Ekstraksi Informasi Berbasis Rule untuk Proceeding, Jurnal, dan Technical Report dengan Memanfaatkan Attribute Font dan Paragraf

Christian Aditya Santoso, *Teknik Informatika, Institut Sains dan Teknologi Terpadu Surabaya*,
Gunawan, *Teknik Informatika, Institut Sains dan Teknologi Terpadu Surabaya*.

Abstrak—**Digital library merupakan solusi yang baik untuk dunia edukasi. Hal ini disebabkan karena buku yang sudah berevolusi menjadi digital. Awalnya dalam bentuk fisik sekarang sudah dalam bentuk digital dengan ekstensi PDF. Namun untuk membangun sebuah digital library merupakan system yang besar dan kompleks, sehingga diperlukan bagian yang banyak. Penelitian ini mengambil satu bagian dari pengembangan system digital library, yaitu pada bagian preprocessing atau persiapan sumber data digital library. Penyediaan sumber data digital library sangat luas dan banyak. Fokus dari penelitian ini adalah penyediaan data dimana data tersebut adalah jurnal, prosiding dan paper. Dokumen tersebut dipilih karena dinilai memiliki manfaat yang besar untuk edukasi karena peneliti mendokumentasikan hasil penelitian pada dokumen tersebut. Dalam 1 paper tentunya ada bagian yang menjadi kunci yang menggambarkan intisari dari penelitian tersebut. Pada penelitian ini diambil informasi Judul, Abstract, Keyword dan penulis. Informasi tersebut dipercaya mampu menggambarkan intisari dari suatu paper. Proses dilakukan dengan terbagi menjadi 3 bagian besar yaitu konversi file mentah dengan ekstensi PDF menjadi file JSON, Proses pengambilan fitur, Proses ekstraksi informasi. Ekstraksi informasi pada penelitian ini menggunakan kumpulan rule yang diimplementasikan pada software. Rule didapatkan dari hasil pengamatan selama penelitian. Hasil dari penelitian dilakukan perhitungan dengan memberikan bobot dimana hal yang terberat memiliki pengaruh yang lebih besar. Ketelitian yang dicapai adalah 81.32% dimana dari hipotesa awal pada ketelitian 80%. Namun masih banyak pengembangan yang bisa dilakukan agar lebih baik lagi pada penelitian selanjutnya**

***Kata Kunci*—Ekstraksi Inforasmi, Rule, Jurnal, Prosiding, paper.**

I. PENDAHULUAN

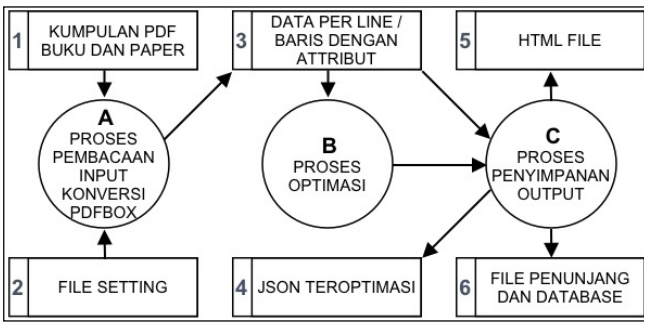
Perpustakaan merupakan media untuk mencari informasi yang lengkap dan menjadi penunjang informasi dalam proses belajar mengajar pada dunia edukasi. Namun dalam hal penyediaan sumber pada perpustakaan tidaklah mudah.

Christian Aditya Santoso, Informatics Department, Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Jawa Timur, Indonesia
Gunawan, Department of Information Technology, Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: gunawan@stts.edu)

Apabila menggunakan metode konvensional tentunya harus membeli banyak sekali buku. Permasalahan lebih rumit ketika sudah mengalami perkembangan teknologi dimana semua sudah serba digital dan kemudahan mendapatkan informasi menjadi prioritas. Hal ini juga terjadi pada perpustakaan maka tercetuslah perpustakaan digital. Perpustakaan digital tidak mudah untuk diwujudkan karena kompleks dan memiliki banyak bagian. Namun apabila tidak dimulai maka juga tidak akan terwujud. Penelitian ini mengambil salah satu bagian dari digital library yaitu penyediaan data sumber untuk diekstraksi informasinya guna dilakukan proses clustering yang digunakan untuk melakukan pencarian yang handal pada mesin pencarian perpustakaan digital. Proses yang akan dilakukan adalah konversi file pdf menjadi data terlebih dahulu pada penelitian ini dipilih JSON. Apabila file dinilai sebagai jurnal ataupun prosiding maka akan ada langkah lebih lanjut yaitu mencari nomor file JSON dari masing-masing prosiding atau jurnal. Langkah selanjutnya dilakukan dengan pengambilan fitur dari masing-masing element yang akan di ekstraksi dalam penelitian ini adalah judul, abstrak, keyword dan penulis. Setelah fitur terkumpul maka dilakukan proses ekstraksi informasi sehingga bisa menghasilkan data yang siap untuk di cluster.

