EKSTRAKSI INFORMASI PADA DOKUMEN SKRIPSI BERBASIS ATURAN

Dimas Mustaqwa[1] , Nelly Indriani Widiastuti[2]

[1,2] Teknik Informatika – Universitas Komputer Indonesia
Jalan Dipatiukur 112-116 Bandung
E-mail : [1] dimasmustaqwa@gmail.com, [2]nelly.indriani@email.unikom.ac.id

ABSTRAK

Pencarian suatu informasi dapat dilakukan salah satunya dengan cara membaca suatu dokumen. Dokumen yang tersedia dalam bentuk softcopy pada umumnya menggunakan media penyimpanan dengan format pdf atau doc. Di perpustakaan penyimpan suatu dokumen harus diberikan suatu informasi atau identitas. Pustakawan memasukan suatu identitas pada suatu dokumen dengan cara mengisi data-data yang diperlukan kedalam sistem. Kelemahan dari cara ini salah satunya adalah adalah membutuhkan waktu relatif lama jika jumlah dokumen banyak yang harus disimpan, masalah lain yang mungkin timbul adalah kesalahan pengetikan identitas dokumen. Kelemahan tersebut dapat ditangani dengan cara mengisi identitas dokumen secara otomatis, salah satu caranya dengan mengektraksi dokumen. Ekstraksi informasi menggunakan *rule based* merupakan suatu metode yang menggunakan aturan berdasarkan fakta dari data yang dianalisis. Untuk pengujian akurasi mengggunakan 50 pada dokumen laporan skripsi cover, abstrak, dan abstract menunjukan hasil yang cukup baik, yang berarti ekstraksi informasi pada penelitian ini bisa digunakan untuk mengekstrak identitas data skripsi yang diinginkan.
Kata Kunci : Ekstraksi Informasi, *rule based*, teks, skripsi.

1. PENDAHULUAN

Pencarian suatu informasi salah satunya dengan cara membaca suatu dokumen. Dokumen yang tersedia dalam bentuk softcopy umumnya menggunakan media penyimpanan berformat pdf atau doc. Pencarian pada suatu dokumen dapat dilakukan jika dokumennya sudah tersimpan.

Di perpustakaan penyimpan suatu dokumen harus diberikan suatu informasi atau identitas. Pustakawan memasukan suatu identitas pada suatu dokumen dengan cara mengisi data-data yang diperlukan kedalam sistem. Kelemahan dari cara ini salah satunya adalah lama jika jumlah dokumen banyak yang harus disimpan, masalah lain yang mungkin timbul adalah kesalahan pengetikan identitas dokumen. Kelemahan tersebut dapat ditangani dengan cara mengisi identitas dokumen secara otomatis, salah satu caranya dengan mengektraksi dokumen. Ekstraksi Informasi adalah pengambilan fakta dan informasi terstruktur dari isi koleksi teks yang besar. Pengertian fakta disini adalah beragam entitas yang diperhitungkan. Secara singkat ekstraksi informasi adalah sebuah proses mendapatkan fakta-fakta terstruktur dari data yang tersedia [1]. Penelitian sebelumnya oleh Ramón yaitu menggunakan ekstraksi informasi untuk mengklasifikasikan iklan koran [2]

Pengekstraksian dokumen dapat dilakukan menggunakan suatu metode. Salah satu metode mengekstraksi adalah *rule based*. Sistem berbasis aturan (*Rule Based System*) adalah suatu program komputer yang memproses informasi yang terdapat di dalam working memory dengan sekumpulan aturan yang terdapat di dalam basis pengetahuan menggunakan mesin inferensi untuk menghasilkan informasi baru [3]. Metode *rule based* dapat digunakan jika suatu dokumen tersebut merupakan dokumen yang terstruktur. Adapun data yang terstruktur adalah data yang telah terorganisir sehingga mudah dalam suatu pencarian data sedangkan tidak terstruktur adalah data yang belum terorganisir [4]. Informasi dokumen yang dapat diperoleh dengan cara mencari struktur suatu dokumen salah satunya adalah skripsi. Skripsi didefinisikan sebagai penulisan karya ilmiah berisi hasil penelitian menyeluruh yang disusun secara sistematis berdasarkan ketentuan metode penelitian ilmIah. Penulisan skripsi ini dimaksudkan sebagai pelatihan bagi mahasiswa untuk menuangkan gagasannya dalam bentuk karya ilmiah [5]. Isi struktur yang diambil dari skripsi yaitu judul skripsi, jenis skripsi, nama penulis, nim, program studi, fakultas, universitas, isi abstrak, dan kata kunci pada abstrak. Pengertian informasi terstuktur ialah suatu kalimat atau teks yang dapat dibagi ke dalam

beberapa kategori seperti topik, fokus, komentar, latar belakang, dan membandingkan informasi lama atau baru [6].

