# ANALISIS PENGGUNAAN METODE HIDDEN MARKOV MODEL DALAM EKSTRAKSI KALIMAT UTAMA SUATU DOKUMEN PADA INFORMATION RETRIEVAL

# ANALYSIS OF HIDDEN MARKOV MODEL METHOD IMPLEMENTATION IN DOCUMENTS TOPIC SENTENCE EXTRACTION FOR INFORMATION RETRIEVAL

**Alfian Akbar Gozali[1], Warih Maharani[2], Yanuar Firdaus A.w.[3]**

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

**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**

**Abstract**

**Document searching in the Internet has special characteristic must be considered. Those are bandwith or limited access speed and searching time spending much longer rather than desktop searching. Therefore, it needs to use indexing at Information Retrieval process that can increase speed and simply searching activities. More indexing terms mean more extra time to searching any term. It needs special methods to cut the indexing terms. One of them is document extraction with Hidden Markov Model. The method using in this extraction system is statistical approach and HMM Hedge for the HMM Model.**

**That method outputs results: tagging can reduce extraction and nnumer of indexed terms signicantly, alpha parameter in decoding reach optimum value in 0,2 and 0,3, extraction can reduce indexing time and number of indexed terms, and corpus kinds influence extraction system accuracy.**

**Keywords : extraction, Hidden Markov Model, indexing, Information Retrieval**

# 1. Pendahuluan

## 1.1 Latar belakang

Dalam pencarian dokumen di *Internet* saat ini telah dapat dilakukan dengan menggunakan search engine seperti Google, Yahoo!, Altavista, dan masih banyak *search engine* lain. Pada umumnya *search engine* tersebut memiliki karakteristik yang sama, yaitu *user* harus memasukkan suatu kata kunci(*keyword*) sebagai *query* yang akan digunakan untuk mencari suatu dokumen. Dokumen yang muncul didasarkan atas algoritma masing-masing *search engine* yang akan mengenali dokumen dengan tingkat kemiripan tertentu dengan *query* yang telah dimasukkan. Namun pendekatan *search engine* yang umumnya dilakukan oleh beberapa *search engine* yang terkenal seperti Google, Yahoo!, Altavista, atau Bing belum ada yang melakukan pencocokan dokumen dengan berdasarkan atas konten dari suatu dokumen atau dengan kata lain mencari suatu dokumen yang mempunyai tingkat kemiripan tertentu terhadap dokumen kunci.

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*. Pencocokan dokumen yang bertujuan untuk mencari dokumen lain yang memiliki tingkat kemiripan tertentu akan menimbulkan konsekuensi banyaknya *Key Word* dan frasa yang akan terekstraksi. Pencarian dokumen yang melibatkan banyak *Key Word* dan frasa akan menyebabkan proses pencarian menjadi sangat lambat jika dilakukan dengan algoritma string matching biasa[15]. Oleh karena itu perlu dilakukan pendekatan statistik untuk memangkas waktu pencarian.

*Hidden Markov Model* (HMM) adalah sebuah model statistik dari sebuah sistem yang diasumsikan sebuah *Markov Process* dengan parameter yang tak diketahui, dan tantangannya adalah menentukan parameter-parameter tersembunyi (*hidden*) dari parameter-parameter yang dapat diamati[5]. Parameter-parameter yang ditentukan kemudian dapat digunakan untuk analisis yang lebih jauh, misalnya untuk aplikasi *Pattern Recognition*. Sebuah HMM dapat dianggap sebagai sebuah *Bayesian Network* dinamis yang paling sederhana[5].

Pada Model Markov umum, *state*-nya langsung dapat diamati, oleh karena itu probabilitas transisi *state* menjadi satu-satunya parameter. Di dalam Model Markov yang *hidden* (tersembunyi), *state*-nya tidak dapat diamati secara langsung, akan tetapi yang dapat diamati adalah variabel-variabel yang terpengaruh oleh *state*. Setiap *state* memiliki distribusi probabilitas atas *token-token* output yang mungkin muncul. Oleh karena itu rangkaian *token* yang dihasilkan oleh HMM memberikan sebagian informasi tentang sekuens *state-state*.

Atas karakteristik itulah *Hidden Markov Model* dapat digunakan untuk mengekstraksi kalimat utama dari suatu paragraf di dalam suatu dokumen[15]. Dalam HMM, bagian yang dapat diamati disebut *observed state* sedangkan bagian yang tersembunyi disebut *hidden state*. HMM memungkinkan pemodelan sistem yang mengandung *observed state* dan *hidden state* yang saling terkait. Pada kasus POS *tagging*, *observed state* adalah urutan kata sedangkan *hidden state* adalah urutan *tag*[15].

