PENGEMBANGAN PAPER CITATION EXTRACTION BAHASA INDONESIA BERBASIS PARSCIT

Resmana Lim [1] , Adi Wibowo [2], Raymond Sutjiadi, Yustus Eko Oktian

[1] Teknik Elektro, Universitas Kristen Petra
Jl. Siwalankerto 121-131, Surabaya 60236 Indonesia
email : resmana@petra.ac.id
[2] Teknik Informatika, Universitas Kristen Petra
Jl. Siwalankerto 121-131, Surabaya 60236 Indonesia
email : adiw@petra.ac.id

ABSTRACS

Indonesian Researchers has been struggling to find paper in Indonesian language that can be used as a reference source. This is due to the lack of Indonesian citation data center that store and link the related papers. This study tried to develop system that can extract citations based on ParsCit that modified to recognize citation in Indonesian languange.As the research that is being developed, the system can recognize the citations in Indonesian language with a good level of accuracy. The study should be develop further to establish the chain of citation between related papers.

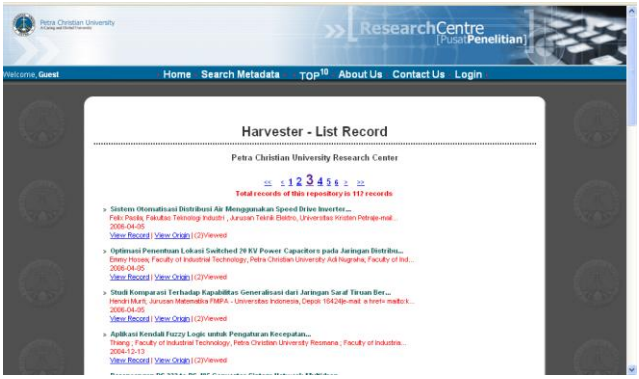
Key words

Citation extraction, Citation Database, ParsCit

1. Pendahuluan

Lembaga penelitian/perguruan tinggi di Indonesia masih belum memilki budaya sharing informasi bagi sesamanya. Mereka lebih banyak menjadi konsumen informasi yang banyak didapatkan dari penyedia informasi luar negeri. Hal ini disebabkan para penulis kesulitan mencari referensi artikel dalam negeri dibandingkan mencari artikel terbitan luar negeri.

Pada penelitian sebelumnya [1], penulis berupaya untuk membangun sebuah situs untuk mengumpulkan database metadata artikel ilmiah. Pengumpulan metadata artikel menggunakan protokol Open Archive Initiative (OAI). Metadata berasal dari berbagai sumber e-journal berbagai perguruan tinggi. Dengan grant penelitian hibah bersaing (PHB) [2], penulis menambahkan fungsi pertukaran metadata pada e-journal tersebut. Metadata tersebut dapat di-*harvest* atau diunduh menggunakan protokol OAI menjadi kumpulan abstraksi artikel jurnal. Sebagai hasil PHB tahun ke 2, penulis mengembangkan metadata harvester artikel jurnal ilmiah [3] seperti pada Gambar 1. Kendati demikian, sistem ini belum menyertakan kemampuan penghitungan sitasi dan inter-linking antar artikel yang saling mensitasi (*citing* & *get cited*).



Gambar 1 Daftar Artikel dalam Jurnal

Untuk mendapatkan kemampuan penghitungan sitasi dan interlinking antar artikel berbasis sitasi dibutuhkan proses *citation indexing*. Pada pengembangan lanjutan, penulis mencoba untuk menambahkan kemampuan citation index sederhana [4] pada sistem e-journal yang ada pada alamat http://puslit2.petra.ac.id/ejournal. Sistem e-journal ini memiliki koleksi 41 jurnal dengan total lebih dari 1.700 artikel full-text di dalamnya. Jurnal berasal dari berbagai perguruan tinggi antara lain UK Petra, Universitas Hang Tuah, STIKOM, Universitas Soegijapranata-Semarang dan Universitas Widya Kartika dan beberapa lagi akan menyusul. Kendati demikian

proses citation indexing artikel masih menggunakan metode semi atomatis. Informasi sitasi diekstrak dari file *full-text* pdf, dan apabila kurang sempurna maka dilakukan penyesuaian manual dengan mengedit satu per satu (*manual data entry/editing*) ke database. Cara seperti ini sangat tidak efektif untuk jumlah artikel dengan volume tinggi. Dibutuhkan suatu mekanisme dengan derajat otomatis lebih tinggi dalam ekstraksi sitasi dari full-text (file PDF) masing-masing artikel ilmiah.

