Abstrak

Pada *World Wide Web* atau yang biasa dikenal dengan WWW memiliki berbagai jenis informasi yang terkandung didalamnya. Biasanya *user* akan menggunakan *search engine* atau mengikuti *link* terkait untuk menemukan informasi yang mereka butuhkan. Akan tetapi, pencarian dengan menggunakan *search engine* terkadang tidak efektif karena menghasilkan banyak data dan *link* terkait yang membutuhkan waktu tidak sedikit untuk menulusurinya satu persatu, bahkan kadangkala hasil yang keluar sama sekali tidak terkait dengan *keyword* yang *user* masukkan. Setelah diteliti ternyata halaman web yang memiliki informasi yang sama memiliki struktur yang sama pula, selain itu diketahui pula bahwa standar dalam menerbitkan halaman web yang tidak ketat menyebabkan si penulis dapat menggunakan kata atau pelabelan yang berbeda untuk halaman web yang memiliki informasi yang sama.

Pada tugas akhir ini akan dibangun sebuah sistem yang dapat mengklasifikasi halaman web berdasarkan kelasnya dengan menggunakan algoritma *label discovery*. Pada awalnya algoritma *label discovery* (LDA) akan mencari label atau kata yang merepresentasikan kelas halaman web sehingga akan didapatkan struktur dari kelas halaman web yang diinginkan. Setelah struktur ditemukan maka akan digunakan untuk mengklasifikasikan halaman web.

Hasil pengujian menunjukkan bahwa LDA dapat menemukan kata atau label yang merepresentasikan suatu kelas dan struktur yang dihasilkan dapat mengklasifikasi halaman web secara akurat.

Kata Kunci : *label discovery*, struktur, klasifikasi, kelas, halaman web, fungsi kesamaan