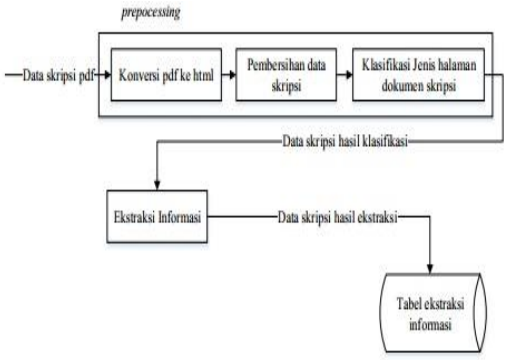
Salah satu penelitian tentang ekstraksi informasi pada rule base berjudul ”ekstraksi informasi dengan metode rule – based untuk evaluasi pemahaman fisika kinematika” ekstraksi informasi yang menggunakan *rule based* mendapatkan inti dari pertanyaan yang diajukan dan menampilkan jawaban yang sesuai dengan pertanyaan yang diajukan. Penggunaan *rule based* pada data adalah pencarian kata kunci pada latihan soal. Kata kunci mencari pertanyaan, angka, besaran, satuan, dan rumus pada pelajaran fisika. Dari kata kunci yang ditemukan akan
dilakukan menyelesaian atau memberikan jawaban yang sesuai dari data yang didapatkan. Hasil akurasi yang diperoleh sebesar 90.6%-95.4% pada penelitan evaluasi pemahaman fisika kinematika [7]. Penelitian lainnya yang berjudul “algoritma ekstraksi informasi berbasis aturan” menggunakan ekstraksi informasi pada dokumen Laporan Hasil Pemeriksaan (LHP) atas Laporan Keuangan Pemerintah Daerah (LKPD) yang hasil ekstraksi di kelompokkan kedalam beberapa klasifikasi. Hasil akurasi pada dokumen LHP LKDP yaitu 89,77% dan 98,27% [8] dan mengekstrak teks untuk menghasilkan kata kunci [9]

Berdasarkan latar belakang tersebut, maka dalam penelitian ini akan dilakukan ekstraksi informasi untuk mendapatkan struktur informasi pada dokumen skripsi dan melakukan pemberian identitas pada dokumen secara otomatis. Proses pengekstraksian informasi akan menggunakan metode *rule based*. Hasil pengekstraksian informasi yang dilakukan adalah mengelompokan isi dokumen ke dalam beberapa klasifikasi. Penelitian ini bertujuan mengetahui judul skripsi, jenis skripsi, nama penulis, nim, program studi, fakultas, universitas, isi abstrak, dan kata kunci pada abstrak. Penelitian data akan menganalisa dan membangun sistem ekstraksi informasi menggunakan *rule based* untuk mendapatkan identitas dari data skripsi yang akan ekstraksi.

2. ISI PENELITIAN

2.1 Metode Ekstraksi Informasi

Analisis sistem dapat didefinisikan sebagai penguraian dari suatu sistem yang utuh kedalam bagian-bagian komponennya dengan maksud untuk mengidentifikasikan kebutuhan-kebutuhan yang diperlukan agar dapat dibangun sebuah aplikasi untuk mengetahui presentasi akurasi analisis dari metode yang digunakan. Analisis sistem tentang ekstraksi informasi dengan metode rule based untuk mengidentifikasi data skripsi yang tidak terstruktur menjadi terstruktur di bagi menjadi beberapa bagian yang dapat dilihat pada gambar 1 berikut



