EKSTRAKSI TEKS OTOMATIS DARI HALAMAN WEB BERBAHASA INDONESIA GUNA MEMBANTU MEMPERCEPAT PENYUSUNAN KORPUS

Fatchurrohman, M. Kom

Zainal Abidin, M. Kom

ABSTRAK

Penelitian ini memandang halaman web sebagai barisan *string-string.* Setiap *string* dari penyusun halaman web mempunyai dua kemungkinan isi, yaitu label-label sintak *html* dan bukan sintak *html.* Kandungan utama dari halaman web merupakan string yang cukup panjang, bahkan mempunyai kecederungan string yan paling panjang. Dalam penelitian ini, string terpanjang dipilih sebagai *string* yang berisikan kandungan utama dari halaman web. Ekstraksi teks secara otomatis didahului dengan penyaringan label-label sintak html. Label-label hasil penyaringan diimpan di database label demi label (kata). Database sebagai pengelola data hasil penyaringan dengan menggunakan SQL. SQL menghitung jumlah kata dalam tiap-tiap baris, kemudian cari baris yang paling banyak. Semua kata yang berada didalam baris dengan jumlah kata terbanyak diambil dari database dan diurutkan berdasarkan pada urutan kata dalam baris. Penelitian ini diujicobakan pada 120 halaman web yang diunduh dari empat situs, yaitu: tempointeaktif.com, mediaindonesia.com, jawapos.co.id, dan cetak.kompas.com. halaman web yang diunduh adalah halaman web yang dipasang pada bulan januari 2009. Uji coba menunjukkan tingkat keberhasilan 76% dapat mengekstraksi teks dengan baik. Halaman web sejumlah 91 buah berhasil diekstraksi dengan otomatis. Ketidakberhasilan ekstraki otomatis disebabkan: penyaringan label html kurang bekerja dengan baik karena penulisan tanpa pemenggalan, adany computer terhadap kandungan utama halaman web dengan panjang melebihi panjang kandungan utama, penulisan kandungan utama dalam beberapa baris.

Kata kunci: ekstraksi, teks, corpus, database