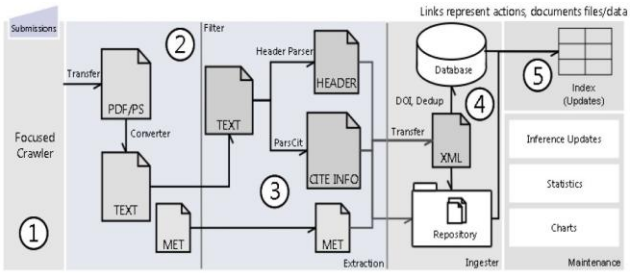
Upaya membuat citation index yang dapat diakses secara bebas di Internet dipelopori oleh Citeseer [5] sejak tahun 1998. Saat ini terdapat lebih dari 1,3 juta artikel dengan jumlah sitasi lebih dari 23 juta. Kendati demikian Citeseer (http://citeseer.ist.psu.edu) lebih dikhususkan pada bidang Computer Science, sehingga jarang menemukan sitasi dari artikel diluar bidang ilmu komputer. Ia melakukan citation indexing secara otomatis terhadap full-text artikel. Citeseer sangat baik melakukan citation indexing untuk artikel berbahasa inggris karena memang algoritmanya dirancang untuk dapat menganalisa artikel berbahasa Inggris. Proses indexing otomatis Citeseer menggunakan pengecekan frasa-frasa dalam bahasa Inggris. Citeseer belum *'friendly'* terhadap artikel ilmiah dalam bahasa Indonesia. Oleh karena itu masih terdapat ruang untuk mengembangkan sistem serupa namun dengan kemampuan analisa sitasi dalam bahasa Indonesia, sekaligus dapat mengakomodasi seluruh bidang ilmu.

2. ParsCit Citation Extraction

Komponen utama dalam sistem ini adalah modul program untuk melakukan ekstraksi sitasi dari input file paper ilmiah berformat PDF. Dari studi referensi yang telah dilakukan, penulis menemukan sebuah modul bernama ParsCit yang terbukti baik dan telah diadopsi oleh beberapa system lainnya misalnya Mendeley.com. ParsCit dikembangkan oleh Information Science and Technology (IST) Penn State University dan National University of Singapore (NUS). Software ini dibuat oleh Min-Yen Kan, Isaac G. Councill, C. Lee Giles, dan Minh-Thang Luong pada tahun 2008 [5]. Software ini digunakan untuk mengolah sebuah paper atau journal elektronik dengan tujuan untuk mengambil informasi-informasi penting secara otomatis dari journal atau paper seperti, judul, penulis, institusi, alamat, tahun, email, abstarak, kata kunci, dan sitasi (daftar referensi). ParsCit efektif untuk mengekstrak sitasi dalam bahasa Inggris, namun gagal untuk ekstrak paper berbahasa Indonesia oleh sebab itu dalam penelitian ini dilakukan modifikasi terhadap ParsCit agar dapat mengekstrak sitasi paper berbahasa Indonesia.

Dalam melakukan proses pengambilan informasi-informasi tersebut, ParsCit menggunakan Conditional Random Field (CRF) sebagai metode pengambilan data dari paper. Selain ParsCit juga terdapat software atau metode-metode lain yang bisa digunakan untuk melakukan pengambilan data tersebut yang sejenis dengan ParsCit, misalnya ParaCite, sebuah software metadata extraction yang menggunakan Perl dan algoritma Regular Expression (Jewell, Michael,) dan FreeCite, sebuah alat yang digunakan untuk memecah-mecah bagian-bagian dari paper berbasis Ruby dengan menggunakan CRF++ library dan CORA dataset [5]. Kendati demikian, penulis memilih ParsCit karena performancenya lebih baik dan kemudahan untuk memodifikasi kode programnya.

Cara kerja system ekstraksi sitasi adalah digambarkan seperti pada diagram blok pada Gambar 2 [5].



Gambar 2 Cara Kerja Sistem ParsCit

Hal pertama yang dilakukan oleh sistem adalah melakukan crawling data PDF dari URL yang dimasukkan oleh user. Setelah itu file PDF tersebut disimpan ke dalam sistem dan dilakukan proses konversi menjadi text. Data text tersebut kemudian akan diekstrak dengan menggunakan ParsCit untuk menghasilkan data-data header dan referensi. Data hasil ekstrak tersebut kemudian akan disimpan ke dalam database. Dari dalam database tersebut paper-paper yang memiliki hubungan di dalam referensinya akan dihubungkan satu dengan yang lainnya sehingga setiap paper akan memiliki relasi yang erat satu dengan yang lainnya.

