ABSTRACT

Plagiarism becomes popular is caused by the ease of obtaining sources of information in writing or doing the task, whether it is done offline or online. This encourages the development of various softwares that can be used to help the task of plagiarism detection. The objective of this study was to design the similarity indication engine (called PlagON) to detect plagiarism indication in Indonesian document. The detection process including preprocessing which consists of stopword removal and stemming Nazief-Adriani, tokenization with word n-gram method, documents title filtration, string matching using Knuth Morris Pratt algorithm, and similarity value was calculated with the Dice coefficient. A web crawler feature was added to collect online link manually and to build offline mirror sites as a detection sources.

The results obtained showed that PlagON has been successfully constructed to deal with cases of plagiarism that was done by copy paste and modifying the contents of the document. High similarity value comes from the small usage of word n-grams (except unigram). On the other hand, document title filterisation through fourgram achieved good result to decrease false positive with average of precision and recall are 0.8 and 1. PlagON has given higher percentage of similarity value rather than Ferret tools and it was proofed to be more complete in giving presentation of similarity indication result in terms of its feature when it is compared to the other plagiarism detection tools like Duplihecker, Plagiarisma, and Plagium. PlagON was still limited to detect plagiarism on duplicated document from the site that have never been through *crawling* before. The failure of filterisation title process through finding relevant documents also leads to the failure of giving similarity indication inter-document.

Keywords : document, similarity, plagiarism, word n-gram.

INTISARI

Maraknya kasus plagiarisme yang terjadi disebabkan kemudahan dalam mendapatkan sumber-sumber informasi dalam mengerjakan tugas atau penulisan, baik *offline* maupun *online.* Hal ini mendorong pengembangan berbagai perangkat lunak yang dapat digunakan untuk membantu tugas deteksi plagiarisme. Tujuan dari penelitian ini adalah merancang mesin pengindikasi kemiripan (disebut PlagON) untuk melakukan deteksi indikasi plagiarisme pada dokumen berbahasa Indonesia. Proses deteksi terdiri atas *preprocessing* yang meliputi *stopword removal* dan *stemming* Nazief-Adriani, tokenisasi dengan metode *word n*-gram, filterisasi judul dokumen, pencocokan *string* menggunakan algoritme *Knuth Morris Pratt,* dan nilai kemiripan dihitung menggunakan *Dice coefficient.* Fitur *web crawler* ditambahkan untuk mengumpulkan tautan *online* secara manual dan berguna untuk membangun *offline mirror sites* sebagai sumber deteksi.

Hasil dari penelitian ini menunjukkan bahwa PlagON sudah berhasil melakukan deteksi plagiarisme dengan kasus salin-menyalin dan modifikasi isi dokumen. Nilai kemiripan yang tinggi dihasilkan oleh penggunaan nilai *word n*-gram yang kecil (kecuali *unigram*), sedangkan filterisasi judul dokumen dengan penggunaan *fourgram* cukup baik untuk mengurangi jumlah *false positive* dengan nilai rata-rata *precision* dan *recall* adalah 0,8 dan 1. PlagON memberikan persentase nilai kemiripan lebih tinggi jika dibandingkan dengan *tools* Ferret dan lebih lengkap dalam memberikan penyajian hasil indikasi kemiripan dilihat dari sisi *feature* jika dibandingkan dengan *tools* deteksi lainnya, yaitu DupliChecker, Plagiarisma, dan Plagium. PlagON masih terbatas untuk mendeteksi dokumen plagiat yang disalin dari situs yang belum pernah di-*crawling* sebelumnya. Kegagalan proses filterisasi judul dalam penemuan dokumen relevan juga menyebabkan PlagON gagal memberikan indikasi kemiripan antardokumen.

Kata kunci – dokumen, kemiripan, plagiarisme, *word n*-gram.