II. PROSES KONVERSI PDF DAN TAGGING PAPER TUNGGAL PADA JURNAL DAN PROSIDING

Seperti yang dibahas pada pendahuluan proses ini merupakan proses pertama yang dilakukan dalam penelitian ini. Konversi PDF dilakukan dengan menggunakan library PDFBox dimana PDFBox merupakan library cross platform. Pada penelitian ini software ditulis dalam Bahasa JAVA. Proses pada bagian ini akan dibahas menjadi 2 hal yaitu konversi PDF dan tagging paper tunggal pada jurnal dan prosiding.



Gambar. 1. Gambar Arsitektur Proses Konversi PDF

A. Proses Konversi PDF

Gambar 1 merupakan alur kerja dari Proses Konversi PDF. Pada gambar 1 terdapat 3 Subproses yang terjadi pada proses Konversi PDF. Subproses pertama adalah Proses Pembacaan Input Konversi PDF Box pada gambar 1 diwakili Proses A. Subproses kedua adalah Subproses Optimasi pada gambar 1 diwakili oleh Proses B. Subproses ketiga adalah Subproses Penyimpanan Output pada gambar 1 diwakili oleh Proses C.

Proses A merupakan proses pertama yang dijalankan oleh system. Input dari proses A adalah file dengan ekstensi PDF pada suatu folder serta file setting dimana file setting ini ditujukan untuk memudahkan penggantian folder pembacaan. Proses konversi dari PDF menjadi file JSON dilakukan dengan bantuan PDF Box. Dimana kinerja PDFBox adalah membaca baris perbaris dari PDF File. Namun dari pembacaan baris-perbaris juga di dapatkan informasi seperti letak dalam x dan y koordinat, serta attribute font family dan font size. Dari informasi tersebutlah digunakan sebagai informasi yang dijadikan input proses B pada gambar 1.

Proses B, merupakan proses proses lanjutan dari proses Konversi PDF setelah menghasilkan array dengan isi text dan attribute untuk masing-masing baris pada masing-masing halaman PDF, maka masih diperlukan langkah selanjutnya yaitu optimasi. Secara umum optimasi yang dilakukan adalah untuk membuat data baris perbaris menjadi kumpulan paragraf, sehingga tidak terpisahkan berdasar baris.

Proses optimasi dilakukan perlu dilakukan karena informasi yang dihasilkan oleh Subsistem 1 ini tidak bisa hanya informasi dalam masing-masing baris contohnya adalah abstract, bisa kita ketahui bahwa abstract pasti memiliki lebih dari satu baris, sehingga akan sangat susah sekali apabila data yang akan dikonversi ini masih dalam wujud perbaris. Contoh lain adalah judul, banyak sekali paper penelitian yang memiliki judul penelitian lebih dari satu baris sangat jarang sekali bisa ditemukan ada judul penelitian yang hanya memilki 1 baris. Dengan permasalahan tersebut maka diputuskan untuk mengembangkan proses optimasi ini. Harapan yang dihasilkan dari optimasi ini adalah data yang awalnya terpisah dari tiap barisnya, akan bisa menjadi bergabung dalam paragraf-paragraf. Hal ini tentunya akan mempermudah proses ekstraksi informasi yang akan dilakukan pada proses selanjutnya.

Proses C merupakan proses terakhir pada gambar 1, bagian ini merupakan proses penyimpanan hasil output. File yang disimpan cukup banyak yaitu, File PDF yang sudah di rename, File JSON dari masing-masing halaman yang telah dikonversi, File txt non optimized, File HTML non Optimized serta gambar dari halaman pertama jurnal, prosiding dan paper. Hal ini dilakukan agar proses pemeriksaan bisa dilakukan dengan mudah dan cepat.

B. Proses Tagging paper Tunggal pada Jurnal dan Prosiding

Jurnal dan prosiding memiliki model perlakuan yang berbeda dengan paper tunggal. Hal ini disebabkan oleh jurnal maupun prosiding merupakan kumpulan dari beberapa paper tunggal hal ini yang menjadi dasar dilakukannya proses ini. Proses ini masih ada melalui proses manual yaitu mendata semua judul-judul paper tunggal pada daftar isi dan dilakukan penyimpanan pada database. Setelah proses manual tersebut dilakukan maka dicarilah File JSON yang merupakan halaman pertama dari masing-masing paper tunggal.