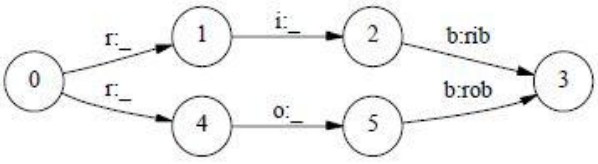
Paper diekstrak berdasarkan model yang dimiliki oleh sistem. Model tersebut didapatkan melalui proses training untuk menghasilkan hasil ekstraksi yang lebih baik. Data-data training yang telah disiapkan akan diproses oleh CRF menghasilkan data model yang baru. Data model yang baru ini akan menggantikan data model yang lama dan kemudian digunakan oleh sistem untuk mengekstrak sebuah paper.

2.1 Conditional Random Field (CRF)

Conditional Random Field adalah sebuah metode yang digunakan untuk melakukan proses segmentasi data

berdasarkan probabilitas. Ide dari CRF ini adalah menggunakan probabilitas untuk melakukan distribusi terhadap pola label berdasarkan pola hasil observasi. Ide yang telah ada sebelum munculnya CRF ini yaitu Hidden Markov Model (HMM), tidak menggunakan sistem probabilitas, hal ini membuat CRF mempunyai keunggulan dibandingkan HMM yaitu CRF mempunyai conditional nature yang menghasilkan asumsi-asumsi yang independen dan juga CRF dapat menghilangkan masalah label bias yang dimiliki oleh HMM [6].

John Lafferty, Andrew McCallum, dan Fernando Pereira [7] membuat sebuah penelitian yang membahas masalah tersebut. Misalkan ada 2 buah kata yang akan dilakukan proses segmentasi yaitu kata rob dan rib, seperti terlihat pada Gambar 3. Sistem mendeteksi huruf pertama adalah sama yaitu 'r' pada state 1 dan 4, kemudian sistem medeteksi lagi huruf 'i' pada state 2 dan 'o' pada state 5, setelah itu sistem kembali mendeteksi huruf yang sama yaitu 'b' pada state 3. Pada kata yang hampir mirip seperti pada contoh di atas, rawan terjadi kesalahan dalam proses segmentasi sehingga kata rib dan rob dianggap sebuah kata yang sama.



Gambar 3 Contoh dari *Label Bias*

2.2 Training Model

Secara umum, cara kerja ParsCit dimulai dengan munculnya data referensi yang handak kita lakukan proses segmentasi untuk mendapatkan informasi-informasi yang ada didalamnya. Hal yang dilakukan adalah memecah-mecah data dari referensi tersebut menjadi beberapa bagian yang dipisahkan dengan token misalnya{r1, r2, ..., rn}. Kemudian setiap token itu akan dimasukkan ke dalam kelas-kelas seperti {c1, c2, ..., cm}. Setiap token yang telah terbentuk akan dimasukkan secara tepat ke dalam kelas sesuai dengan kriteria label yang telah ditentukan. Gambar 4 menjelaskan gambaran bagaimana proses pemecahan data referensi tersebut terjadi.

Secara keseluruhan ParsCit menggunakan beberapa label untuk mengumpulkan informasi-informasi yang didapatkan, yaitu:

a. Author, yaitu pengarang dari sebuah buku, jurnal, atau artikel.
b. Title, yaitu judul dari artikel atau issue.
c. Journal, yaitu judul jurnal yang digunakan di dalam referensi
d. Booktitle, merupakan judul buku yang digunakan.
e. Publisher, nama penerbit dari buku atau jurnal.
f. Location, merupakan tempat terbit dari buku atau jurnal.
g. Date, yaitu tanggal terbit dari buku atau jurnal
h. Tech, keterngan mengenai jurnal
i. Institution, lembaga yang terlibat di dalam referensi
j. Note, keterangan tambahan yang ada di dalam referensi.
k. Editor, yaitu nama dari editor yang ada.
l. Pages, menunjukkan halaman yang dikutip
m. Volume, menunjukkan seri dari sebuah jurnal atau buku.

2.3 Pre-processing

Setelah data referensi berhasil dikelompokkan ke dalam masing-masing label, langkah berikutnya dinamakan dengan Pre-Processing, yaitu mencari bagian referensi atau daftar pustaka di dalam sebuah paper. ParsCit akan mencari bagian Referensi dari paper tersebut. Di dalam format Institute of Electrical and Electronics Engineers (IEEE) biasanya setiap baris dari referensi ditandai dengan angka [1], [2], dst. Sistem dari ParsCit dapat mengenali format tersebut dan mengambil data dari referensi tiap barisnya. Bila tidak ada angka atau penanda yang menunjukan baris dari riap-tiap referensi tersebut, seperti pada format American Psychological Association (APA), maka beberapa pendekatan perlu dilakukan untuk menentukan awal dan akhir dari masing-masing baris referensi. Biasanya, kata-kata yang dideteksi sebagai pengarang terletak pada awal baris, sedangkan tanda baca biasanya terletak di akhir baris. [7]. Sistem perlu melakukan proses training terlebih dahulu agar dapat mengenali pola tersebut dengan baik. Setelah data dari masing-masing baris diketahui, maka selanjutnya dilakukan pengaplikasian CRF dengan

