IMPLEMENTASI EKSTRAKSI INFORMASI PADA DOKUMEN TEKS SKRIPSI MENGGUNAKAN METODE RULED BASED

Dimas Mustaqwa,Nelly Indriani

Program Studi Teknik Informatika
Fakultas Teknik dan Ilmu Komputer Universitas Komputer Indonesia
Jl. Dipati Ukur 114 Bandung
E-mail : dimasmustaqwa@gmail.com

ABSTRAK

Pencarian suatu informasi salah satunya dengan cara membaca suatu dokumen. Dokumen yang tersedia dalam bentuk softcopy umumnya menggunakan media penyimpanan berformat 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]

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. . [2] Metode *rule based* dapat digunakan jika suatu dokumen tersebut merupakan dokumen yang terstruktur. 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 ilmuah. Penulisan skripsi ini dimaksudkan sebagai pelatihan bagi mahasiswa untuk menuangkan gagasannya dalam bentuk karya ilmiah. [3] 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. [4]

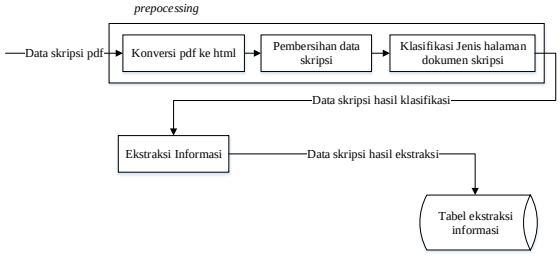
Salah satu penelitian tentang ekstraksi informasi pada rule base berjudul ”ekstraksi

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 memeberikan jawaban yang sesuai dari data yang didapatkan. Hasil akurasi yang diperoleh sebesar 90.6%-95.4% pada penelitan evaluasi pemahaman fisika kinematika. [5] 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%. [6]

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. 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.

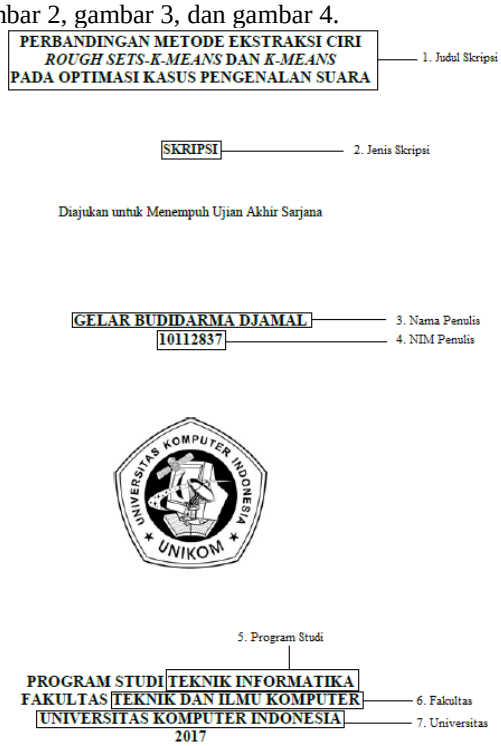


Gambar 1. Alur Proses ekstraksi data skripsi

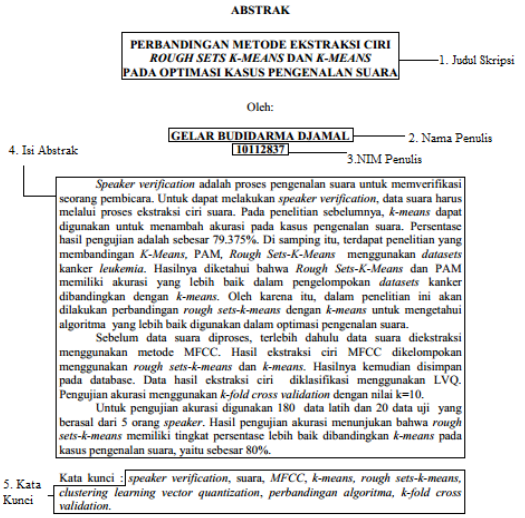
Sebelum melakukan analisis sistem dilakukan analisis bertujuan membuat rule based pada data skripsi cover, abstrak, dan abstract

