there are privacy concerns about the user along with the biasness of these aggregators. In order to benefit more from the available information, the objective of the research is to develop a technical framework, gathering online news and approach to recognize most important latest news and display the recognized news items that society is interested in without any bias. Presenting crawled news items in a way that it displays the trending topics in society will increase the awareness of the reader. In order to do that news classification and ranking is a needed.

News items for the framework will be gathered through (*RSS*) feeds. Gathered news feeds will be stored and will be preprocessed. Keywords will be extracted from an algorithm that can be worked with any language that has basic Morphological tools for language processing. Category classification of the news items will be done using a method that is based on the keyword extraction algorithm. Topic detection and classification of the news items will be done to the category classified news items using an algorithm that requires no corpus for statistics or training data. The ranking of the news article, topic and source will be done using an approach which is based on the virtual graph model. In the ranking process, similarity between articles are calculated manually and it will be automated using the cosine similarity.

Keywords: Information retrieval, Category classification, Topic classification, News ranking