<author> A. Cau, R. Kuiper, and W.-P. de Roever. </author> <title> Formalising Dijkstra's development strategy within Stark's formalism. </title> <editor> In C. B. Jones, R. C. Shaw, and T. Denvir, editors, </editor> <booktitle> Proc. 5th. BCS-FACS Refinement Workshop, </booktitle> <date> 1992. </date>

<author> M. Kitsuregawa, H. Tanaka, and T. Moto-oka. </author> <title> Application of hash to data base machine and its architecture. </title> <journal> New Generation Computing, </journal> <volume> 1(1), </volume> <date> 1983. </date>

<author> Alexander Vrchoticky. </author> <title> Modula/R language definition. </title> <tech> Technical Report TU Wien rr-02-92, version 2.0, </tech> <institution> Dept. for Real-Time Systems, Technical University of Vienna, </institution> <date> May 1993. </date>

Gambar 4 Contoh dari *Label Bias*

menggunakan CRF++ seperti yang telah dijelaskan sebelumnya.

2.4 Post-processing

Langkah berikutnya yang dilakukan dinamakan dengan Post-Proccessing. Data per baris yang telah dilakukan proses segmentasi akan terkumpul sesuai dengan label yang ada (mis, author, title, dll.). Pada label author, data yang terkumpul masih tercampur misalnya "Hermanto, Raymond, dan Susan Ratilee". Pada data seperti ini dilakukan proses naturalisasi sehingga data menjadi "Raymond Hermanto" dan " Susan Ratilee". Contoh yang lainnya adalah pada label volume biasanya ditulis "Vol. 5", data seperti ini akan dilakukan proses naturalisasi menjadi "5". Pada label date akan dilakukan proses naturalisasi sehingga data yang diambil hanyalah data tahun, misalnya kita ada data "Maret 2010" maka yang akan tersimpan adalah data "2010" [7].

3. Citation Extraction Berbasis ParsCit untuk Bahasa Indonesia

Penulis membuat website untuk percobaan ekstrasi sistasi dari input paper ilmiah berformat pdf. Beberapa langkah yang telah dilakukan adalah:

a. Melakukan instalasi parscit-110505b pada server berbasis Linux.
b. Mengedit (modifikasi) sistem parsCit agar bisa mendeteksi paper-paper yang berbahasa Indonesia.
c. Memasukkan data-data referensi dalam bahasa Indonesia dan melakukan proses training ulang parsCit.
d. Mengedit sistem ParsHed agar bisa mendeteksi paper-paper yang berbahasa Indonesia.
e. Memasukkan data-data inti paper seperti, author, affiliation, abstract, keyword dalam bahasa Indonesia dan melakukan proses training ulang parsHed.
f. Mengedit file index.php agar user bisa memasukkan entri daftar referensi yang ingin dideteksi oleh parscit.
g. Membuat database dalam phpmyadmin, mengedit file index.php agar daftar hasil ekstraksi dari parsCit bisa dimasukkan ke dalam database.
h. Membuat fungsi untuk menghapus header dari paper.

3.1 Mengedit (modifikasi) sistem parsCit agar bisa mendeteksi paper-paper yang berbahasa Indonesia

Sistem ParsCit masih belum bisa mendeteksi paper ilmiah yang ditulis dengan menggunakan Bahasa Indonesia. ParsCit hanya bisa mengenali dengan baik untuk paper berbahasa Inggris, karena ParsCit sebenarnya dikembangkan untuk paper berbahasa Inggris. Ketika dicoba untuk memasukkan paper berbahasa Indonesia, ParsCit tidak bisa mengenali tiap-tiap bagian dari paper seperti contohnya Abstrak, Kata kunci, Pendahuluan, Daftar Pustaka. ParsCit mengenali paper berbahasa Inggris yang biasanya di dalam paper dituliskan sebagai Abstract, Keywords, Introduction, References.