III. PROSES EKSTRAKSI FITUR

Proses Ekstraksi Fitur merupakan proses yang akan dijalankan terlebih dahulu sebelum dilakukan ekstraksi informasi [1]. Secara software yang diimplementasikan menjadi satu software mandiri dari software yang ada di section B, bahkan juga menjadi software yang berbeda dari Proses-proses sebelumnya. Ekstraksi Fitur ini merupakan bagian yang mencoba mengenali bagian-bagian element yang akan diekstraksi informasinya. Element yang akan diekstraksi adalah sebagai berikut :

- Judul Paper : merupakan nama yang dipakai untuk buku atau bab dalam buku yang dapat menyiratkan secara pendek isi atau maksud buku atau bab itu.
- Abstract Paper : merupakan sebuah ringkasan isi dari sebuah karya tulis ilmiah yang ditujukan untuk membantu seorang pembaca agar dapat dengan mudah dan cepat untuk melihat tujuan dari penulisannya.
- Kata Kunci Paper : merupakan kata-kata yang mengandung konsep pokok yang dibahas dalam karya tulis.
- Penulis : merupakan seseorang atau beberapa orang yang melakukan penelitian dan membuat laporan hasil penelitian dalam bentuk paper / karya ilmiah.

Dengan penjelasan tersebut maka dipilih empat informasi tersebut untuk dipilih karena tujuan dari penelitian ini adalah untuk menghasilkan informasi yang menggambar intisari dari isi suatu paper ilmiah sehingga bisa dilakukan clustering atau pengelompokan data pada subsistem tiga. Dari empat informasi tersebut hal yang dapat menggambarkan mengenai intisari dari paper pada sesungguhnya hanya pada informasi judul paper, abstrak paper, dan kata kunci paper. Penulis di ekstraksi untuk tujuan pembuatan front end aplikasi pencarian data. Namun pada penelitian ini hanya sebatas hanya bisa mengekstrasi data penulis.

Proses Ekstraksi fitur ini menggunakan metode pengenalan akan kata kunci. Masing-masing element yang akan diambil informasinya sehingga software memiliki kumpulan kata kunci untuk masing-masing element yang akan diekstraksi informasinya. Kata kunci tersebut didapatkan dengan proses pengamatan selama penelitian, sehingga ditemukan

kumpulan kata kunci tersebut untuk mengenali masing-masing element.

TABEL I
DAFTAR KATA KUNCI

Pendahuluan	Keyword	Abstrak
Introduction	.key.word	abstract
1. Introduction	key.word	abstracts
1. Pendahuluan	(.*)key(.*)word(.*)	a b s t r a c t
1. Pendahuluan	keyword	a b s t r a c t s
I	keywords	abstrak
i	key-words	
I. Introduction	key-word	
i. Introduction	key words	
I. Pendahuluan	key word	
i. Pendahuluan	index terms	
Background	index term	
	indexterms	
	indexterm	
	index-terms	
	index-term	

Tabel 1 merupakan daftar kata kunci yang diterapkan untuk pencarian element. Pada table tidak terdapat untuk kata kunci judul dan juga kata kunci penulis. Hal ini dikarenakan untuk dua element tersebut dilakukan metode dan cara berbeda untuk mengekstraksi fitur yang dibutuhkan. Setelah dilakukan proses ekstraksi fitur maka dilakukan ekstraksi informasi pada masing-masing element.

IV. PROSES EKSTRAKSI INFORMASI

Proses Ekstraksi informasi merupakan inti dari penelitian ini. Proses ekstraksi informasi terbagi menjadi 4 bagian besar yaitu ekstraksi informasi pada bagian judul, abstract, keyword dan author. Proses ekstraksi informasi menggunakan rule dimana rule ini dihasilkan dan disimpulkan dari proses penelitian. Berbagai penelitian sebelumnya telah dilakukan untuk pengolahan teks [2], [3] dan ekstraksi informasi [4]

A. Ekstraksi Informasi Judul

Proses ekstraksi informasi dari element judul berbeda dengan element yang lain karena element ini tidak dilakukan pengenalan element terlebih dahulu [5]–[8]. Element judul langsung dilakukan ekstraksi informasi. Namun untuk melakukan ekstraksi tersebut selama penelitian tetap mengumpulkan fakta-fakta yang bisa dijadikan acuan untuk mengindentifikasi judul. Fakta-fakta yang berhasil dikumpulkan adalah sebagai berikut :

- Judul selalu berada dibagian awal
- Judul selalu memiliki font yang paling besar
- Judul selalu memiliki tipe font Bold
- Judul selalu dituliskan sebelum element author
- Judul ditulisan dengan Upper case.