Gambar 1. Sistem Ekstrasi Data Skripsi

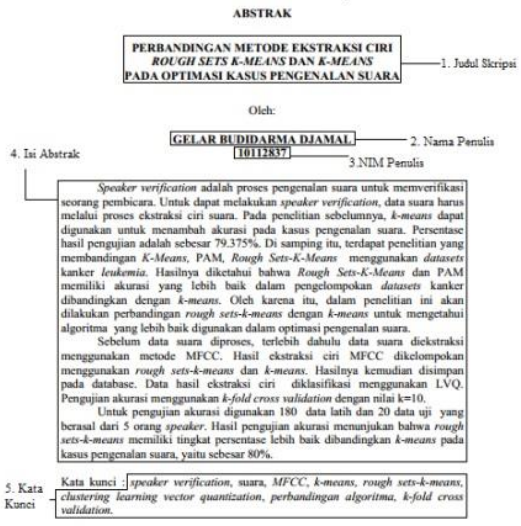
Gambaran alur sistem ekstraksi dokumen skripsi akan dijelaskan sebagai berikut ini.

1. Proses yang pertama adalah konversi file pdf dokumen skripsi menjadi file berformat html.
2. Selanjutnya hasil proses konversi menghasilkan file yang perlu dibersihkan dari simbol atau karakter yang tidak relevan.
3. Tahap terakhir adalah menyusun aturan berdasarkan seluruh dokumen skripsi yang ada.

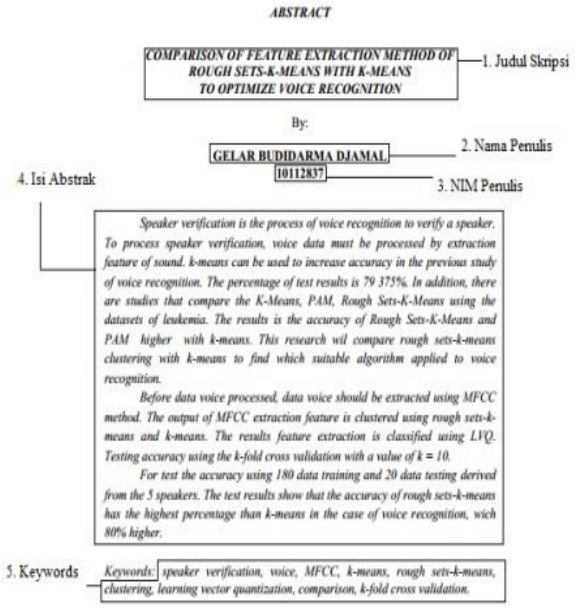
Sebelum melakukan analisis sistem dilakukan analisis bertujuan membuat rule based pada data skripsi cover, abstrak, dan abstract

2.2 Analisis data masukan

Pada penelitian ini menggunakan data skripsi yang berformat .pdf. Analisis berfokus pada pencarian kata kunci dan aturan yang akan digunakan untuk mengekstraksi informasi data skripsi. Data skripsi yang akan dianalisis yaitu data cover, abstrak bahasa indonesia dan abstrak bahasa inggris. Data skripsi cover, abstrak, dan abstract dapat dilihat pada gambar 2, gambar 3, dan gambar 4.



Gambar 2. Data masukan abstrak



Gambar 3. Data Masukan Abstract

Gambar 4. Data Masukan Cover

Sampel data skripsi akan diidentifikasi data skripsi. Hasil kata kunci dan aturan ekstraksi data skripsi yang didapatkan dari identifikasi data dapat dilihat pada tabel 1 sebagai berikut.

Tabel 1. Hasil Kata Kunci dan Aturan Ekstraksi Data Skripsi

Identifikasi	Kata kunci	Keterangan
Data Cover Skripsi		
Judul skripsi	-	Mengambil judul skripsi dari awal kata hingga menemukan kata kunci jenis skripsi yang terakhir dari data skripsi
Jenis Skipsi	Skripsi, Tesis, Disertasi	Mencari kata “Skripsi, Tesis, Disertasi” yang terakhir dari data skripsi
Nama Penulis	Sarjana	Mencari kata “Sarjana” yang terakhir dari data skripsi mengambil kata setelah kata “Sarjana” hingga menemukan angka atau jika terdapat posisi kata “Teknik dan Ilmu Komputer” yang berbeda mengambil kata setelah kata “Teknik dan Ilmu Komputer” pada posisi yang pertama
NIM		Angka setelah nama penulis sebanyak 8 angka
Program Studi	Program	Mencari kata “Program” yang terakhir dari data skripsi mengambil kata setelah kata “Program” hingga menemukan kata kunci fakultas
Fakultas	Fakultas	Mencari kata “Fakultas” yang terakhir dari data skripsi mengambil kata setelah kata “Fakultas” hingga menemukan kata kunci Universitas
Kampus	Universitas	Mencari kata “Universitas” yang terakhir dari