Agar ParsCit bisa mengenali paper berbahasa Indonesia maka dilakukan proses pembelajaran untuk mencari tahu bagaimana cara kerja dari ParsCit tersebut. ParsCit menggunakan library dalam pengenalan kata-kata kunci dalam paper. Kode program dari library perlu diubah seperti pada Gambar 5 dan 6 agar bisa mengenali kata-kata dalam Bahasa Indonesia. Pada Gambar 5 ParsCit menggunakan kata-kata References, References and Notes, Cited References, dan Bibiolgraphy untuk mengenali bagian dari Daftar Pustaka pada sebuah paper. Tulisan Besar atau kecil juga mempengaruhi kinerja dari program tersebut (Case Sensitive). Bibliography dan BIBLIOGRAPHY akan dikenali sebagai kata yang berbeda oleh sistem ParsCit karena itu semua kemungkinan besar atau kecilnya huruf harus ditulis juga pada program tersebut. Pada Gambar 6 ditambahkan istilah-istilah sejenis dalam bahasa Indonesia, yaitu Daftar Pustaka, Daftar Referensi, Referensi, dan Daftar Kutipan.

```
while ($text =~ m/\b(References?|REFERENCES?|Bibliography|BIBLIOGRAPHY|
References?\s+and\s+Notes?|References?\s+Cited|REFERENCES?\s+CITED|
REFERENCES?\s+AND\s+NOTES?|LITERATURE?\s+CITED?):?\s*\n+/sg)
{
$bodytext = substr $text, 0, pos $text;
$citetext = substr $text, pos $text unless (pos $text < 1);
}
```

Gambar 5 Kode Asli Program PreProcess dari ParsCit

```
while ($text =~ m/\b(Referensi?|REFERENSI?|References?|REFERENCES?|DaftarPustaka|
DAFTARPUSTAKA|Daftar\s+Pustaka|Daftar\s+Referensi|DAFTAR\s+REFERENSI|DAFTAR\s+PUSTAKA|
Bibliography|BIBLIOGRAPHY|References?\s+and\s+Notes?|References?\s+Cited|REFERENCES?\s+CITED|
REFERENCES?\s+AND\s+NOTES?|Daftar\s+Kutipan?|DAFTAR\s+KUTIPAN?):?\s*\n+/sg)
{
$bodytext = substr $text, 0, pos $text;
$citetext = substr $text, pos $text unless (pos $text < 1);
}
```

Gambar 6 Kode Program yang Diubah dari ProProcess

3.2 Mengedit sistem ParsHed agar bisa mendeteksi paper-paper yang berbahasa Indonesia.

Sistem ParsHed di dalam ParsCit mengenali kata-kata Introduction, apabila sistem mendeteksi kata Introduction tersebut pada saat sistem sedang memproses data-data header maka data-data yang didapatkan setelah kata-kata Introduction tersebut tidak akan diambil oleh sistem dan dibuang sehingga hal ini membuat data header akan berhenti sampai pada data kata kunci karena biasanya

data kata kunci terletak sebelum bab pertama ditulis. Untuk dapat mengenali paper yang ditulis dengan menggunakan Bahasa Indonesia maka kode program tersebut harus disesuaikan dengan cara menambahkan kata-kata Pendahuluan pada kode program tersebut seperti yang terlihat pada Gambar 7. Dengan ditambahkannya kode tersebut maka sistem ParsHed dapat mengenali dua bahasa paper yaitu Bahasa Inggris dan Bahasa Indonesia. Untuk paper bahasa Inggirs, sistem mengenali kata-kata Introduction sedangkan untuk paper bahasa Indonesia, sistem mengenali kata-kata Pendahuluan, ini terjadi karena didalam kode program tersebut digunakan logika atau. Logika atau ini juga dipakai dalam kode program yang digunakan untuk mengolah referensi. Sehingga sekarang sistem ParsCit secara keseluruhan dapat mengenali bagian-bagian dari paper berbahasa Inggris dan Indonesia.

```
else
    {
      # sample RE for header stop.
      #edited by Yustus 29 Juni 2012
          if (/INTRODUCTION/i|/PENDAHULUAN/i) { last; }

          if($is_token_level)
      {
        $buf .= "$_";
        $buf .= " +L+ ";
          }
      else
      {
        $buf .= "$_\n";
          }

      }
```

Gambar 7 Kode Program yang Diubah dari ParsHed

3.3 Proses Training Data References dan Header

Untuk melakukan proses *training* ParsCit, sistem memerlukan sebuah data referensi sebagai bahan untuk dilakukan proses *training*. Data referensi tersebut disimpan oleh sistem di dalam direktori traindata dengan nama file tagged_references.txt yang ditunjukkan pada Gambar 8. Sistem juga membutuhkan file tagged_headers.txt sebagai data header dalam proses training ParsHed seperti ditunjukkan pada Gambar 9.

```
<author> Gardner, J.K., and Knopoff, L., </author> <title> Is the Sequence
of Earthquakes in Southern California, with Aftershocks removed, Poissonian,
</title> <journal> Bulletin of the Seismological Society of America,
</journal> <volume> Vol. 64, No. 5, </volume> <date> 1974, </date> <pages>
pp. 1363-1367. </pages>
```