Dengan didapatkan fakta mengenai tata cara penulisan serta peletakan element judul maka di berikan beberapa rules yang digunakan untuk melakukan ekstraksi informasi element judul. Berdasar 5 fakta yang disimpulkan dari semua fakta tersebut diintepretasikan kedalam software. Implementasi ke dalam software dilakukan dengan membuat rule-rule yang apabila di eksekusi akan mampu menghasilkan output yang baik. Rule yang diterapkan pada bagian judul disajikan pada table 2.

TABEL 2
DAFTAR RULE

No	Rule	Keterangan
1	Pencarian Font Size Terbesar pada halaman pertama	Disimpulkan bahwa judul di tuliskan dalam Font Size terbesar pada halaman pertama
2	Apakah Abstrak sudah terdeteksi?	Rule ini untuk menentukan batas pencarian
3	Apakah Keyword sudah terdeteksi?	Rule ini untuk menentukan batas pencarian
4	Apakah Introduction sudah terdektesi?	Rule ini untuk menentukan batas pencarian
5	Pencarian paragraf dengan font size terbesar	Rule untuk mengenali paragraph yang memiliki font terbesar dan di jadikan sebagai Judul
6	Apakah hasil dari no 5 memiliki karakter lebih dari 30	Rule ini digunakan untuk mendeteksi apakah yang terdeteksi merupakan benar judul, karena judul tidak mungkin hanya 30 karakter
7	Pencarian dilanjutkan dengan mengambil fontsize ke 2 terbesar	Proses ini dilakukan sampai Langkah 6 menghasilkan jumlah karakter cukup

Dengan menggunakan 7 rule tersebut maka pencarian judul dilakukan. Masing-masing rule memiliki peran tersendiri. Rule ini masih terus bisa di explorasi dan ditambahkan agar ketelitian bisa terus meningkat.

B. Ekstraksi Informasi Abstrak

Metode yang digunakan untuk pengambilan informasi pada abstrak berbeda dengan metode pengambilan informasi pada judul [9]. Hal itu disebabkan pengenalan pada abstrak menggunakan output informasi dari subproses B. Namun prediksi dari subproses B masih belum selesai karena ketika diambil informasinya masih belum 100% lengkap dan benar. Berikut adalah masalah-masalah yang harus diselesaikan ketika melakukan ekstraksi informasi pada bagian abstrak.

- Terdapat penulisan kata kunci abstrak yang tidak tergabung dengan abstrak, sehingga hal ini akan membuat pengenalan dari subproses B masih belum memuat konten abstrak
- Terdapat penulisan kata kunci abstrak namun memiliki style berbeda dengan konten abstrak
- Terdapat penulisan abstrak yang diikuti oleh copyright penulisan
- Terdapat penulisan abstrak yang bergabung dengan keyword
- Terdapat penulisan abstrak yang keyword penanda abstrak bergabung dengan konten abstrak.

Dari 5 masalah tersebut yang harus diselesaikan pada subproses ekstraksi informasi untuk abstrak ini. Sehingga dapat dihasilkan abstrak yang rapi dan memuat semua konten dari abstrak. Dengan baiknya output yang dihasilkan tentunya akan membantu proses selanjutnya untuk bisa menghasilkan output yang optimal.

Permasalahan yang cukup banyak ditemui adalah banyak sekali keyword abstrak yang dituliskan sebagai judul subbab pada suatu paper ilmiah, penanganan untuk permasalahan ini adalah software akan membaca terlebih dahulu data dari table paperfeature dan melakukan filter data untuk pdf yang akan diekstrak informasinya kemudian dilakukan filter dengan jenisfeature abstract. Langkah pertama setelah mendapatkan data tersebut adalah dengan cara melihat jumlah karakternya apabila jumlah karakternya sedikit (dalam kasus ini di

lakukan pengecekan lebih kecil dari 10). Berikut adalah rule yang digunakan untuk mengekstraksi bagian abstrak.