Identifikasi	Kata kunci	Keterangan
		data skripsi mengambil kata setelah kata "Universitas" hingga menemukan angka
Data Abstrak dan Abstract		
Judul Indonesia atau Judul Inggris	-	Setelah Kata "Abstrak atau Kata Abstract" sampai menemukan awal kata kunci Nama penulis
Nama Penulis Oleh, By Mencari kata	"Oleh, By"	yang terakhir dari data skripsi mengambil kata setelah kata "Oleh, By" hingga menemukan angka
NIM	-	Angka setelah nama penulis
Isi Abstrak Indonesia atau Bahasa Inggris	-	Setelah Nim dan sampai menemukan kata Kata Kunci Indonesia atau Bahasa Inggris yang terakhir dari data skripsi
Kata Kunci Indonesia atau Inggris	Kata Kunci, Keywords, Keyword	Mencari kata "Kata Kunci, Keywords, Keyword" yang terakhir dari data skripsi mengambil setelah kata kunci hingga akhir kata data srkipsi

2.3 Preprocessing

Berdasarkan analisis data yang dilakukan perlu dilakukannya tahap preprocessing. Tahap preprocessing yang akan dilakukan yaitu konversi pdf ke html, pembersihan data skripsi, klasifikasi jenis dokumen skripsi.

2.3.1 Konversi pdf ke html

Konversi pdf ke html adalah proses mengubah data yang berformatan pdf menjadi html. Hal ini dilakukan supaya data dapat dianalisis. Sebagai gambaran konversi pdf ke html ditampilkan pada gambar 5 berikut ini.

Sesudah Konversi
i   ABSTRAK   PERBANDINGAN METODE   EK STRAKSI CIRI   ROUGH SETS K - MEANS   DAN K - MEANS   PADA OPTIMASI KASUS PENGENALAN SUARA   Oleh: GELAR BUDIDARMA DJAMAL   10112837   Speaker verification   adalah proses pengenalan suara untuk memverifikasi seorang pembicara. Untuk dapat   melakukan speaker verification , data suara harus mela lui proses ekstraksi ciri suara . Pada   penelitian sebelumnya ,   k - means   dapat digunakan untuk menambah akurasi pada kasus pengenalan suara . Persentase   h asil pengujian adalah Sesudah Konversi i   ABSTRAK   PERBANDINGAN METODE   EK STRAKSI CIRI   ROUGH SETS K - MEANS   DAN K - MEANS   PADA OPTIMASI KASUS PENGENALAN SUARA   Oleh: GELAR BUDIDARMA DJAMAL   10112837   Speaker verification   adalah proses pengenalan suara untuk memverifikasi seorang pembicara. Untuk dapat   melakukan speaker verification , data suara harus mela lui proses ekstraksi ciri suara . Pada   penelitian sebelumnya ,   k - means   dapat digunakan untuk menambah akurasi pada kasus pengenalan suara . Persentase   h asil pengujian adalah

Gambar 5. Hasil Konversi data skripsi abstrak

2.3.2 Pembersihan data skripsi

Hasil dari konversi pdf ke txt memberikan data yang masih tidak sesuai. paragraf baru muncul walaupun kata-kata tersebut masih dalam satu kalimat yang sama. Beberapa paragraf baru muncul setelah akhir kalimat. Oleh karena itu perlu dilakukan pembersihan data skripsi. Berikut langkah-langkah tahapan pembersihan data skripsi.

1. Jika terdapat 3 line yang kosong maka buat menjadi 1 line kosong.

2. Jika line baru tidak kosong maka line baru digabung dengan line sebelumnya.

Hasil pembersihan data skripsi dapat dilihat pada gambar 6 berikut ini.