2.1. Analisis Kata Kunci dan Aturan

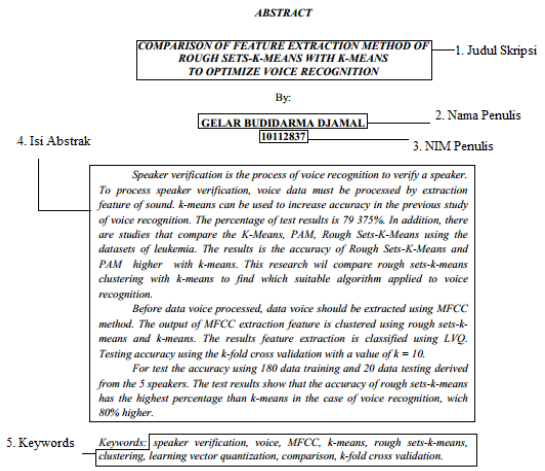
Pada penelitian ini menggunakan data skripsi yang berformat .pdf. Analisis berfokus pada pencarian kata kunci dan aturan yang akan digunakan untuk mengekstraksi informasi data skripsi. Data skripsi yang akan dianalisis yaitu data cover, abstrak bahasa indonesia dan abstrak bahasa inggris. Data skripsi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



Gambar 2. Data Skripsi Cover



Gambar 3. Data Skripsi Abstrak



Gambar 4. Data Skripsi Abstract

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 Skri**psi**

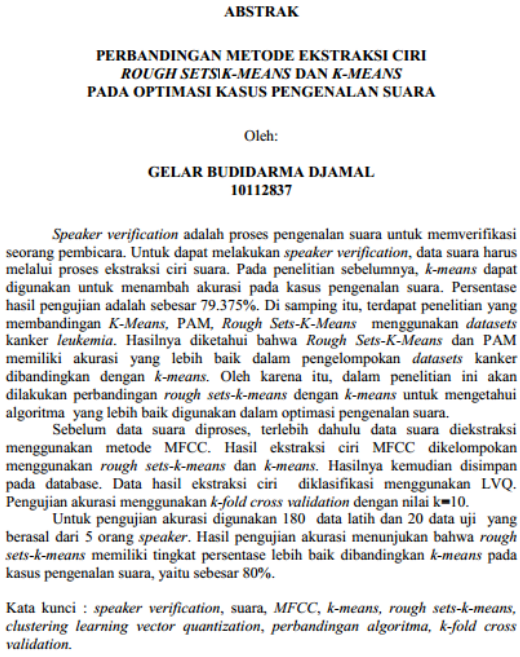
Identifikasi	**Kata kunci**	**Keterangan**
Data Skripsi Cover		
Judul skripsi	-	Mengambil judul skripsi dari awal kata hingga menemukan kata kunci jenis skripsi yang terakhir dari data skripsi
Jenis Skipsi	Skripsi, Tesis, Disertasi	Mencari kata “Skripsi, Tesis, Disertasi” yang terakhir dari data skripsi
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
Identifikasi	**Kata kunci**	**Keterangan**
Kampus	Universitas	Mencari kata “Universitas” yang terakhir dari data skripsi mengambil kata setelah kata “Universitas” hingga menemukan angka
Data Skripsi 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

2.2 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.

2.2.1 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 dan gambar 6 berikut ini.



ABSTRAK

PERBANDINGAN METODE EKSTRAKSI CIRI *ROUGH SETSK-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 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 Sets-K-Means* dan PAM memiliki akurasi yang lebih baik dalam pengelompokan *datasets* kanker dibandingkan dengan *k-means*. Oleh karena itu, dalam penelitian ini akan dilakukan perbandingan *rough sets-k-means* dengan *k-means* untuk mengetahui algoritma yang lebih baik digunakan dalam optimasi pengenalan suara.

