

BAB I

PENDAHULUAN

1.1 Latar Belakang

Kearsipan adalah penempatan kertas-kertas dalam tempat-tempat penyimpanan yang baik menurut aturan yang telah ditentukan terlebih dahulu sedemikian rupa sehingga setiap kertas (surat) apabila diperlukan dapat diketemukan kembali dengan mudah dan cepat (Moekijat, 2002). Masalah dalam kearsipan bersifat dinamis, karena arsip akan terus bertambah seiring dengan perkembangan suatu organisasi. Apabila arsip yang terus bertambah tersebut tidak diikuti dengan sistem pencarian yang baik, arsip akan memerlukan waktu yang lama untuk ditemukan jika sewaktu-waktu diperlukan. Oleh karena itu dibutuhkan sistem pencarian dokumen yang akan mencari dokumen sesuai kebutuhan user.

Algoritma Aho Corasick adalah sebuah algoritma pencarian string yang diperkenalkan oleh Alfred V. Aho dan Margaret J. Corasick. Algoritma ini merupakan jenis algoritma pencocokan kata yang mencari lokasi kata yang dicari dari sekumpulan elemen *string* (kamus) dalam teks (Aho & Corasick, 1975). Pencocokan kata dilakukan secara bersamaan, sehingga pencocokan dapat ditemukan selama beberapa kali untuk setiap substring yang terdapat pada kata yang dicari. Namun algoritma Aho-Corasick dirancang untuk set pola ukuran sedang, dan jika untuk set pola ukuran besar, maka waktu pencarian juga akan bertambah lama, untuk itu maka digunakan algoritma AC-NFA (Aho Corasick with Non-deterministic Finite Automation). Definisi formal dari nondeterministic finite automaton (NFA) mirip dengan deterministic finite automaton. Keduanya memiliki states, input abjad, fungsi transisi, states awal, dan kumpulan states yang dapat diterima. Namun, mereka berbeda dalam satu cara penting: dalam jenis fungsi transisi. Dalam DFA, fungsi transisi mengambil states dan input symbol kemudian menghasilkan states selanjutnya. Dalam NFA, fungsi transisi mengambil states dan input symbol atau string kosong kemudian menghasilkan kumpulan kemungkinan states berikutnya (Sipser, 2012). Berbeda dengan Aho-Corasick, NFA tidak menggunakan banyak *memory* dalam proses penggunaannya.

Algoritma AC-NFA merupakan algoritma yang memungkinkan pemeriksaan beberapa karakter secara paralel dengan transisi *multicharacter*. Pada penelitian Chen (Chen & Wang, 2013) AC-trie dikonversi ke AC-NFA dengan memperbolehkan aktivasi bersamaan dari beberapa *states* dan kemudian mengubah AC-NFA menjadi *k*-karakter AC-NFA menggunakan algoritma dengan operasi penggabungan dan transisi asisten. Berdasarkan uraian tersebut, maka diajukan penelitian tugas akhir dengan judul **“PENCARIAN DOKUMEN ARSIP DENGAN METODE MULTICHARACTER AC-NFA BERBASIS GPU”**

1.2 Rumusan Masalah

Berdasarkan latar belakang diatas, maka yang menjadi permasalahan pada tugas akhir ini adalah membandingkan sistem pencarian Aho-Corasick dengan sistem pencarian AC-NFA.

1.3 Tujuan

Tujuan dari penulisan tugas akhir ini adalah mengembangkan aplikasi yang dapat mempercepat proses pencarian dokumen yang telah tersip dengan menggunakan metode Multicharacter AC-NFA Berbasis GPU.

1.4 Manfaat

Manfaat dari penulisan tugas akhir ini adalah untuk mempercepat proses pencarian dokumen yang telah diarsip.