Sesudah Pembersihan
i ABSTRAK PERBANDINGAN METODE EKSTRAKSI CIRI ROUGH SETS KMEANS DAN K-MEANS PADA OPTIMASI KASUS PENGENALAN SUARA Oleh: GELAR BUDIDARMA DJAMAL 10112837 Speaker verification adalah proses pengenalan suara untuk memverifikasi seorang pembicara. Untuk dapat melakukan speaker verification, data suara harus melalui proses ekstraksi ciri suara. Pada penelitian sebelumnya, k-means dapat digunakan untuk menambah akurasi pada kasus pengenalan suara. Persentase hasil pengujian adalah sebesar 79.375%. Di samping itu, terdapat penelitian yang membandingan K-Means, PAM, Rough Sets-K-Means menggunakan datasets kanker leukemia. Hasilnya diketahui bahwa Rough SetsK-Means dan PAM memiliki akurasi yang

Gambar 6. Hasil pembersihan data skripsi

2.3.3 Klasifikasi Jenis Halaman Dokumen Skripsi

Klasifikasi jenis halaman dokumen skripsi pada preprocessing merupakan klasifikasi data jenis skripsi yang akan diekstraksi apakah cover atau abstrak setelah proses pembersihan data skripsi. Tabel klasifikasi jenis halaman dokumen skripsi dapat dilihat pada tabel 2 berikut.

Tabel 2. Klasifikasi jenis halaman dokumen skripsi

Klasifikasi	Status data	Keterangan
1	Data cover skripsi	Mencari kata program studi, fakultas dan universitas pada data skripsi
2	Data abstrak skripsi	Mencari kata abstrak, atau abstract pada data skrips
3	Bukan data skripsi cover atau abstrak.	Tidak ditemukan kata pada pencarian yang dilakukan pada klasifikasi

2.4 Ekstraksi Dokumen Skripsi

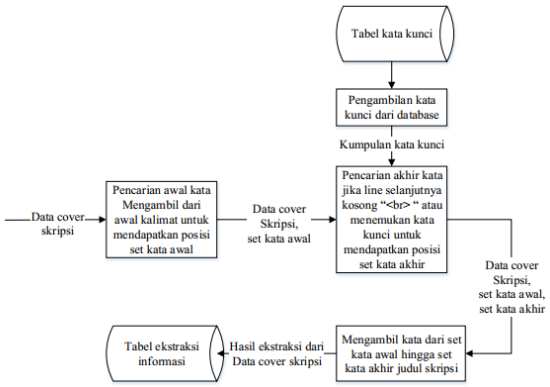
Ekstraksi Informasi akan dilakukan menggunakan rule base berdasarkan kata kunci dan aturan yang sudah dianalsis. Tahapan yang akan dilakukan pada ekstraksi informasi yaitu ekstraksi data cover skripsi dan ekstraksi data abstrak skripsi.

2.4.1 Ekstraksi Informasi Data Cover Skripsi

Ekstraksi informasi data cover menggunakan rule based. Data yang dicari adalah sebagai berikut:

a. Judul

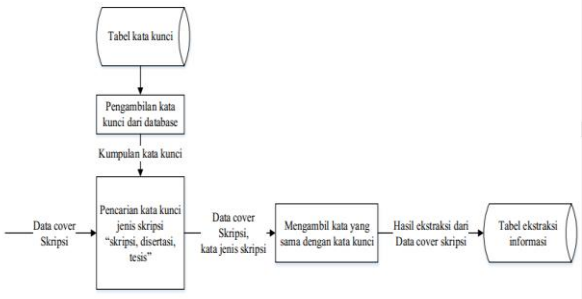
Judul merupakan bagain skripsi yang pertama kali dicari, berikut ini adalah gambar 7 yang mengilustrasikan proses mendeteksi judul dalam dokumen cover skripsi. Judul umumnya terdapat pada awal baris lembar cover. Sehingga pendeteksian dilakukan dengan mencari baris awal hingga kata kunci.



Gambar 7. Proses ekstraksi judul

b. Jenis

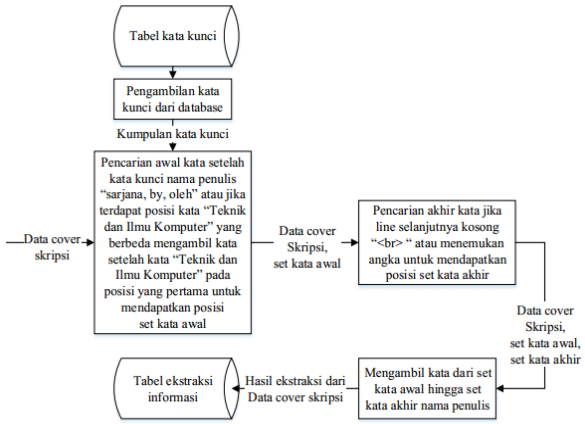
Pada bagian ini, yang akan dideteksi adalah jenis dokumen. Dalam sistem yang dibangun, jenis dokumen yang dapat dideteksi adalah skripsi, tesis dan disertasi. Ilustrasi proses dapat diluhat pada gambar 8. Proses ini dilakukan dengan menggunakan kata kunci "skripsi", "tesis", atau "disertasi.



