Ekstraksi Tabel di Internet : dalam Format HTML dan PDF

[1]Detty Purnamasari, [2]I Wayan Simri Wicaksana, [3]Lintang Yuniar Banowosari
[1, 2] Program Doktor Teknologi Informasi, [3] Program Diploma Manajemen Informatika
Universitas Gunadarma
Jl. Margonda Raya No. 100 Depok Indonesia
[1,2,3]{detty, iwayan, lintang}@staff.gunadarma.ac.id

ABSTRAK

Internet menyediakan data dalam berbagai format, salah satunya adalah tabel yang dapat dalam format HTML dan PDF. Suatu ekstraksi semi otomatis pada tabel dibutuhkan untuk mengambil data, sehingga dapat digunakan untuk proses lebih lanjut bersama dengan data lain. Hal ini dapat dilakukan dengan cara *copy-paste*, tetapi tidak efektif karena membutuhkan lebih banyak waktu dan pekerjaan berulang untuk melakukannya. Tabel terdiri dari struktur fisik dan struktur logik. Artikel ini menyajikan ekstraksi tabel dilihat dari struktur logik-nya yaitu dengan algoritma yang sudah dikembangkan pada tabel HTML dan suatu tinjauan pustaka yang akan digunakan untuk penelitian selanjutnya dalam mengembangkan algoritma ekstraksi dari struktur logik tabel PDF. Melalui artikel ini, menelaah algoritma yang sudah dikembangkan untuk ekstraksi struktur logik dari tabel HTML dan akan menjadi acuan dalam mengembangkan metode/pendekatan untuk ekstraksi tabel dalam format PDF pada penelitian selanjutnya.

Kata Kunci : *Ekstraksi tabel,struktur logik tabel, tabel HTML, tabel PDF*

PENDAHULUAN

Tabel merupakan salah satu cara yang digunakan untuk menampilkan data dalam bentuk baris dan kolom yang saling berhubungan. Menurut Liu et.al. (2008) tabel menampilkan data struktur dan informasi yang berhubungan dalam bentuk dua dimensi dan meringkaskan isinya.

Data yang tersaji di Internet, dalam hal ini adalah tabel dapat dalam format HTML (H*ypertext Markup Language)* dan PDF (*Portable Document Format*). HTML adalah bahasa yang digunakan untuk membuat website yang terdiri dari kumpulan tag. (Ronggobramantyo, 2007), sedangkan PDF merupakan format dokumen dimana merupakan alat independen, seperti file dapat dibuka di komputer dengan *platform* apa saja. (Cohene, 2002).

Pengambilan data di internet untuk digunakan pada proses lebih lanjut yang tersaji pada tabel dalam bentuk HTML dan PDF dapat dilakukan dengan cara *copy-paste*, tetapi hal ini membutuhkan banyak waktu dalam mengerjakannya. Sehingga dibutuhkan suatu teknik ekstraksi tabel secara semi otomatis.

Artikel ini menyajikan algoritma yang dikembangkan untuk ekstraksi tabel dari struktur logik pada tabel HTML, dan tinjauan pustaka mengenai ekstraksi tabel PDF yang akan digunakan sebagai pustaka dalam

Tabel 1. Transformasi pada Struktur Logik Tabel

STRUKTUR TABEL	TRANSFORMASI	KOMPONEN
Struktur logik	1. merging/splitting of region	1. *cell*s
		2. tables
		3. splitting region at detected separators
	2. graph/tree transformation	1. to correct structural errors
		2. join regions into a table region
	3. filtering	1. small region for noise reduction
		2. texture, images and half-tones
	4. sorting and indexing	1. sorting (ex : boxes by geometric attributes)
		2. indexing (ex : of *cell*)
	5. translation	1. HTML to character matrix
		2. map strings to regular expression
		3. transform token of a single class to a uniform representation
		4. encoding recognized form data
		5. indexing relation of a table

Sumber : Zanibbi,et.al. (2003)

melakukan penelitian selanjutnya untuk mengembangkan suatu metode/pendekatan ekstraksi tabel bentuk PDF.

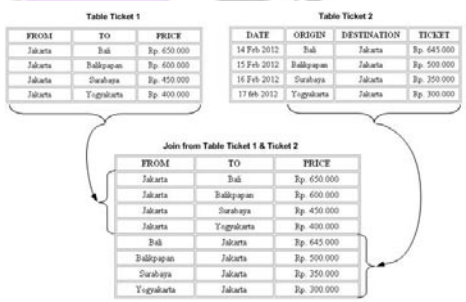
TINJAUAN PUSTAKA

Tabel terdiri dari dua struktur, yaitu : i). Struktur fisik, menjelaskan tempat tabel berada pada gambar atau file teks, dan ii). struktur logik, menjelaskan tipe dari lokasi tabel dan bagaimana pembentukan tabel, serta dapat dikodekan menggunakan *Markup language* seperti HTML. (Zanibbi, 2003).

Pada tabel 1. disajikan transformasi yang dapat dilakukan pada struktur logik tabel., beberapa diantaranya adalah *merging/splitting* untuk *cell/table*, transformasi graf/*tree* untuk memperbaiki struktur yang salah, filtering untuk mengurangi gangguan/*noise, sorting* dan *indexing*, serta *translation* untuk mengubah HTML ke bentuk matriks. Transformasi pada struktur logik merupakan salah satu proses yang dilakukan untuk menyusun ulang tabel.

Pada salah satu algoritma yang sudah dikembangkan mengenai ekstraksi tabel HTML adalah dengan memperhatikan adanya *merging* pada sel.