Pada *Information Retrieval*, proses *indexing* adalah proses yang pertama kali dilakukan. Pada proses ini, seluruh dokumen diambil kata-kata kuncinya untuk dimasukkan ke dalam *database*. Penyimpanan kata kunci ke dalam *database* ini dilakukan untuk mempermudah pencarian data ke depannya. Jika pencarian menjadi lebih mudah, imbasnya tentu waktu pencarian suatu dokumen akan lebih cepat. Ini adalah tujuan dari dilakukkannya proses *indexing*.

Kasus *indexing* pada *Information Retrieval* merupakan kasus ekstraksi kalimat utama yang khusus. Hal ini karena tidak hanya ekstraksi kalimat utama yang dilakukan namun juga ada penanganan pada proses pengindeksan untuk dimasukkan ke dalam database. Pengindeksan dilakukan untuk mempercepat waktu pencarian selanjutnya. Proses pengindeksan yang akan dilakukan terdiri dari 2 tahap, yaitu *Key Topic Sentence Extracting* (Peringkasan Kalimat Utama Kunci) dan *Key Word Indexing* (Pengindeksan Kata Kunci). Diharapkan dengan adalanya pengindeksan dua tahap ini akan menghasilkan waktu pencarian yang lebih berkurang daripada proses pencarian umum yang hanya menggunakan kata kunci.

Dalam tugas akhir ini akan dilakukan analisis terhadap proses pengekstraksian kalimat utama dalam suatu artikel dengan metode *Hidden Markov Model* dan performansi dari proses pengindeksan kalimat utama hasil ekstraksi ke dalam *Key Topic Sentence Extracting* untuk kemudian akan dipecah lagi menjadi *Key Word Indexing*. Metode ekstraksi yang akan digunakan adalah algoritma *Hidden Markov Model* yang diadaptasi dari model *testing translation* dan HMM-Hedge. HMM-Hedge adalah struktur *Hidden Markov Model* yang ditujukan khusus untuk meng-*generate* judul sebuah berita. Algoritma Viterbi digunakan untuk mencari susunan kata yang paling optimal[15].

## 1.2 Perumusan masalah

Masalah yang dirumuskan dalam Tugas Akhir ini berdasar dari latar belakang yang telah dijelaskan sebelumnya, yaitu:

1. Bagaimana cara melakukan *cleaning* pada dokumen elektronik dari *Internet* untuk mengekstrak artikel utamanya?
2. Bagaimana cara kerja Algoritma *Hidden Markov Model* dalam melakukan ekstraksi kalimat utama dari suatu artikel?
3. Bagaimana kinerja Algoritma *Hidden Markov Model* dalam melakukan ekstraksi dokumen berdasarkan atas kecepatan dan akurasi (tingkat kebenaran hasil ekstraksi) yang didasarkan atas perubahan parameter pada Algoritma *Hidden Markov Model*?
4. Bagaimana perbandingan hasil *indexing* biasa (*parsing* langsung dari konten dokumen) dengan hasil *indexing* dengan ekstraksi kalimat utama pada besar jumlah *term* yang tersimpan pada database *indexing*.
5. Apakah ada pengaruh jenis dokumen terhadap kinerja Algoritma *Hidden Markov Model*?

## 1.3 Batasan masalah

Batasan masalah pada Tugas Akhir ini adalah :

1. Hanya menangani masalah *preprocessing* pada *Information Retrieval* khususnya masalah *indexing* .
2. Dokumen elektronik yang dipakai sebagai *training set* adalah halaman situs yang sebagian besar kontennya merupakan artikel utama yang diwakili oleh penggunaan simbol <p> dalam *body* paragrafnya dan dalam bahasa Inggris.
3. Hanya menangani masalah *cleaning*, *extracting* artikel dan *indexing* dokumen ke dalam *database*.
4. Yang menjadi parameter uji untuk melakukan analisis hasil adalah kecepatan *indexing* dokumen elektronik yang didasarkan atas perubahan parameter pada Algoritma *Hidden Markov Model*, tingkat kebenaran semantik hasil ekstraksi dengan persamaan ROGUE-2, dan jumlah *term* pada database *indexing* pada *indexing* sebelum dan setelah dengan ekstraksi.
5. *Text preprocessing* diimplementasikan dalam Tugas Akhir ini namun tidak menjadi fokus permasalahan dalam Tugas Akhir ini.