Sebelum data suara diproses, terlebih dahulu data suara diekstraksi menggunakan metode MFCC. Hasil ekstraksi ciri MFCC dikelompokan menggunakan *rough sets-k-means* dan *k-means*. Hasilnya kemudian disimpan pada database. Data hasil ekstraksi ciri diklasifikasi menggunakan LVQ. Pengujian akurasi menggunakan *k-fold cross validation* dengan nilai k=10.

Untuk pengujian akurasi digunakan 180 data latih dan 20 data uji yang berasal dari 5 orang *speaker*. Hasil pengujian akurasi menunjukan bahwa *rough sets-k-means* memiliki tingkat persentase lebih baik dibandingkan *k-means* pada kasus pengenalan suara, yaitu sebesar 80%.

Kata kunci : *speaker verification*, suara, *MFCC*, *k-means*, *rough sets-k-means*, *clustering learning vector quantization*, *perbandingan algoritma*, *k-fold cross validation*.

Gambar 5. Data Skripsi Abstrak berupa pdf

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
 sebesar 79.375% .   Di   samping itu ,   terdapat penelitian yang mem bandingan   K - Means, PAM , Rough Sets - K - Means   menggunakan datasets   kanker   leukemia . H asil nya   diketahui   bahwa   R ough S ets - K - M eans   dan PAM memiliki akurasi   yang

Gambar 6. Hasil Konversi data skripsi abstrak

2.2.2 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 7 berikut ini.

Sesudah Pembersihan
i ABSTRAK PERBANDINGAN METODE EKSTRAKSI 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 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 Sets-K-Means dan PAM memiliki akurasi yang

Gambar 7. Hasil pembersihan data skripsi

2.2.3 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 skripsi
Klasifikasi	**Status data**	**Keterangan**
3	Bukan data skripsi cover atau abstrak.	Tidak ditemukan kata pada pencarian yang dilakukan pada klasifikasi

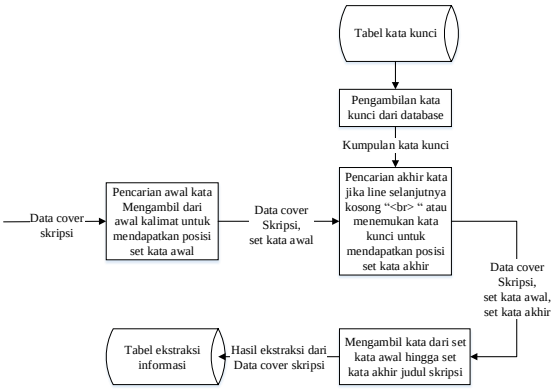
2.3 Ekstraksi Informasi Data

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.

2.3.1 Ekstraksi Informasi Data Cover Skripsi

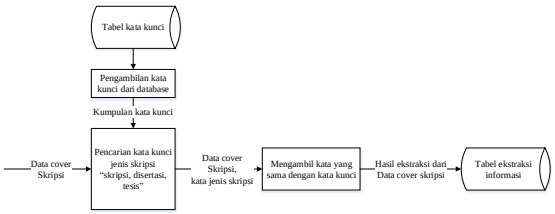
Ekstraksi informasi data cover menggunakan rule based dapat dilihat pada gambar 8 sampai gambar 14.

