2019 5th International Conference for Convergence in Technology (I2CT)
Pune, India. Mar 29-31, 2019

A Review On Language Specific Multi Document Similarity Detection

Achala Piyarathna
Informatics Institute of Technology
No 57, Ramakrishna Road,
Colombo 6, Sri Lanka
+94 (0) 71 8686836
achalayasas007@gmail.com

Guhanathan Poravi
Informatics Institute of Technology
No 57, Ramakrishna Road,
Colombo 6, Sri Lanka
+94 (0) 77 342330
guhanathan.p@gmail.com

***Abstract* - Plagiarism is exploitation of others work and presents them as your own without referencing the original work. There are various detection tools that are being developed in order to detect these plagiarized content. Most of the available detection tools are based on the English language. Though there are language independent and language-specific detection tools, there is no comprehensive multi-document plagiarism detection mechanism. If the already available work on other language-specific tools and Similarity detection tools are analyzed and find what has been missing, it will be a stepping stone to continue research on this area. This paper contains the underlying piece of a continuous research, and later on, we plan to use this learning to present a comprehensive research on the subject area.**

Keywords - Plagiarism Detection, Language dependent model, Asian languages, Natural Language Processing, Text preprocessing

I. INTRODUCTION

Plagiarism has been a serious problem mainly academia. This is mostly due to the easy way of copying the electronic documents and to the difficulty of detecting similar documents in a sufficiently large database [1].

With the explosion of using online data in the forms of documents, articles, research papers and assignments etc. the need of plagiarism detection emerged in order to analyze and identify the plagiarized content. Plagiarism detection can be classified in to two areas as language-dependent and language independent. When it comes to language specific plagiarism detection, there are many factors to be considered and is dedicated to a particular natural language [2].

There are researches and tools developed to detect plagiarism in English language and they are only vary in performance and algorithm vice. Considering the already available tools for language independent similarity detection tools the main drawback is that these tools only consider the common patterns between the two text corpuses rather than checking the language specific mechanisms.

However when considering the Asian languages there are only few researches done on similarity checking between two text corpuses and there is no comprehensive multi document plagiarism detection tool is available.

By analyzing and identifying the key research features from the already available researches on language independent language dependent and specific similarity detection mechanisms, we are trying to provide a solution for the multi document plagiarism detection for Asian languages. This analysis will help to identify current approaches, and possibilities for new approaches in future.

In the following section we have addressed on the methodology that we have used in our initial phase of the research to gather the knowledge on the domain. Section 3 will elaborate the background and the details in depth about the Language Specific Plagiarism Detection followed by the problem and the gap.

Section 4 will present the available work on Asian language based similarity detection, WordNet and other language specific plagiarism detections approaches. Section 5 discusses about what is missing and what are the needful to create comprehensive language specific Plagiarism Detection tool. A complete analysis of the available work has been discussed in the section 6 followed by the future work and conclusion of the paper in the section 7.

II. METHODOLOGY

We started with the literature survey to get the domain knowledge on how plagiarism works and the algorithms underneath it. Primarily, the techniques used to detect plagiarism are classified in to: global and local. Global methods evaluate the characteristics from large data chunks and local methods go through the pre-defined data from the documents [3].

For language specific plagiarism detection, there should be a powerful algorithm that should consider the linguistic properties of a particular language. Though there are language independent tools that works well with many languages they usually perform poorly since they do not take the language specific properties in to consideration [4].

Since our focus is on developing a comprehensive language specific plagiarism detection tool, we conducted our research on the Asian languages and the linguistic features that should be taken in to consideration. As for any natural language project we found that a WordNet for each of the languages is extremely important for our purpose [5].

978-1-5386-8075-9/19/$31.00 ©2019 IEEE