Gambar 8 Data Training References

Data dalam tagged_headers.txt dipisahkan berdasarkan label seperti pada Gambar 9. Informasi judul dari artikel dipisahkan dengan menggunakan label <title> dan </title>, informasi pengarang dipisahkan dengan menggunakan label <author> dan </author>, dsb. Jika dalam data tersebut berpindah baris dari baris pertama menuju ke baris kedua maka di dalam penulisannya pada file tagged_headers.txt diberi tanda +L+, tanda tersebut menunjukkan bahwa teks berakhir dan dilanjutkan pada baris berikutnya. Tanda ini sangat penting untuk ditulis karena hal tersebut juga merupakan sebuah pola. Judul yang ditulis dalam dua baris tentu saja berbeda dengan judul yang ditulis hanya satu baris. Sedangkan bila data tersebut berpindah ke halaman berikutnya pada paper maka penulisannya diberi tanda +PAGE+. Tanda tersebut menunjukkan bahwa teks telah berakhir pada halaman tersebut dan dilanjutkan pada halaman berikutnya.

3.3 Menghilangkan Header Note dari Paper

Dalam sebuah paper ada kalanya memiliki header note yang dituliskan di dalam paper tersebut. Penulisan header note sebenarnya bukan merupakan suatu ketentuan umum yang harus ada di dalam sebuah paper. Namun hanyalah sebuah keterangan yang dituliskan untuk mendukung kemudahan dalam mengakses sebuah paper di dalam journal. Header note ini biasanya berisi

```
<title> PERBANDINGAN TINGKAT KEPUASAN KELUARGA PASIEN GAWAT DARURAT +L+ DAN
GAWAT NON DARURAT TERHADAP MUTU PELAYANAN KESEHATAN DI +L+ UGD RS. BAPTIS
BATU +L+ </title> <author> Erlin Kurnia +L+ Diba Yusanto +L+ </author>
<abstract> ABSTRACT +L+ Background : The differences of the satisfaction
level among emergency and non emergency +L+ patient family become one of
important problems in Emergency Unit Batu Baptist Hospital. The +L+ purpose
of this research is for comparing the satisfaction level among emergency and
non +L+ emergency patients family to the health service quality given by
Emergency Unit at Batu Baptist +L+ Hospital. +L+ Method : this research was
comparative study. The population are all of the patient family +L+ in
Emergency Unit. The quantity of the sample are 100 respondences that was
taken by accidental +L+ sampling. The independent variable is health service
quality. The dependent variable was +L+ satifaction level of emergency and
non emergency patient is family.The data was analized by +L+ â€œWilcoxon
Testâ€ with mean level ï¡ ï‚£ 0,05. The result of this research showed
that there was no +L+ difference of satisfaction level between emergency and
non emergency patients family. The result +L+ of statistic test â
€œWilcoxonâ€ is p = 0,465 and p ï‚³ ï¡ , so that HO is accepted. +L+
Conclusion : The conclusion of this research is the same satisfaction
between emergency and +L+ non emergency patient family. It is very satisfy
result so that the health service quality at +L+ Emergency Unit in Batu
Baptist Hospital have been suitable with the patient and customer hopes. +L+
</abstract> <keyword> Key word : The satisfaction of emergency and non
emergency patient is family, health +L+ service quality, Emergency Unit. +L+
</keyword> <intro> PENDAHULUAN +L+ </intro>
```

Gambar 9 Data Training ParsHed

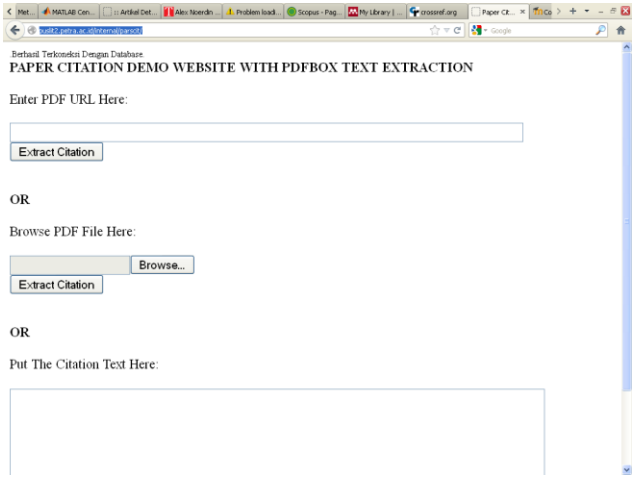
keterangan umum dari paper yang bersangkutan misalnya nama pengarang dan judul atau bisa juga berisi keterangan umum dari journal yang menerbitkan paper tersebut misalnya judul dan volume journal. Manusia bisa membedakan manakah bagian dari paper yang merupakan header note dan manakah bagian isi dari paper. Namun hal tersebut tidak berlaku oleh sistem PDFBOX yang dimiliki oleh sistem.

```
function hapus_header()
{
  $file = "uploads/proses2.txt";
  $text = file($file) or die("could not read file");
  $unique = array_unique($text);
  $handle = fopen($file, "w") or die("could not open file for writing");
  foreach($unique as $line)
  {
    fwrite($handle, $line);
  }
  fclose($handle);
}
```