Gambar 8. Proses ekstraksi jenis dokumen

c. Nama penulis

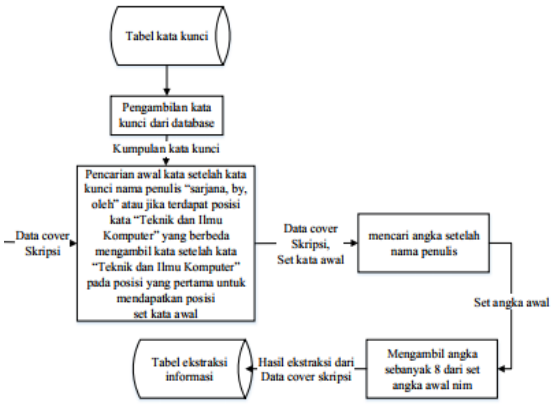
Nama penulis dalam sistem ini adalah nama yang menyusun skripsi. Berikut pada gambar 9 /adalah ilustrasi ekstraksi mendeteksi penulis. Nama penulis dideteksi dengan menggunakan kata kunci "sarjana", "oleh", atau "Teknik dan Ilmu Komputer".



Gambar 9. Proses deteksi penulis

d. NIM

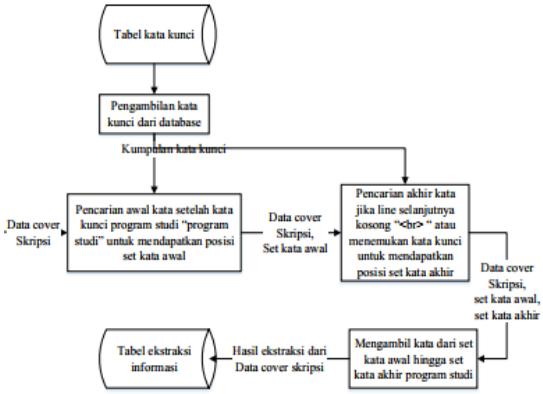
Setelah mendeteksi nama, berikutnya adalah mendeteksi NIM yaitu Nomor Induk Mahasiswa. Pada gambar 9 dapat dilihat ilustrasi proses mengekstraksi NIM dalah suatu dokumen skripsi. NIM dideteksi dengan menggunakan pencarian angka setelah nama penulis.



Gambar 9. Proses ekstraksi NIM

e. Program studi

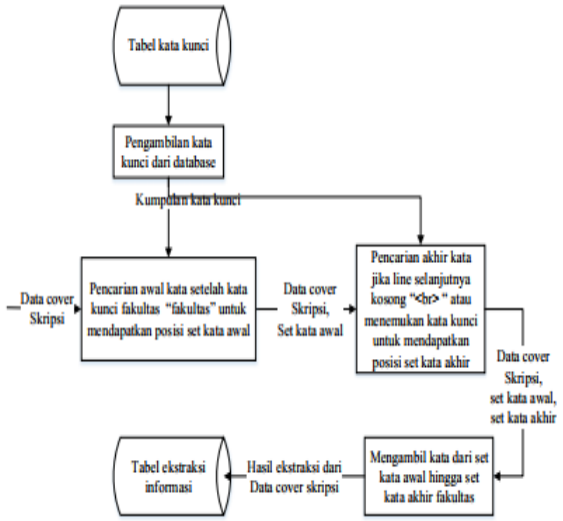
Pada bagian ini nama program studi akan dideteksi dengan menggunakan aturan pencarian kata kunci "Program studi", maka kata berikutnya adalah nama program studi hingga akhir baris tersebut. Proses tersebut dapat dilihat pada gambar 10.



