INTISARI

EKSTRAKSI METADATA BERITA *ONLINE* BERBAHASA INDONESIA MENGGUNAKAN *TEXT DETECTION FRAMEWORK*

Oleh

Firly Armanda

12/331320/PA/14590

Media *web* berita mengandung sumber teks yang menyediakan informasi yang luas dan kaya untuk aplikasi penggalian teks salah satunya adalah ekstraksi informasi. Tetapi konten utama dari berita biasanya dikelilingi berbagai konten yang tidak berelasi dengan topik seperti iklan, spanduk dan tautan navigasi. Kemampuan untuk memilah dan memilih informasi tetapi tetap dapat menjaga keutuhan informasi yang penting akan berguna untuk membangun teks corpora yang bersih dan berpotensi meningkatkan sistem pencarian teks.

Pada penelitian ini dilakukan sebuah pendekatan untuk ekstraksi metadata berita *online* berbahasa Indonesia dengan judul berita, tanggal rilis dan konten berita sebagai target ekstraksi, dengan menggunakan metode *text detection framework*. Metode ini menghitung nilai *Compound text-tag difference* (CTTD) yang dijadikan sebagai acuan untuk mengekstrak metadata berita. Data set yang digunakan berjumlah 102 halaman HTML yang berasal dari 21 kanal berita yang tersebar pada 10 portal berita. Berdasarkan hasil pengujian didaptkan nilai presisi sebesar 94.5%, *recall* sebesar 99.3% dan *f-measure* sebesar 96.6%.

Kata Kunci: ekstraksi informasi, ekstraksi metadata, *text detection framework*

ABSTRACT

METADATA EXTRACTION ONLINE INDONESIAN NEWS BASED ON TEXT DETECTION FRAMEWORK

by

Firly Armanda

12/331320/PA/14590

Web news media contains a huge amount of text sources that provide wide coverage and rich text information for many text mining application such as information extraction. However the main text in web news is usually surrounded by various unrelated content, such as commercial ads, banners, link navigation and user comments. The ability to filter noisy information while preserving the important information is useful for constructing clean text corpora and potentially improving text retrieval system.

In this research was conducted on approach for metadata extraction of Indonesian online news, with news title, release date and the main content as targets, using a method based on text detection. This proposed method calculates the value of Compound Text-Tag Differences (CTTD) which is used as a reference for news metadata extracting. Dataset are used as much as 102 HTML documents which came from 21 kanal news which spread from 10 online news sites. Based on the evaluation results obtained from precision is 94.5%, recall is 99.3% and f-measure is 96.6%.

Keywords: information extraction, metadata extraction, text detection framework