

Abstrak

Saat ini halaman *web* bertambah sangat banyak dan berkembang begitu cepat dan menjadi salah satu sarana penyebaran informasi baik itu personal, sosial maupun komersial. Semakin banyak pula orang yang membutuhkan informasi mengenai topik-topik tertentu misalnya tentang olahraga namun mengalami kesulitan untuk mendapatkan informasi yang relevan. Untuk itu dibutuhkan *Web Crawler* khusus untuk membantu pengguna internet mencari halaman yang relevan. *Web crawler* sendiri adalah suatu program yang melakukan proses scanning ke semua halaman-halaman internet untuk dibuat indexnya dan mendukung sebuah *search engine*.

Berbeda dengan *crawler* yang dipakai oleh *search engine* komersial yang pada umumnya bertujuan untuk mengumpulkan halaman Web sebanyak mungkin, *focused crawler* (juga sering disebut dengan *topical crawler*) secara selektif menelusuri dan mengambil halaman Web yang relevan dengan topik tertentu. Dalam tugas akhir ini, digunakan *classifier Naïve Bayes* untuk membedakan halaman web olahraga dan bukan olahraga, serta menggunakan *Best First Search* sebagai algoritma penelusuran antrian. Pemilihan nilai yang terbaik dilakukan dengan membandingkan skor hasil perhitungan *Cosine Similarity*.

Ditunjukkan bahwa algoritma *best first search* dan *classifier Naïve Bayes* akan membantu menelusuri halaman yang relevan terlebih dahulu.

Kata kunci: focused crawler, web olahraga, naïve bayes, best first search