Gambar 10. Kode Program untuk Menghapus Header Note.

4. Analisa Hasil Sistem Ekstraksi Sitasi

Telah dibuat demo *Website* untuk menampilkannya hasil ekstraksi dengan alamat http://puslit2.petra.ac.id/int



Gambar 11. Website Demo Ekstraksi Sitasi

ernal/parscit/. *Website* ini menjadi modal penting sebagai alat yang menjembatani antara *user* dengan sistem seperti gambar 11 dan 12.

Sistem ParsCit telah berjalan dengan lebih baik daripada sebelumnya. Saat ini ParsCit telah bisa mengenali paper

CITATION(S) :

Marker	Name	Title	Date	Journal	Volume	Issue	Pages	Book Title	Publisher	Location	Editor	Institution	Tech	Note
[1]	Soebagio	Model mesin AC pada koordinat d-qn,	2006									ITS,		Materi Kuliah Mesin Listrik Lanjut,
[2]	D Casadei G Serra A Tani L Zarri	Assessment of direct torque control for induction motor drivesâ€, Buletin of the Polish academy of science tech. sciences,	2006		54									
[3]	Bose	Modern Power Electronics and AC drives,	2002						Prentice Hall PTR,					
[4]	Sri Kusumadewi	Membangun Jaringan Syaraf Tiruan Menggunakan Matlab dan Excellink,	2004		1				Graha Ilmu,	Yogyakarta				
[5]		Mauridhi Hery Purnomo dan Agus Kurniawan, Supervised Neural Networks dan Aplikasinya,	2006						Graha Ilmu,	Yogyakarta :				
[6]	A Damiano P Vas etal	Comparison of speed sensorless DTC induction motor drivesâ€,	1997				1--11	Proc. PCIM,		Nuremberg, Germany,				
[7]	D Casadei Giovanni Serra	FOC and DTC: two variable scheme for induction motors torque controlâ€,	2002	Trans. On Power Electronics,	17									

Gambar 12. Contoh Hasil Ektraksi Sitasi

berbahasa Indonesia dan Inggris. ParsCit juga semakin baik kinerjanya dalam mendeteksi bagian-bagian header

dan referensi. Selama proses proyek penelitian ini, telah ditambahkan sekitar 500 data referensi berbahasa Indonesia ke dalam tagged_references.txt untuk memperkuat hasil training dari ParsCit. Selain itu juga ditambahkan sekitar 40 data header berbahasa Indonesia ke dalam tagged_headers.txt untuk memperkuat hasil training dari ParsHed. Dengan ditambahkannya data-data baru tersebut, sistem menjadi lebih baik kinerjanya saat mendeteksi bagian-bagian dari header dan referensi. Saat ini ParsCit telah bisa mendeteksi paper-paper dalam website Puslit dengan baik. Sistem ParsCit memiliki beberapa keunggulan yaitu:

a. ParsCit mendeteksi pola dari header dan referensi dengan menggunakan CRF sehingga kita tidak perlu memberikan banyak data training ke dalam sistem. Untuk suatu pola data yang sama cukup hanya diberikan satu sample data training maka sistem akan dapat mendeteksi pola lain yang sejenis.
b. Sistem dibagi menjadi dua bagian yaitu ParsHed dan ParsCit yang bekerja secara independen. Hal ini membuat sistem menjadi sangat diandalkan ketika terjadi kesalahan dalam salah satu bagian sistem tidak akan mempengaruhi bagian sistem yang lain. Selain itu dengan pemisahan ini maka dapat dipilih apakah ingin menggunakan sistem ParsCit untuk mendeteksi data header saja atau untuk mendeteksi referensi saja atau bahkan keduanya.

