Abstrak

Sebuah halaman web biasanya mengandung berbagai jenis content seperti navigasi, dekorasi, dan bagian-bagian lain yang tidak berhubungan dengan inti informasi dari halaman web tersebut. Di sisi lain, kadang pengguna sebenarnya hanya membutuhkan informasi inti dari halaman tersebut. Dari sinilah muncul kebutuhan akan sistem yang dapat mengekstrak informasi dari suatu halaman web.

User melihat sebuah halaman web melalui web browser dan mendapatkan representasi 2D yang mempunyai banyak *visual cues* (penanda visual) untuk membantu membedakan bagian bagian yang berbeda dari halaman tersebut. Seorang *web designer* biasanya mengorganisasi *content* dari sebuah halaman web agar mudah untuk dibaca/dipahami oleh user. Oleh karena itu, *content-content* yang berhubungan secara semantik biasanya diletakkan dalam satu kelompok dan halaman web tersebut dibagi menjadi region-region untuk *content* yang berbeda dengan menggunakan pembeda visual seperti garis, ukuran *font*, warna, dll. Content-content yang sejenis biasanya akan ditampilkan dengan bentuk visual yang sama atau sejenis pula. *Visual cues* inilah yang akan dimanfaatkan untuk proses identifikasi dan ekstraksi data. Metode *Visual-Based Page Segmentation* akan memanfaatkan penanda visual (*visual cues*) dari halaman web untuk mengekstrak data dari halaman web tersebut.

Tahap analisis dan pengujian memberikan hasil bahwa *pattern visual cues* yang tepat terbukti dapat dimanfaatkan untuk membuat sistem ekstraksi informasi dari halaman web meskipun masih terdapat *noise*.

Kata kunci: halaman web, ekstraksi, *visual cues*, *pattern*, *noise*.