TABEL 3
DAFTAR RULE

No	Rule	Keterangan
1	Pencarian Keyword Penentu Abstrak, daftar keyword terdapat pada tabel 5.5	Disimpulkan bahwa judul di tuliskan dalam Font Size terbesar pada halaman pertama
2	Apakah Index Abstrak Lebih Kecil dari Index Introduction?	Rule ini untuk menentukan batas pencarian
3	Apakah Akhir pencarian lebih besar dari index keyword dan index keyword lebih besar dari index abstrak?	Rule ini untuk menentukan batas pencarian
4	Apakah index keyword lebih besar dari index abstrak?	Rule ini untuk menentukan batas pencarian
5	Apakah start karakter keyword = 0?	Apabila tidak 0 maka artinya keyword di tuliskan Bersama dengan keyword
6	Apakah index abstrak sama dengan index pencarian?	Apabila true maka system akan mencari sampai index abstrak + 1
7	Apakah Index Abstrak = 0	Kasus ini jarang terjadi namun apabila terjadi maka system akan menentukan pencarian adalah index email + 1
8	Setelah di temukan start dan akhir pencarian system akan menggabungkan data text menjadi satu kesatuan	Proses ini digunakan apabila abstrak terpecah menjadi beberapa object JSON
9	Apakah Keyword kata kunci ditemukan dalam text gabungan?	Apabila true maka akan ada proses substring data dari awal sampai index kata kunci
10	Apakah Keyword abstrak ditemukan dalam text gabungan?	Apabila ditemukan maka akan ada proses substring untuk menghilangkan keyword tersebut
11	Apakah ada karakter © pada text gabungan?	Apabila ditemukan maka akan ada proses substring karena karakter tersebut merupakan informasi copyright dari publisher dimana hal tersebut tidak termasuk di dalam abstrak

Abstract merupakan rule yang paling banyak, karena bagian ini adalah bagian yang paling sukar untuk diekstraksi. Dengan menggunakan rule tersebut mendapat hasil yang cukup memuaskan. Dengan ditambah rule tentunya akan meningkatkan ketelitian dari ekstraksi informasi bagian abstrak.

C. Ekstraksi Informasi Keyword

Keyword merupakan element yang juga bisa menggambarkan isi dari suatu paper ilmiah sehingga bagian ini juga penting untuk dilakukan ekstraksi informasi. Element keyword pada sesungguhnya juga merupakan bagian dari abstrak dan tidak selalu pada setiap paper ilmiah.
Sama halnya terjadi pada abstrak element ini hasil output dari proses belum bisa langsung dilakukan ekstraksi informasi. Hal itu dikarenakan konsep tata tulis yang tidak 100% standar, sehingga pada bagian ini masih akan dilakukan pengolahan lebih lanjut agar hasil output untuk element ini lebih optimal. Hasil yang optimal tentunya akan meningkatkan performansi untuk proses selanjutnya. Berikut adalah permasalahan yang akan diselesaikan pada proses ektraksi element keyword.

- Penulisan kata kunci penanda element keyword yang menjadi satu dengan konten keyword
- Penulisan keyword yang bergabung dengan abstrak

Selama proses penelitian ini dilakukan terdapat dua permasalahan ini yang harus diselesaikan proses ekstraksi informasi untuk element keyword. Pada subproses ini akan menyelesaikan dua permasalahan tersebut sistem mampu menghasilkan output yang baik. Berikut adalah rule yang digunakan untuk mengekstraksi keyword

TABEL 4
DAFTAR RULE

No	Rule	Keterangan
1	Pencarian Keyword Penentu Abstrak, daftar keyword terdapat pada tabel 5.3	Disimpulkan terdapat keyword tertentu yang selalu dituliskan sebelum dituliskan kata kunci
2	Apakah index kata kunci == 0?	Apabila True maka pada paper tersebut disimpulkan tidak ada kata kunci karena terdapat paper yang memang tidak ada kata kunci sebagai element informasi.
3	Apakah index kata kunci memiliki karakter < 30?	Hal ini digunakan untuk memutuskan apakah kata kunci terbagi menjadi 2 baris atau tidak
4	Apakah index introduction di temukan?	Index introduction di butuhkan untuk menentukan batas pencarian.
5	Sistem akan melakukan iterasi atau tidak sampai pada batas tertentu berdasar hasil dari rule nomor 3	Iterasi ini digunakan untuk mengambil kumpulan text bagian kata kunci
6	Apakah keyword kata kunci dimulai dari index 0?	Apabila false maka dilakukan substring start dari awal keyword sampai akhir text
7	Apakah terdapat keyword kata kunci pada text?	Jika true maka system akan melakukan substring kembali dari akhir