## 1.4 Tujuan

Tujuan yang ingin dicapai dalam pengerjaan Tugas Akhir ini adalah sebagai berikut :

1. Menganalisis performansi Algoritma *Hidden Markov Model* dalam mengekstraksi kalimat utama pada artikel.
2. Menganalisis pengaruh parameter-parameter yang terdapat didalam Algoritma *Hidden Markov Model* terhadap performansi dari *Key Topic Sentence* yang didapatkan.
3. Menganalisis hasil hasil *indexing* biasa (*parsing* langsung dari konten dokumen) dengan hasil *indexing* dengan ekstraksi kalimat utama pada besar jumlah *term* yang tersimpan pada database *indexing*.
4. Mengetahui pengaruh jenis dokumen terhadap kinerja Algoritma *Hidden Markov*

## 1.5 Metodologi penyelesaian masalah

Langkah-langkah yang dilakukan dalam penyelesaian Tugas Akhir ini adalah:

1. **Pengumpulan data dan studi literatur**
   Pengumpulan data meliputi pengumpulan data berupa halaman-halaman situs dari *Internet* yang memenuhi batasan masalah. Sedangkan studi literatur meliputi Algoritma *Hidden Markov Model*, Algortima *Viterbi*, dan *Information Retrieval* khususnya untuk proses *indexing* yang sumbernya berasal dari buku dan artikel dari *Internet*.

2. **Perancangan Sistem**
   Pada tahap ini dilakukan perancangan sistem untuk pencarian dokumen dengan tingkat kemiripan tertentu dari dokumen sumber. Digunakan metode ekstraksi kalimat utama dengan *Hidden Markov Model* dan Viterbi pada *preprocessing*nya. Namun analisis hanya dilakukan pada proses *indexing*nya dan tidak akan dibahas lebih lanjut tentang *postprocessing*-nya.
   Jika di-*breakdown* perancangan sistemnya adalah sebagai berikut:

a. Menerima input data berupa dokumen artikel berita berbahasa inggris yang bersifat *offline* dalam bentuk *file* *.htm/*.html
b. Melakukan proses *evaluation*/*training*
c. Melakukan proses ekstraksi dokumen menjadi *free-text* berekstensi *.txt.
d. Melakukan proses ekstraksi dengan *preprocessing* tertentu
e. Melakukan pengukuran tingkat akurasi hasil ekstraksi
f. Melakukan proses *indexing*

**3. Implementasi Sistem**

Dalam implementasi ini, rancangan sistem yang dibuat akan diimplementasikan ke dalam bahasa pemrograman *server-side* PHP 5.2.2 dengan *database* MySQL 5.0.41.

**4. Pengujian Sistem**

Pengujian sistem dilakukan untuk mengetahui apakah sistem yang dibangun sudah tepat dalam mencapai tujuan yang telah dibuat.
Pengujian dilakukan dengan:

a. Menguji kecepatan ekstraksi dan akurasi dalam penggunaan *tag* {NAMA} dan {NUMERIK} dalam *preprocessing*
b. Menguji akurasi sistem dengan mengganti nilai parameter α dalam proses ekstraksi
c. Menguji pengaruh proses ekstraksi pada kecepatan dan jumlah *term* yang terindeks
d. Menguji pengaruh beberapa jenis *corpus* pada kecepatan dan akurasi ekstraksi

**5. Analisis Hasil**

Analisis hasil dilakukan pada hasil-hasil yang telah diperoleh dari pengujian sistem dan berdasarkan rumusan masalah yang telah ditentukan.

**6. Penyusunan Laporan Tugas Akhir**

Langkah terakhir, yaitu pembuatan laporan tugas akhir, yang meliputi hasil analisa dan langkah-langkah yang lainnya yang telah dilakukan. Pembuatan laporan berfungsi sebagai dokumentasi dari apa yang selama ini telah dikerjakan untuk menyelesaian sistem.

# 5. Kesimpulan dan Saran

## 5.1 Kesimpulan

Dari hasil pengujian dan analisis yang telah dilakukan pada bab sebelumnya dalam Tugas Akhir ini, maka didapatkan kesimpulan:

1. Penggunaan *preprocessing* berupa *tag* untuk entitas nama dan numerik dapat menurunkan waktu eksekusi dan jumlah *term* yang perindeks secara signifikan namun jika tidak diimbangi dengan jumlah *training* dataset yang besar, akan menurunkan tingkat akurasi sistem ke level yang tidak dapat ditoleransi.
2. Nilai parameter *alpha* untuk parameter *decoding* berpengaruh dalam menentukan akurasi dari hasil ekstraksi dan nilai *alpha* yang paling optimum untuk jenis *corpus* yang dibahas dalam tugas akhir ini adalah 0,2 dan 0,3.
3. Penggunaan sistem ekstraksi dokumen dapat mereduksi waktu eksekusi dan jumlah *term* dengan cukup signifikan.
4. Spesifikasi *corpus* yang digunakan sebagai *training* dataset sangat berpengaruh terhadap tingkat akurasi *testing* dataset yang menjadi input sistem. Semakin jauh karakteristik *testing* dataset dari *training* dataset, akurasinya akan semakin rendah. Sebaliknya, semakin dekat karakteristik *testing* dataset dengan *training* dataset, akurasinya akan semakin tinggi.