a. Ekstraksi judul skripsi



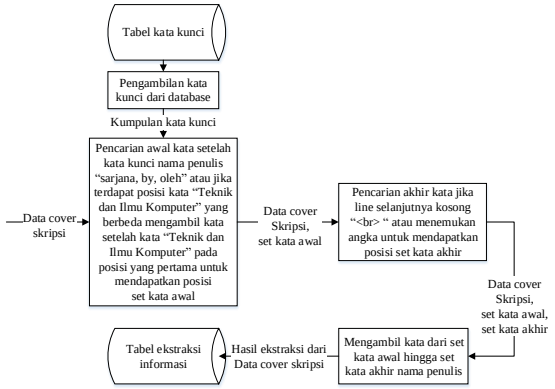
Gambar 8. Ekstraksi judul skripsi

b. Ekstraksi jenis skripsi



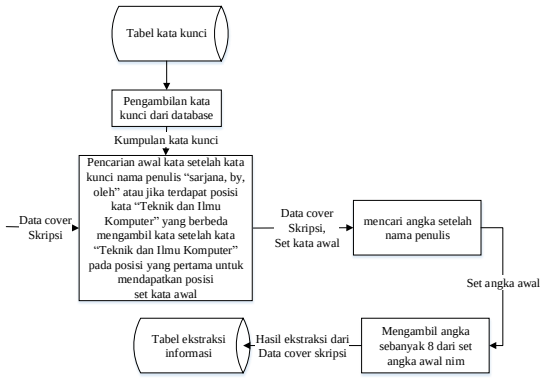
Gambar 9. Ekstraksi jenis skripsi

c. Ekstraksi nama penulis



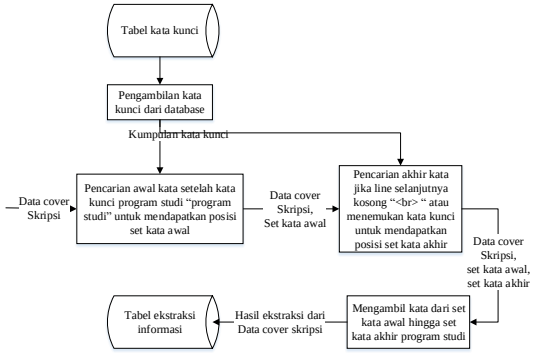
Gambar 10. Ekstraksi nama penulis

d. Ekstraksi nim



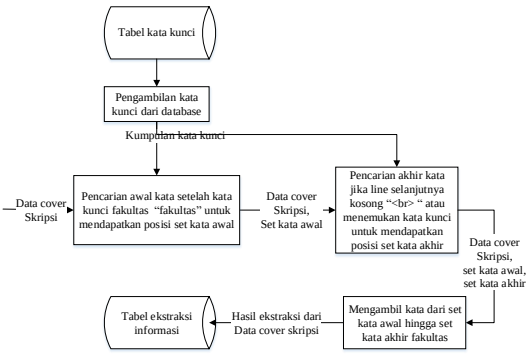
Gambar 11. Ekstraksi nim

e. Ekstraksi program studi



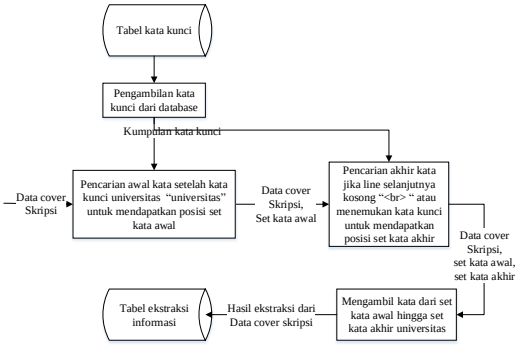
Gambar 12. Ekstraksi program studi

f. Ekstraksi fakultas



Gambar 13. Ekstraksi fakultas

g. Ekstraksi universitas

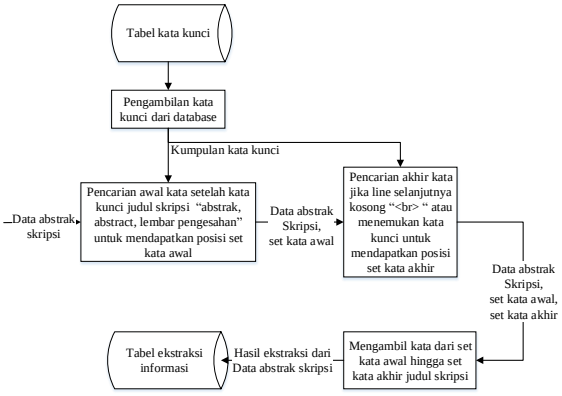


Gambar 14. Ekstraksi universitas

2.3.2 Ekstraksi Informasi Data Abstrak Skripsi

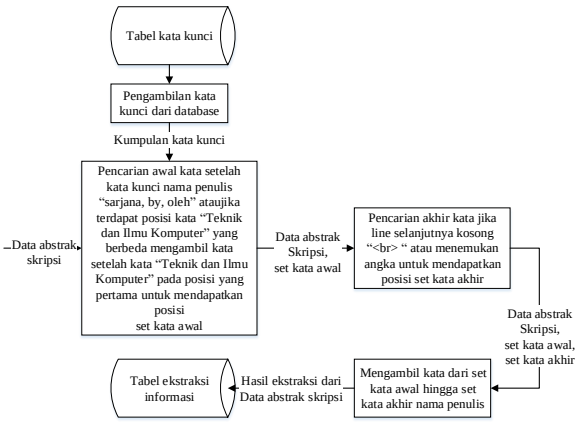
Ekstraksi informasi data abstrak menggunakan rule based dapat dilihat pada gambar 15 sampai gambar 19.