Ekstraksi Keyword membutuhkan 7 rule. Keyword merupakan salah element yang memiliki nilai ketelitian baik. Namun masih ditemukan beberapa kendala dan tentunya pengembangan masih bisa terus dilakukan

D. Ekstraksi Informasi Author

Proses ekstraksi element author merupakan ekstraksi informasi terakhir. Output dari proses ini adalah berupa satu paragraf yang berisi informasi tentang detil dari author. Untuk mengidentifikasi element ini menggunakan bantuan NER (Name Entity Recognizer) [10]. Berbeda dengan metode yang akan dipakai pada bagian ini. Sebelum dilakukan ekstraksi informasi terlebih dahulu dilakukan pengamatan tata letak penulisan element author ini. Berikut adalah beberapa fakta yang disimpulkan untuk element author.

- Author ditulis setelah penulisan title
- Author ditulis sebelum penulisan abstrak
- Pada element author juga terdapat email dari masing-masing author

Dengan fakta tersebut maka disusunlah software untuk mengadopsi fakta-fakta yang ditemukan sehingga bisa menjadi dasar penunjang ekstraksi informasi yang

akan dilakukan. Berikut adalah table rule yang digunakan untuk ekstraksi informasi pada author

TABEL 5
DAFTAR RULE

No	Rule	Keterangan
1	Apakah sudah mendapat index judul?	Syarat agar hasil menjadi optimal adalah proses ini dilakukan setelah proses ekstraksi judul
2	Apakah indexabstrak < indexkeyword	Jika true maka batas pencarian sampai abstrak apabila false batas pencarian sampai pada keyword.
3	Proses iterasi penggabungan informasi author menjadi 1 text	Proses ini digunakan untuk menggabungkan text author dimana hasil selalu

Rule untuk bagian author merupakan rule yang paling sedikit. Author di ekstraksi namun tidak terlalu mendetail karena informasi mengenai author hanya digunakan sebagai pelengkap informasi. Bagian ini masih bisa dibedah lebih jauh lagi.

V. UJI COBA

Pada bagian ini akan membahas mengenai hasil dari penelitian yang telah dilakukan. Berawal dari data. Data yang dijadikan uji coba adalah sebesar 1080 data dimana dari 1080 data terdapat 4 pdf yang tidak berhasil di konversi sehingga menjadi 1076 data. Berikut adalah rincian mengenai data tersebut.

TABEL 6
KOMPOSISI DATA

No	Jenis Dokumen	Jumlah Dokumen
1	Jurnal / Prosiding	3 (167 Paper Tunggal)
2	Paper Tunggal	909
	TOTAL	1076

Dari 1076 dokument yang terekstraksi informasinya memiliki beberapa model penulisan paper yang diekstraksi. Variasi tersebut juga menjadi salah satu bagian dari tantangan pengujian system yang diteliti berikut adalah variasi yang dilakukan.

TABEL 7
KOMPOSISI DATA

No	Jenis Dokumen	Jumlah Dokumen
1	Tipe 1	91
2	Tipe 2	21
3	Tipe 3	55
4	Tipe 4	119
5	Tipe 5	408
6	Tipe 6	65
7	Tipe 7	36
8	Tipe 8	11
9	Tipe 9	270
	TOTAL	1076

Pada Tabel 7 terdapat 9 tipe penulisan paper yang menjadi target ujicoba pada penelitian ini. Tipe 1 sampai tipe 3 mewakili masing-masing jurnal ataupun prosiding, Tipe 4 paper tunggal namun juga memiliki kemiripan, Tipe 6, namun pada tipe 6 masih terdapat varian di dalamnya namun dikategorikan menjadi 1 kategori. Tipe 7 juga merupakan paper tunggal namun memiliki kemiripan dan bukan hanya dari 1 publisher. Hal sama juga pada tipe 7. Berbeda dengan Tipe 8 terdapat 11 dokumen namun dari 11 dokumen tersebut memiliki ciri khas model format penulisan yang berbeda-beda. Tipe 8 merupakan tipe yang digunakan sebagai kategori penulisan format namun pada waktu pencarian sumber data, paper yang masuk ke dalam tipe 8 ini penulis tidak menemukan banyak, sehingga di kategorikan menjadi 1 kategori, tipe 9 merupakan tipe terakhir. Berikut adalah table sampling dari hasil pengujian mengenai score yang diberikan pada masing-masing data.

