

Penerapan Algoritma *Enhanced Confix Stripping* Dan Rabin-Karp Untuk Mendeteksi Kemiripan Dokumen Teks Bahasa Indonesia

Penulis

Ainur Rofik

150411100029

Dosen Pembimbing 1

Dosen Pembimbing 2

Sigit Susanto Putro, S.Kom., M.Kom.

Husni, S.Kom., M.T.

19790313 200604 1 002

19770722 200312 1 006

ABSTRAK

Information retrieval merupakan kegiatan memperoleh data informasi yang relevan, pencarian informasi dapat didasarkan pada teks lengkap atau pengindeksan berbasis konten lainnya. *Information retrieval* terdiri dari tiga komponen utama, yaitu masukan (*input*), pemroses (*processor*), dan keluaran (*output*). Plagiarisme merupakan pelanggaran utama akademik, namun sebagian dari institusi perguruan tinggi yang telah berkembang dengan baik sudah menerapkan kebijakan dan prosedur untuk menangani masalah tersebut. Selain itu, plagiarisme juga digambarkan sebagai tindakan pencurian karya orang lain dan membuat pekerjaan tersebut adalah pekerjaan sendiri. Pendeteksian plagiarisme di Indonesia hanya menggunakan cara sederhana pra-pemrosesan teks seperti menghilangkan spasi, dan karakter yang tidak dibutuhkan. Metode semacam ini hanya bisa mendeteksi plagiarisme dengan cara sederhana tetapi tidak dapat mendeteksi bentuk yang lebih kompleks. Metode *Enhanced Confix Stripping* merupakan pengembangan algoritma terdahulunya dengan beberapa perbaikan kesalahan pada metode sebelumnya dan metode *Rabin-Karp* sebagai tahap membandingkan nilai *hash* antara string masukan dan substring dalam teks. Berdasarkan hasil pengujian sistem menggunakan metode *Rabin-Karp* dikombinasikan dengan *Enhanced Confix Stripping Stemmer* dengan akurasi rata-rata 24.71% dan tanpa menggunakan *Stemming* dengan akurasi rata-rata 26.28%.

Kata Kunci: *Information retrieval, kemiripan dokumen, Enhanced Confix Stripping, Rabin-Karp, Stemming*

Application Of Enhanced Confix Stripping And Rabin-Karp Algorithm To Detect Similarity Of Indonesian Text Documents

Penulis

Ainur Rofik

150411100029

Dosen Pembimbing 1

Dosen Pembimbing 2

Sigit Susanto Putro, S.Kom., M.Kom.

Husni, S.Kom., M.T.

19790313 200604 1 002

19770722 200312 1 006

ABSTRACT

Information retrieval is an activity to obtain relevant information data, information search can be based on full text or other content based indexing. The information retrieval system consists of three main components, namely input, processing, and output. Plagiarism is a major academic violation, however, a number of well-developed tertiary institutions have implemented policies and procedures to deal with the problem. In addition, plagiarism is also portrayed as an act of stealing other people's work and making work as if it were one's own work. The detection of plagiarism in Indonesia only uses simple methods of text pre-processing such as removing spaces, and unneeded characters. Such methods can only detect plagiarism in a simple way but cannot detect more complex forms. The Enhanced Confix Stripping method is the development of the previous algorithm with some error correction in the previous method and the Rabin-Karp method as a stage of comparing hash values between input strings and substrings in the text. Based on the results of system testing using the Rabin-Karp method combined with Enhanced Confix Stripping Stemmer with an average accuracy of 24.71% and without using Stemming with an average accuracy of 26.28%.

Keywords : *Information Retrieval, Document Similarity, Enhanced Confix Stripping, Rabin-Karp, Stemming*