a. Ekstraksi judul skripsi



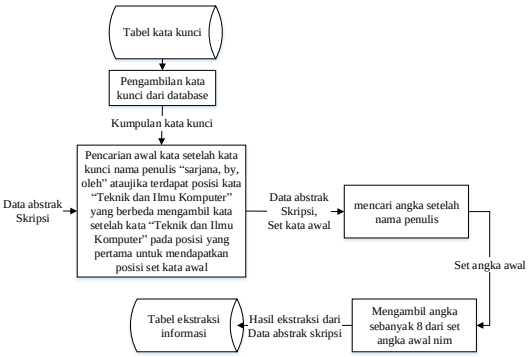
Gambar 15. Ekstraksi judul skripsi

b. Ekstraksi nama penulis



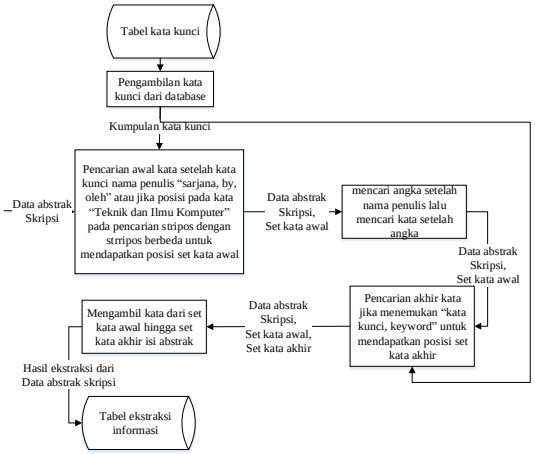
Gambar 16. Ekstraksi nama penulis

c. Ekstraksi nim



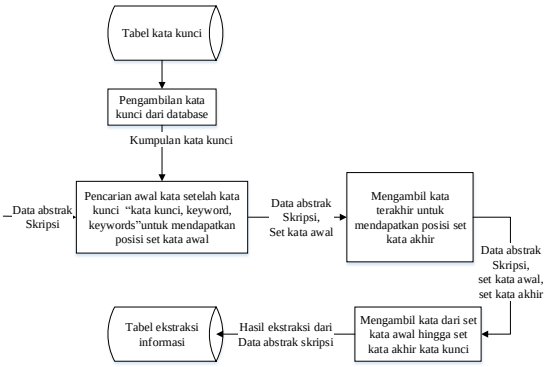
Gambar 17. Ekstraksi nim

d. Ekstraksi isi abstrak



Gambar 18. Ekstraksi isi abstrak

e. Ekstraksi kata kunci



Gambar 19. Ekstraksi kata kunci

3. PENGUJIAN DAN PEMBAHASAN

Pengujian ekstraksi merupakan tahap yang memiliki tujuan untuk mengetahui performa dari metode seleksi fitur yang digunakan pada sistem yang dibangun yaitu metode mencocokan data hasil ekstraksi informasi dengan pencarian manual. Dibagi menjadi 3 pengujian yaitu pengujian perbandingan dokumen dengan data sistem pada 50 data skripsi cover, data skripsi abstrak, dan data skripsi abstact.