TABEL 8
SCORE SAMPLE

No	Id PDF	Id TOC	Judul (0/1/2)	Penulis (0/1/2)	Abstrak (0/1/2)	Kata kunci (0/1/2)
1	1	0	2	2	2	2
2	2	0	2	2	2	2
3	3	0	2	2	2	2
4	4	0	2	2	2	2
5	5	0	2	2	2	2
6	6	0	2	2	2	2
7	7	0	2	2	2	2
8	8	0	2	2	1	0
9	9	0	2	2	2	2
10	10	0	2	2	2	2
****	****	****	****			
290	265	31	2	2	2	2
291	265	32	2	2	2	2
292	265	33	2	2	2	2
293	265	34	2	2	2	2
294	265	35	2	2	0	2
295	265	36	2	2	2	2
296	265	37	2	2	2	2
297	265	38	2	2	2	2
298	265	39	2	2	2	2
299	265	40	2	2	2	2
****	****	****	****			
850	712	0	2	2	2	2
851	713	0	2	2	2	2
852	714	0	2	2	2	2
853	715	0	2	2	2	2
854	716	0	2	2	2	2
855	717	0	2	2	2	2
856	718	0	2	2	2	2
857	719	0	2	2	2	2
858	720	0	2	2	2	2
859	721	0	2	2	2	2

Tabel 6.6 merupakan potongan dari hasil pencocokan / koreksi data ekstraksi informasi dari system. no 1 – 10 merupakan paper tunggal ditandai dengan id_toc bernilai 0, hal serupa sama dengan no 850 – 859 merupakan paper tunggal. Sedangkan no 500 – 509 memiliki id_pdf yang sama namun id_tocnya tidak 0 menandakan data pada bagian itu adalah prosiding atau jurnal. Poin terpenting ada pada kolom score. Pengisian score dilakukan dengan cara sebagai berikut:

- Apabila judul tidak terdeteksi sama sekali maka akan mendapat score sebesar 0
- Apabila judul terdeteksi namun tidak sempurna maka akan mendapat score sebesar 1
- Apabila judul terdekteksi dan menghasilkan informasi yang benar maka akan mendapat score sebesar 2

Dengan menggunakan acuan tersebut maka score akan dihitung dan diolah untuk menghasilkan nilai ketelitiannya. Pada proses ujicoba dilakukan lebih dari 1x running da (1)
pengolahan dan pendeteksian 1000 data. Namun t
ditentukan dulu rumus untuk melakukan pehitungan ketelitian. Rumus yang dimaksud adalah

$$Percentage\ (\%) = \frac{Sum(Score\ Judul)}{2 * Jumlah\ Data}\ x\ 100$$

Pada Rumus 6.1 merupakan cara perhitungan untuk mencari nilai percentage ketelitian. Secara umum perhitungan dilakukan dengan cara menjumlah semua score yang didapat dan kemudian dibagi dengan jumlah score penuh dalam kasus ini adalah 2 karena poin maximum untuk masing-masing judul dan jumlah data yang digunakan pada penelitian ini adalah 1076 data. Setelah dihitung maka dikalikan dengan 100 untuk mendapat nilai persen untuk ketelitian judul. Berikut adalah perhitungan final untuk penelitian ini

TABEL 8
HASIL FINAL

Keterangan	Judul	Penulis	Abstrak	Kata Kunci
Score	1761	1873	1862	1933
Score Penuh	2160	2160	2160	2160
Percent	81,53	86,71	86,20	89,49
Rule	7	3	11	7
Total Rule	28	28	28	28
Percent Bobot	20,38	9,29	33,87	22,37
			TOTAL	**85,91**

Konsep perhitungan yang disajikan adalah pesentase yang didapatkan dihitung dengan bobot jumlah rule yang mempengaruhi masing-masing elemen, sehingga elemen yang rumit menjadi penentu porsi paling besar dalam perhitungan. Dalam Tabel 6.8 dapat disimpulan ketelitian yang dicapai pada penelitian ini adalah sebesar 85,91 %. Hipotesa dari penelitian ini adalah ketelitian mencapai angka 80 %. Sehingga disimpulkan bahwa penelitian ini memenuhi kriteria pada hipotesa awal.