Gambar 10. Proses ekstraksi nama prodi

f. Fakultas

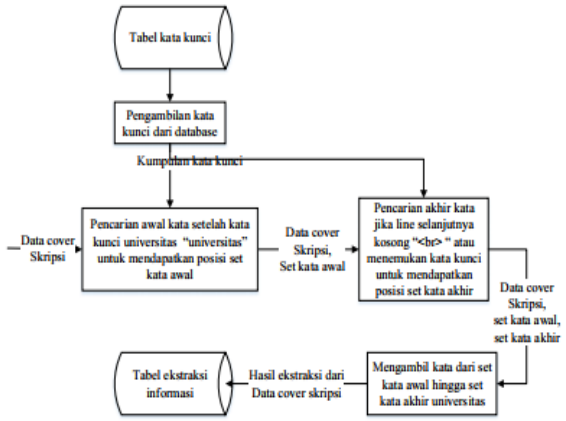
Hampir serupa dengan ekstraksi nama program studi, nama fakultas juga dilakukan dengan menggunakan kata kunci "Fakultas". Hingga akhir baris akan dideteksi sebagai nama fakultas tersebut. Proses ini dapat dilihat pada gambar 11.



Gambar 11. Proses ekstraksi nama fakultas

g. Universitas

Nama universitas di deteksi dengan menggunakan aturan yang hampir sama dengan mendeteksi nama program studi dan fakultas. Nama universitas selalu muncul pada baris akhir sebelum tahun. Selain itu juga dapat dideteksi dengan kata kunci "Universitas". Proses ini dapat dilihat pada gambar 12.



Gambar 12. Proses ektraksi nama universitas

2.4.2 Ekstraksi Informasi Data Abstrak Skripsi

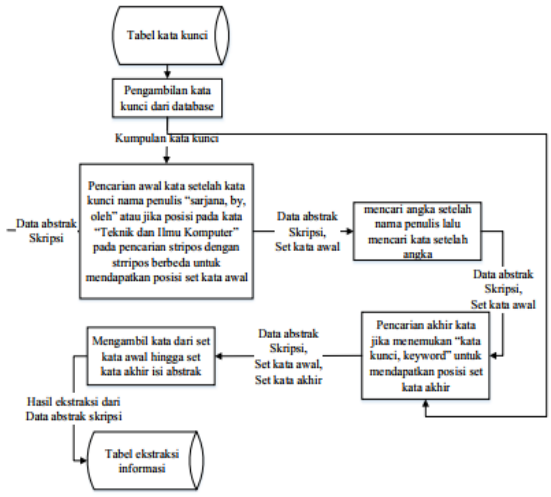
Ekstraksi informasi data abstrak menggunakan rule based. Data yang dicari adalah sebagai berikut:

a. Judul
b. Nama
c. NIM
d. Isi abstrak
e. Kata kunci

Pendeteksian judul, nama dan NIM dilakukan dengan cara yang sama seperti pendeteksian judul, nama dan NIM pada dokumen cover. Dalam mengekstraksi dokumen abstrak sedikit berbeda pada bagian ekstraksi isi abstrak dan kata kunci. Berikut proses pendeteksian isi abstrak dan kata kunci.

a. Isi abstrak

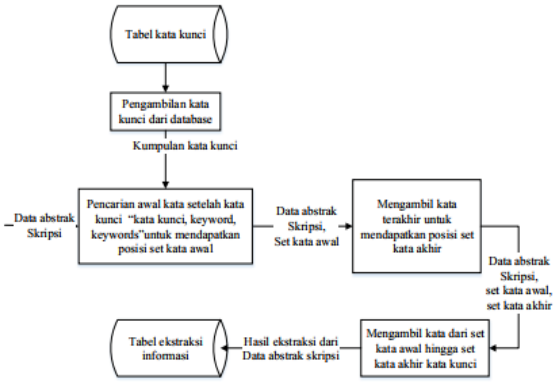
Untuk mengekstraksi isi abstrak, ada beberapa hal yang harus diperhatikan. Awal kata setelah mendeteksi nama penulis dan NIM sebagai titik awal. Setelah itu dicari kata kunci "kata kunci : ", kata tersebut menjadi tanda akhir sebuah abstrak. Sebagai ilustrasi, proses ekstraksi isi abstrak dapat dilihat pada gambar 13.