3.1 Hasil Perbandingan Data Skripsi Cover

Perbandingan data skripsi cover merupakan pengujian perbandingan dokumen dengan data sistem pada data skripsi cover. Hasil perbandingan yang dilakukan dilihat pada tabel 3 berikut.

Tabel 3. Hasil pengujian perbandingan dokumen dengan data sistem data skripsi cover

No	Pengujian Data Skripsi Cover	Tahun Lulus	Keterangan	Kesimpulan
1	gelar budidarma djamal	2017	Ekstraksi sesuai	Ekstraksi Berhasil
2	idwar halid	2017	Ekstraksi sesuai	Ekstraksi Berhasil
3	narji jaariah	2017	Ekstraksi sesuai	Ekstraksi Berhasil
4	faisal muslim	2017	Ekstraksi sesuai	Ekstraksi Berhasil
5	dede juniawan suri	2017	Ekstraksi sesuai	Ekstraksi Berhasil

3.2 Hasil Perbandingan Data Skripsi Abstrak

Perbandingan data skripsi abstrak merupakan pengujian perbandingan dokumen dengan data sistem pada data skripsi abstrak. Hasil perbandingan yang dilakukan dilihat pada tabel 4 berikut.

Tabel 4. Hasil pengujian perbandingan dokumen dengan data sistem data skripsi abstrak

No	Pengujian Data Skripsi Cover	Tahun Lulus	Keterangan	Kesimpulan
1	gelar budidarma djamal	2017	Ekstraksi sesuai	Ekstraksi Berhasil
2	idwar halid	2017	Ekstraksi sesuai	Ekstraksi Berhasil
3	narji jaariah	2017	Ekstraksi sesuai	Ekstraksi Berhasil
4	faisal muslim	2017	Ekstraksi sesuai	Ekstraksi Berhasil
5	dede juniawan suri	2017	Ekstraksi sesuai	Ekstraksi Berhasil

3.3 Hasil Perbandingan Data Skripsi Abstract

Perbandingan data skripsi abstract merupakan pengujian perbandingan dokumen dengan data sistem pada data skripsi abstract. Hasil perbandingan yang dilakukan dilihat pada tabel 5 berikut.

Tabel 5. Hasil pengujian perbandingan dokumen dengan data sistem data skripsi abstract

No	Pengujian Data Skripsi Abstract	Tahun Lulus	Keterangan	Kesimpulan
1	gelar budidarma djamal	2017	Ekstraksi sesuai	Ekstraksi Berhasil
2	idwar halid	2017	Ekstraksi sesuai	Ekstraksi Berhasil
3	narji jaariah	2017	Ekstraksi sesuai	Ekstraksi Berhasil
4	faisal muslim	2017	Ekstraksi sesuai	Ekstraksi Berhasil
5	dede juniawan suri	2017	Ekstraksi sesuai	Ekstraksi Berhasil

4. KESIMPULAN DAN SARAN

4.1 Kesimpulan

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.

4.2 Saran

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.

1. Memakai library yang lebih sesuai sehingga bisa mengurangi kesalahan pada ekstraksi.

DAFTAR PUSTAKA

[1] J. P. a. R. Yangarber, “Informastion Extraction: Past, Present and Future,” 2013.

[2] B. G. Buchanan, Principles of Rule-Based Expert Systems, 1982.

[3] A. D. Nindyati, “Buku Panduan Penulisan Skripsi atau Tugas Akhir Universitas Parmadina,” 2010.

[4] M. Krifka, Basic notions of information structure, 2008.

[5] M. H. Taqwim, “Ekstraksi Informasi Dengan Metode Rule - Based untuk Evaluasi Pemahaman Fisika Kinematika,” *Jurnal Informatika,* 2016.

[6] A. Ismaya, “Algoritma Ekstraksi Informasi Berbasis Aturan,” *Jurnal Informatika,* 2014.