VI. KESIMPULAN

Setelah melakukan proses penelitian yang panjang dan menemukan beberapa solusi. Penelitian yang diawali dengan ide dan terealisasi dalam system yang dikembangkan. Kemudian juga dilakukan pengujian. Hal tersebut menghasilkan beberapa kesimpulan diantaranya adalah:

1. Penulisan paper / karya ilmiah memiliki format yang bisa di generalisasi dan memiliki kemiripan yang signifikan dari masing-masing penerbit. Hal ini menjadi penunjang yang baik untuk penyusunan rule base system karena system tidak dituntut untuk terus menyesuaikan dengan perubahan yang sangat dinamis.
2. Selain format penulisan urutan dari masing-masing elemen paper yang memiliki aturan yang general juga sangat membantu dalam proses ekstraksi informasi pada system.
3. Pengambilan informasi judul, abstrak, dan kata kunci menjadi pilihan yang sangat tepat untuk proses clustering dalam pencarian kesamaan isi dari paper. Pada 3 informasi tersebutlah tersimpan ini dan isi dari suatu paper.
4. Kumpulan Rule dan Algoritma yang disusun oleh pakar dengan menggunakan Attribut Font dan juga Susunana Paragraf terbukti mampu mengekstraksi informasi untuk judul, abstrak, katakunci dan penulis dengan baik.
5. Fitur pencarian kata kunci juga menjadi penunjang yang sangat baik untuk mengenali bagian-bagian dari paper.
6. Deteksi Penulis dengan hanya menggunakan NER tidak akan pernah maksimal karena tata cara penulisan bagian penulis sangat bervariasi font attributnya sehingga mengakibatkan menjadi lebih rumit untuk di gabungkan menjadi 1 kesatuan
7. Subsistem konversi PDF dengan mengoptimasi pengenalan paragraph dengan menggunakan Attribut Font terbukti membantu mempermudah proses pengenalan element dan ekstraksi informasi
8. Subsistem Ekstraksi Informasi menggunakan rule dapat menjadi sebuah solusi yang cukup baik untuk ekstraksi informasi pada paper karena sudah memilik struktur yang cukup baku, dengan ketelitian yang berhasil dicapai adalah 82.17%
9. Alat bantu PDFBox yang digunakan sangat membantu dalam proses konversi file sumber dengan ekstensi .pdf menjadi data yang bisa diolah lebih baik.
10. Encoding yang digunakan untuk penyimpanan pada database harusnya UTF-8 karena pada paper terdapat karakter-karakter khusus yang tentunya akan mengganggu proses penyimpanan.
11. Format JSON yang bisa di sematkan dalam bentuk object sangat membantu dalam proses ekstraksi informasi karena proses pencarian tidak diperlukan lagi query ke database namun bisa langsung dilakukan pada level program dengan memanfaatkan data yang sudah berbentuk JSON.

DAFTAR PUSTAKA

[1] L. Chiticariu, Y. Li, and F. R. Reiss, “Rule-based information extraction is dead! Long live rule-based information extraction systems!,” 2013.

[2] E. Lim, E. I. Setiawan, and J. Santoso, “Stance Classification Post Kesehatan di Media Sosial Dengan FastText Embedding dan Deep Learning,” *J. Intell. Syst. Comput.*, vol. 1, no. 2, pp. 65–73, 2019.

[3] M. A. Rahman, H. Budianto, and E. I. Setiawan, “Aspect Based Sentimen Analysis Opini Publik Pada Instagram dengan Convolutional Neural Network,” *J. Intell. Syst. Comput.*, vol. 1, no. 2, pp. 50–57, 2019.

[4] S. N. Soenardjo and G. Gunawan, “Information Extraction Berbasis Rule Untuk Soal Ujian,” *J. Intell. Syst. Comput.*, vol. 2, no. 1, pp. 28–33, 2020.

[5] K. Yao, “Header Extraction from Scientific Documents.”

[6] J. Beel, B. Gipp, A. Shaker, and N. Friedrich, “SciPlore Xtract: extracting titles from scientific PDF documents by analyzing style information (Font Size),” in *International Conference on Theory and Practice of Digital Libraries*, 2010, pp. 413–416.

[7] J. Beel, S. Langer, M. Genzmehr, and C. Müller, “Docear’s PDF Inspector: Title Extraction from PDF Files,” in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2013, pp. 443–444, doi: 10.1145/2467696.2467789.

[8] D. Meyerzon, Y. Cao, H. Li, Q. Zheng, and Y. Hu, “Automatic extraction of titles from general documents using machine learning,” in *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL ’05)*, 2005, pp. 145–154, doi: 10.1145/1065385.1065418.

[9] L. Kovriguina, A. Shipilo, F. Kozlov, M. Kolchin, and E. Cherny, “Metadata extraction from conference proceedings using template-based approach,” in *Communications in Computer and*

Information Science, 2015, vol. 548, doi: 10.1007/978-3-319-25518-7_13.

[10] F. Peng and A. McCallum, “Information extraction from research papers using conditional random fields,” *Inf. Process. \& Manag.*, vol. 42, no. 4, pp. 963–979, 2006.