Abstract No: SO-05 **Software Intensive Systems**

An algorithm for plagiarism detection in Sinhala language

Subash F. Basnayake*, Himesha Wijekoon and T. K. Wijayasiriwardhane

Department of Industrial Management, Faculty of Science,
University of Kelaniya, Sri Lanka
yadsb1@stu.kln.ac.lk

According to the Merriam-Webster dictionary, the simple definition of the verb plagiarize is, "to use the words or ideas of another person as if they were your own words or ideas". Many software tools to aid in detecting plagiarism is available for English language, but equivalent tools are not yet available specifically for Sinhala language. Though language independent tools that work on many languages are available, they generally give poor results as they do not consider language specific features.

There are some detection methods proposed for Asian languages like Hindi, Malayalam, Arabic and Persian which have some close relationship and similar properties of Sinhala language. All of those methods use language specific rules and they even outperform the commercially available tools. These findings are evidence that the language specific plagiarism detection is more effective than the language independent plagiarism detection as some paraphrasing techniques can be used to mislead the language independent systems.Sinhala language is constitutionally recognized as the official language of Sri Lanka, along with Tamil. Due to the complexity of the language structure and rules of grammar, the language independent tools seem to provide poor results when used for plagiarism detection in Sinhala documents.

In this research, we propose a novel plagiarism detection algorithm built around content based methods specific to Sinhala language. The methodology of this study follows both experimental and build approaches. The proposed plagiarism detection system has two modules namely, text pre-processing module and the similarity detection module. The text pre-processing module pre-process the text files to standardize the text sources using techniques such as stop word removal, number replacement, lemmatization, synonym recognition and creating n-grams. Then the similarity detection module analyses the pre-processed text using *Jaccard* coefficient and cosine similarity coefficient to measure the similarity between two documents. A prototype of Sinhala language plagiarism detection system will be implemented using the proposed method and several combinations of the above techniques will be used to discover the best combination. Testing and statistical performance evaluation will be carried out using a sample of source text files and plagiarized text files in Sinhala language by taking expert judgements also into the consideration. The final outcome of this research study is to develop an effective software application for plagiarism detection in Sinhala language documents.

Keywords: Natural language processing, Plagiarism detection, Sinhala language