Abstrak

Pencarian dokumen di *Internet* memiliki karakteristik khusus yang harus dipertimbangkan yaitu *bandwidth* atau kecepatan akses yang terbatas serta waktu pencarian relatif lebih lambat daripada pencarian di *desktop*. Karena itu perlu dilakukan *indexing* pada proses *Information Retrieval* agar dapat mempercepat dan mempermudah pencarian. Makin banyak *term* yang terindeks akan makin membutuhkan waktu ekstra untuk mencari sebuah *term*. Sehingga diperlukan metode khusus untuk memangkas jumlah *term* dalam indeks. Salah satunya dengan melakukan ekstraksi dokumen menggunakan algoritma *Hidden Markov Model*. Metode yang dipakai dalam sistem ekstraksi ini adalah dengan melakukan pendekatan statistik dan *HMM Hedge* sebagai model HMM.

Metode yang digunakan tersebut mengeluarkan hasil: penggunaan *tagging* dapat memangkas waktu ekstraksi dan jumlah *term* terindeks secara signifikan, parameter *alpha* pada proses *decoding* mencapai nilai optimum pada 0,2 dan 0,3, ekstraksi dapat mengurangi waktu proses *indexing* dan jumlah term yang terindeks, serta jenis *corpus* mempengaruhi nilai akurasi dari sistem ekstraksi.

Kata kunci: *Hidden Markov Model, indexing, Information Retrieval, ekstraksi*