## 5.2 Saran

1. Performansi *Hidden Markov Model* dalam pemodelan ekstraksi dokumen model statistik dapat dibandingkan dengan model bahasa yang lebih taat tata bahasa
2. Karena sifatnya yang tidak terpengaruh adanya tata bahasa, sistem ini dapat dicoba dalam bahasa lain di samping bahasa Inggris
3. Dilakukan proses hilir *Information Retrieval* agar dapat diketahui tingkat kepuasan pengguna terhadap sistem yang telah dibuat
4. Perlu dilakukan riset lebih lanjut untuk identifikasi *term* yang layak dikenai proses *tagging*

# Daftar Pustaka


[1] B. H. Juang; L. R. Rabiner. *Hidden Markov Models for Speech Recognition*. Technometrics, Vol. 33, No. 3. (Aug. 1991):251-272

[2] Banko, M. Mittal, V. O. Witbrock, M. J., "*Headline Generation Based on Testing Translation*", Annual Meeting- Association For Computational Linguistics,Vol 38; Part 1, 2000: 318 – 325

[3] Broder, Andrei. *A Taxonomy of Web Search*. IBM Research, 2002

[4] Doran, W., Stokes, N., Newman, E., Dunnion, J., Carthy,J., Toolan, F., "*News Story Gising at University College Dublin*", Document Understanding Conference, DUC 2004.

[5] *Hidden Markov Model* – Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/*Hidden*_Markov_*Model* diunduh 20 Oktober 2009

[6] J. Beal, Matthew; Ghahramani, Zoubin; Edward Rasmussen, Carl. *The Infinite Hidden Markov Model*. Gatsby Computational Neuroscience Unit University College London, 2008.

[7] Jing H, *Sentence Reduction For Automatic Text Summarization*-Proceedings of the 6th Applied Natural Language Processing, 2000.

[8] Knight, K., and Marcu, D, *Statistics Based Summarization: Step One: Sentence Compression*. Proceedings of the 17th National Conference of theAmerican Association for Artificial Intelligence AAAI2000, Austin, Texas, July 30-August 3, 2000.

[9] L. Rabiner. *A tutorial on Hidden Markov Models and selected applications in speech recognition*. Proc. of IEEE, 77(2):257-286, 1989.

[10] C.Y. Lin, *Recall-Oriented Understudy for Gisting Evaluation*, 2003.

[11] Marlow, Kit. 2003. *Information Retrieval Methods*, http://www.seas.upenn.edu/~zives/03s/cis650/ir.pdf diunduh 20 Oktober 2009.

[12] M Le Nguyen, S Horiguchi, A Shimazu, BT Ho, *Example-Based sentence Reduction* using the *Hidden Markov Model*, ACM Transactions on Asian Language Information *Process*ing, 2004.

[13] SF Chen, J Goodman, *An Empirical Study of Smoothing Techniques for Language Modeling*, Proceedings of the 34th annual meeting on Association for Computational Linguistik, 1996

[14] Waiyamai, Kitsana. *Introduction to Text Mining*. Dept of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok, Thailand.

[15] Wibisono, Yudi, *Penggunaan Hidden Markov Model untuk Kompresi Kalimat*, Tesis Institut Teknologi Bandung, 2008.

[16] Witbrock, M.J., Mittal, V.O, "*Ultra-Summarization: A Testing Approach to Generating Highly Condensed Non-Extractive Summaries (poster abstract)*",1999.

[17] *Research and Development in Information Retrieval*, pages 315-316

[18] Wright, Jan 1998. *An Overview Of Indexing Methods* http://www.stcsig.org/idx/articles/methods.pdf diunduh 20 Oktober 2009.

[19] Zajik D, Dorr B, *Automatic Headline Generation for Newspaper Stories*, 2002.

[20] Zajic,D.etal., *Multi-candidate Reduction*:*Sentence Compression as a tool*, *Information Process*ing *and Management*, 2007.