1.5 Batasan Masalah

Batasan masalah yang ditetapkan dalam menyusun tugas akhir ini adalah :

1. Hardware yang digunakan berbasis Graphics Processing Unit (GPU) dengan tipe *GeForce GT 650M (CUDA)*.
2. Mengubah surat yang telah dipindai menjadi teks dengan menggunakan library Tesseract-OCR.
3. Penginputan kata kunci hanya menggunakan karakter dengan ASCII 0-122.

4. Maksimum panjang kata kunci yang dapat diinput pada bagian pencarian (Tahun 4 karakter dan Search 13 karakter). Minimum panjang kata kunci yang dapat diinput adalah 1 karakter.
5. Surat yang dipindai merupakan *file* .jpg.
6. Surat yang dipindai memiliki ketentuan:
 - a. Ukuran kertas A4.
 - b. Font *style* yang dapat dibaca adalah Times New Roman, Arial.
 - c. Font *size* yang dapat dibaca adalah dari *range* 11-60.
 - d. *Header* (kop surat) berada di bagian atas.
 - e. Posisi surat tegak (*portrait*).

1.6 Metodologi Pengembangan Sistem

Metodologi yang digunakan pada sistem ini adalah metode *System Development Life Cycle* (SDLC). SDLC adalah suatu pendekatan yang memiliki tahap atau bertahap untuk melakukan analisa dan membangun suatu rancangan sistem dengan menggunakan siklus yang lebih spesifik terhadap kegiatan pengguna (Kendall & Kendall, 2006). Tahapan - tahapan yang dilakukan dalam penyelesaian tugas akhir ini adalah sebagai berikut:

1. Perencanaan : Identifying Problems, Opportunities, Objectives

Masalah yang dapat diuraikan adalah bagaimana arsip dapat disimpan secara komputerisasi dan ketika arsip diperlukan, pengguna dapat mencari arsip tanpa membutuhkan waktu yang lama.

2. Analisis Sistem

a. Determining Information Requirements

Tahap ini melakukan kajian literatur dari buku, jurnal dan data – data dari internet. Selain itu, meminta izin kepada pihak Sekolah Dharma Bakti untuk mengumpulkan data berupa dokumen / surat dari tata usaha Sekolah Dharma Bakti yang akan dijadikan sebagai *data set*.

b. Analyzing System Needs

Tahapan selanjutnya adalah menganalisis sistem yaitu analisis proses dan analisis kebutuhan dari sistem secara fungsional dan non fungsional. Kebutuhan fungsional digambarkan dengan *usecase diagram*

sedangkan, untuk kebutuhan non fungsional digambarkan dengan kerangka PIECES. Untuk analisis proses dijelaskan cara kerja dari metode Multicharacter AC-NFA yang digunakan dengan menggunakan *flowchart*.

3. Desain : Designing the Recommended System

Tahap ini merancang tampilan dasar (*mockup*) menggunakan *Balsamiq* dan merancang desain basis data menggunakan *Entity Relationship Diagram* (ERD).

4. Implementasi dan Pengujian

a. Developing and Documenting Software

Tahap ini dilakukan untuk mengimplementasikan rancangan tampilan *user interface* menggunakan bahasa pemrograman *Visual C#* sedangkan untuk rancangan database menggunakan MySQL, untuk mengenali teks pada gambar menggunakan OCR dengan library *tesseract*, untuk melakukan pencarian digunakan pemrograman secara paralel dengan CUDA.

b. Testing and Maintaining the System

Pada tahapan pengujian, dilakukan pengujian dengan *data set* dari surat yang telah dikumpulkan, yaitu surat internal sebanyak 137 lembar dan surat eksternal sebanyak 112 lembar. Selain itu, dilakukan pengujian hasil perbandingan waktu antara algoritma Aho-Corasick dengan algoritma AC-NFA.

c. Implementing and Evaluating the System

Tahap ini dilakukan untuk mengimplementasi algoritma AC-NFA dengan pendekatan secara multicharacter pada sistem yang telah dibangun. Selain itu, menampilkan *chart* hasil perbandingan antara pencarian Aho-Corasick dengan pencarian AC-NFA menggunakan Microsoft Excel.