Abstrak

Suatu halaman berita website biasanya banyak mengandung konten informasi dalam tiap-tiap blok halaman yang ditampilkan. Kadangkala konten berita yang ditampilkan pada halaman berita di suatu website tidak sepenuhnya memberikan informasi yang relevan atau tidak berhubungan dengan konten utama misalnya, panel navigasi, *copyright*, *user guide*, *links*, sinopsis suatu berita, berbagai macam iklan dan lain-lain. Blok-blok informasi yang tidak relevan dengan konten utama tersebut dikenal sebagai *web pages noise.*

Dalam tugas akhir ini akan digunakan teknik *Style Tree* untuk mendapatkan presentation style (layout) secara umum dan konten aktual dari halaman web dengan menggunakan *sampling* beberapa halaman website.Pertama kali seluruh halaman web akan dimodelkan dengan *DOM* tree , lalu penggabungan *DOM* menjadi *Style Tree* untuk memperoleh struktur umum dan pemecahan blok-blok informasi dalam website.Informasi yang didapatkan digunakan untuk melakukan pengukuran dan mengevaluasi tingkat kepentingan dari masing-masing node hingga pemberian bobot pada masing-masing individual word *(feature)* pada masing-masing blok konten. Hasil pembobotan *(weighting)* akan digunakan untuk mengukur performansi hasil *preprocessing* dengan cara klasifikasi untuk mendapatkan nilai F-measure.

Kata Kunci : *Style Tree, DOM, weighting, identifikasi noise, web mining.*