Kelemahan dari ParsCit adalah:

a. Jika di dalam penulisan referensi di dalam paper tidak menggunakan marker, maka hasil dari pengolahan referensi menjadi semakin banyak menghasilkan data yang salah. ParsCit mendeteksi dengan lebih baik bila penulisan referensi dalam sebuah paper diberi marker.
b. Jika dalam penulisan paper ada pola referensi yang salah ketik misalnya, setelah koma tidak diberi spasi, atau setelah titik tidak diberi spasi, seperti pada Gambar 3.20, maka keluaran yang dihasilkan akan menghasilkan data yang salah. Tahun 1990 seharusnya ditulis dengan diberi spasi terlebih dahulu. Jika tidak diberi spasi maka tahun tersebut masuk ke dalam data publisher. Ini merupakan data yang salah. Data yang benar ketika penulisan tahun 1990 diberi spasi terlebih dahulu
c. Jika user hendak menggunakan ParsCit dengan cara melakukan upload file pdf maka nama file tidak boleh ada spasi, contohnya jika nama file adalah "paper ilmiah.pdf" maka sistem ParsCit tidak bisa mendeteksi file tersebut, proses pdf_to_text tidak bisa berjalan. Seharusnya nama file tersebut diubah menjadi "paper_ilmiah.pdf".
d. Jika terjadi kesalahan ketik di dalam penulisan paper misalnya kata-kata Daftar Referensi salah ketik menjadi Daftar Rererensi, maka sistem ParsCit tidak bisa mendeteksi kata-kata tersebut dan referensi pun menjadi tidak bisa diolah oleh sistem.
e. Jika di dalam sebuah paper terdapat lampiran. Maka lampiran tersebut tidak dapat dideteksi oleh ParsCit sebagai sebuah lampiran. Biasanya lampiran diletakkan pada akhir dari sebuah paper, biasanya setelah bagian Daftar Pustaka dari sebuah paper. ParsCit hanya bisa mendeteksi kata Daftar Pustaka dan menganggap bahwa bagian setelah kata tersebut adalah bagian referensi sehingga hal ini membuat lampiran akan masuk ke dalam data referensi dan membuat kekacauan pada data referensi sehingga ParsCit akan menghasilkan data yang salah.

4. Kesimpulan

Modifikasi terhadap ParsCit telah mempu mengekstraksi paper paper berbahasa indonesia. Dengan penambahan data referensi dan header membuat modul ParsCit semakin baik ketika mendeteksi data-data dari artikel berbahasa Indonesia. Keberhasilan ekstraksi sangat bergantung dengan data-data training yang ditambahkan ke dalam sistem tersebut. Modul ekstraksi sitasi telah banyak ditambahkan dengan data-data paper yang berasal dari website e-journal Puslit UK Petra (sebagai model), sehingga sistem tidak mengalami masalah ketika mendeteksi paper-paper yang berasal dari website tersebut karena sistem sudah memahami format penulisan atau pola yang dimiliki oleh paper-paper tersebut. Masalah terjadi ketika Modul ekstraksi dihadapkan dengan paper yang belum pernah diketahui polanya, sehingga sistem akan mendeteksi paper tersebut dengan kesalahan. Namun masalah tersebut bisa diatasi dengan menambahkan data training baru untuk paper yang belum terdeteksi dengan baik tersebut.

DAFTAR PUSTAKA

[1] Resmana Lim, Tonny Prawiro, Sayoga Senior, Poedi Soenarjo Wartono, 2008, "Pengembangan Data Provider Jaringan Repositori Digital Antar Lembaga Penelitian Memanfaatkan Teknologi Open Archives", Seminar Nasional Sistem dan Aplikasi Teknologi Informasi, 22 Oktober 2008

[2] Resmana Lim, Tonny Prawiro, Sayoga Senior, Poedi Soenarjo Wartono, 2007, "Pengembangan Situs Web Jaringan Repositori Digital Antar Lembaga Penelitian Memanfaatkan Teknologi Open Archives", Laporan Penelitian Hibah Bersaing, Universitas Kristen Petra.

[3] Iwan Handoyo Putro, Resmana Lim, Rocky Y. Dillak, 2009, "Aplikasi Web Direktori Jurnal Menggunakan Feature Harvester Metadata Artikel". SENTIA 2009. Malang, 12 Maret 2009

[4] Resmana Lim, Bernard A.D., 2008, "Pengembangan Aplikasi Citation Indexer Sederhana dengan fitur Semi Automatic pada E-Journal UK Petra", Project Report, Universitas Kristen Petra.
[5] Councill, Isaac G., Giles, C. Lee., Kan, Min-Yen., 2008, "ParsCit: An open-source CRF reference string parsing package", Proceedings of LREC, pp. 661-667
[6] Charles M. Sutton, Andrew McCallum, 2002, "An Introduction to Conditional Random Fields for Relational Learning". Graphical Models, pp. 93
[7] Lafferty, J., McCallum, A., Pereira, F., 1999, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", Computer, pp. 282-289