The Construction of Indonesian-English cross language plagiarism detection system using fingerprinting technique

Zakiy Firdaus Alfikri, author

Abstrak

Deteksi plagiarisme lintas bahasa merupakan hal yang penting untuk melindungi hak kekayaan intelektual. Bahasa Inggris adalah bahasa internasional yang paling populer, karenanya peneliti mengusulkan deteksi plagiarisme lintas bahasa Indonesia-Inggris untuk menangani masalah tersebut di mana domain dokumen yang diduga plagiat ditulis dalam bahasa Indonesia dan dokumen sumber ditulis dalam bahasa Inggris. Untuk meminimalkan kesalahan terjemahan, peneliti membangun sistem dengan menerjemahkan dokumen bahasa Indonesia ke bahasa Inggris dan kemudian membandingkan dokumen yang diterjemahkan dengan koleksi dokumen bahasa Inggris. Sistem pendeteksian ini terdiri dari komponen preprocess, komponen pencarian heuristik, dan komponen analisis detail. Teknik utama yang digunakan dalam temu kembali informasi adalah fingerprinting yang dapat mengekstrak fitur leksikal dari teks yang cocok digunakan untuk mendeteksi plagiarisme dengan menggunakan metode terjemahan harfiah. Dalam tulisan ini, peneliti juga mengusulkan metode-metode tambahan yang akan diimplementasikan dalam komponen pengambilan heuristik untuk meningkatkan kinerja system seperti chunking frase, penghilangan stop word, stemming, dan pemilihan sinonim. Peneliti mengevaluasi kinerja sistem dan efek dari metode tambahan untuk kinerja sistem, dengan menyediakan sekumpulan skenario tes beberapa data yang merepresentasikan plagiarisme. Dari pengujian diperoleh kesimpulan bahwa sistem bekerja pada 83,33% kasus uji. Peneliti juga menyimpulkan bahwa terutama semua metode tambahan kecuali chunking frase memiliki efek yang baik dalam meningkatkan akurasi sistem.

<hr>

Abstract

Cross language plagiarism detection is an important task since it can protect person intellectual property right. Since English is the most popular international language, we proposed an Indonesian-English cross language plagiarism detection to handle such problem in Indonesian-English domain where the suspected plagiarism document is written in Indonesian and the source document is written in English. To minimize translation error, we build the system by translating the Indonesian document into English and then compare the translated document with the English document collection. The detection system consists of preprocess component, heuristic retrieval component, and detailed analysis component. The main technique used in retrieval process is fingerprinting which can extract lexical features from text which is suitable to be used to detect plagiarism done using literal translation method. In this paper, we also propose additional methods to be implemented in heuristic retrieval component to increase the performance of the system: phrase chunking, stop word removal, stemming, and synonym selection. We evaluated system?s performance and the effects of additional methods to system?s performance, provided several data test sets which represents a plagiarism type. From the experiments, we concluded that the system works on 83.33% of test cases. We also concluded that mainly all additional methods except the phrase chunking have good effects in enhancing the system accuracy.