Gambar 13. Proses ekstraksi isi abstrak

b. Kata kunci

Dalam ekstraksi kata kunci merupakan langkah terakhir dalam sistem ini. Pendeteksian kata kunci dilakukan dengan menggunakan kata kunci "kata kunci : ". setelah kata kunci tersebut, semua kata diambil hingga akhir kalimat. Proses ekstraksi kata kunci dapat dilihat pada gambar 14.



Gambar 14. Proses ekstraksi kata kunci

Seluruh proses ekstraksi yang dilakukan pada dokumen abstrak, juga dilakukan pada dokumen abstract. Hanya saja kata kunci yang digunakan, semuanya dalam bahasa Inggris.

2.5 Pengujian

Pengujian ekstraksi merupakan tahap yang memiliki tujuan untuk mengetahui performa dari aturan yang dihasilkan berdasarkan data yang digunakan. Aturan tersebut digunakan pada sistem yang dibangun. Metode pengujian yang digunakan yaitu metode mencocokan data hasil ekstraksi informasi dengan pencarian manual. Pengujian tersebut dibagi menjadi 3 pengujian yaitu pengujian dengan membandingkan 5 dokumen uji dengan dokumen skripsi yang ada pada sistem sejumlah 50 data skripsi cover, data skripsi abstrak, dan data skripsi abstract.

3 PENUTUP

Kesimpulan yang didapat dari penelitian yang telah dilakukan diketahui bahwa dari 50 dokumen skripsi cover, abstrak, dan abstract tidak ada yang ekstraksi yang gagal atau tidak sesuai sehingga akurasi ekstraksi yang dilakukan pada 3 dokumen skripsi cover, abstrak, dan abstract skripsi yaitu 100%. Maka disimpulkan ekstraksi informasi penelitian ini dapat digunakan untuk mengekstraksi data skripsi pada dokumen cover, abstrak, dan abstract.

Berdasarkan hasil penelitian yang telah dilakukan, masalah yang muncul pada penelitian ini adanya hasil konversi dari pdf ke html menemukan

simbol yang tidak beraturan sehingga ekstraksi informasi yang dilakukan tidak berhasil. Adapun saran untuk kajian lebih lanjut sebagai berikut:

a. Memakai library yang lebih sesuai sehingga bisa mengurangi kesalahan pada ekstraksi.
b. Cara lain mengatasi hasil konversi yang tidak sesuai adalah dengan menambah metode atau algoritma pendeteksian kesalahan teks pada dokumen bahasa indonesia
c. Untuk pengembangan selanjutnya, menambahkan lebih banyak format dokumen skripsi yang berbeda untuk memperkaya aturan sehingga mendukung penanganan variasi data yang berbeda.

DAFTAR PUSTAKA

[1] J. Piskorski and Y. Roman. Information Extraction: Past, Present and Future," in Multi-source, Multilingual Information Extraction and Summarization. Berlin Haidelberg. Springer. pp. 23 - 49. 2013.

[2] P. R. Aragues, J.-C. Chappelier and M. Rajmn. *Using Imformation Extraction to Classify Newspapers Advertisements*, Journal International on Statistical Analysis of Textual Data (JADT), 2000.

[3] B. G. Buchanan and R. O. Duda. *Principles of Rule-Based Expert Systems*. Advances in Computer*s*. vol. 22, pp. 163 - 236, 1983.

[4] J. Jing. Information Extraction from Text. Spinger Science and Business Media, 2012..

[5] A. D. d. Nindyati. Panduan Penulisan Skripsi Atau Laporan Tugas Akhir, Jakarta: Universitas Paramadina. 2015.

[6] M. Krifka. Basic Notion of Information Structure," Humboldt Universitat zu Berlin. Barlin. 2006.

[7] M. H. Taqwim. Ekstraksi Informasi Dengan Metode Rule - Based untuk Evaluasi Pemahaman Fisika Kinematika.[Online].Available: https://elib.unikom.ac.id/files/disk1/719/jbptunikompp-gdl-muhammadha-35901-5-unikom_m-r.pdf. 2016.

[8] A. Ismaya. Algoritma Ekstraksi Informasi Berbasis Aturan. Journal Universitas Gajah Mada. . Yogyakarta, 2014.

[9] H. Anette and B. B. Megyesi, " A Study on Automatically Extracted Keywords in Text Categorization," *Association for Computational Linguistics*, pp. 537-544, 2006.