Pengambilan data jika dari satu tabel yang berasal dari satu sumber maka proses dengan *copy-paste* sudah memadai, maka ekstraksi tabel pada HTML akan bermanfaat jika mengambil beberapa tabel dari berbagai sumber di Internet, ilustrasi dapat di lihat pada Gambar 1. (Purnamasari, et.al., 2012)



Gambar 1. Ilustrasi Penggabungan Tabel Tiket
Sumber : Purnamasari,et.al, (2012)

Pada Gambar 1. terdapat dua bentuk tabel yang memberikan informasi harga tiket dengan nama *property* yang berbeda tetapi mempunyai arti yang sama, yang kemudian isi kedua tabel tersebut digabungkan menjadi satu tabel saja.

Dikembangkan suatu algoritma untuk melakukan ekstraksi tabel dalam format HTML sederhana menjadi bentuk *database* dengan

mempertimbangkan factor *property* dan *record*. (Purnamasari, et.al., 2012)

Craven (2003) dan Gatterbauer et.al. (2007) melakukan ekstraksi tabel di web. Pohon *Document Object Model* (DOM) merupakan penyusun suatu halaman web yang digunakan dalam pengembangan metode ekstraksi tabel yang ada di web, salah satunya digunakan oleh Lin et.al.(2009), serta Gultom et.al. (2011) dengan aplikasi Xtractors-nya, dimana selain untuk mengekstrak tabel juga untuk mashup. Algoritma dibuat menggunakan teknik rekursif dengan GUI yang *user-friendly*.

Ekstraksi dokumen PDF dilakukan oleh beberapa peneliti diantaranya: Chao (2003), Dejean, et.al. (2006), dan Liu, et. al. (2006).

Penelitian yang dilakukan oleh Ramel, et.al (2003) mengembangkan metode untuk deteksi dan ekstrak tabel dengan melakukan analisa *graphic lines*, dimana penelitian Ramel et.al ini juga dapat menjadi salah satu acuan dalam mengembangkan metode ekstraksi tabel dalam bentuk PDF.

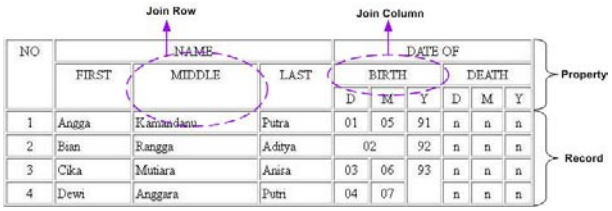
PEMBAHASAN

Ekstraksi Tabel HTML

Penelitian yang pernah dikembangkan sebelumnya pada ekstraksi tabel HTML adalah algoritma untuk melakukan ekstraksi untuk tiga bentuk tabel, yaitu : tabel bentuk standar, tabel bentuk penggabungan baris, dan tabel bentuk penggabungan *cell*/kolom. (Purnamasari, et.al., 2012).

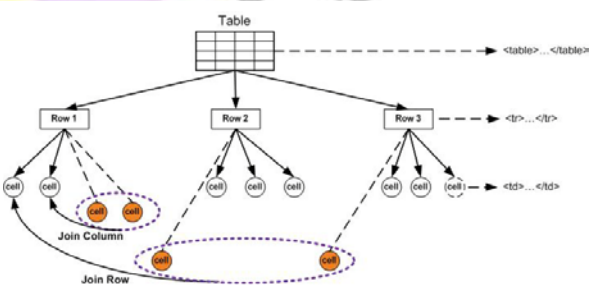
Kemudian, penelitian dilanjutkan dengan menggunakan bentuk tabel yang lebih kompleks, memperhitungkan sampai baris ke berapa disebut *property*, dan mana yang disebut sebagai isi tabel atau *record*, selain itu isi tabel tersebut juga ada yang mengalami penggabungan baris dan penggabungan kolom.

Tabel 2. Bentuk Tabel dengan Penggabungan Baris dan Kolom pada *Property* dan *Record*



Proses penggabungan baris dan kolom diilustrasikan dengan menggunakan pohon *class* logika, seperti terlihat pada gambar 2.

Ada empat algoritma yang dikembangkan, yaitu : i).menghitung jumlah total kolom dan baris sebenarnya, ii).mencari nilai *rowspan* terbesar, dan jumlah baris sebagai batas *property*, iii).mencari isi *property*, dan iv). mendapatkan isi *record*.



Gambar 2. Pohon *Class* Logika Penggabungan Baris dan Kolom
Sumber : Purnamasari,et.al, (2012)

Berikut ini adalah algoritma yang sudah dikembangkan, detail dapat dilihat pada (Purnamasari, 2012).

1. Tag dan string yang ditemukan setelah tag <table> dan sebelum tag </table> merupakan penyusun tabel.
2. Cari nilai rowspan terbesar dari tiap tag <td>...</td> pada tag <tr>...</tr> ke-s sampai tidak ditemukan nilai rowspan > 1 untuk mendapatkan jumlah baris sebagai property. (rowmax_pro = batas baris property)
3. Mengambil isi property, dilakukan mulai dari batas akhir tag <tr>...</tr> yang ke- rowmax_pro

down to 1 untuk mendapatkan posisi cell/kolom jika terjadi penggabungan kolom.

4. Isi record diambil mulai dari baris terakhir pada tabel / tag <tr>...</tr> ke-RsTotal sampai dengan baris ke rowmaxpro + 1.
5. Ada 3 kondisi yaitu : i). Jika colspan = 1 dan rowspan = 1, ii). Jika colspan = 1 dan rowspan >1, iii). Jika colspan >1 dan rowspan = 1.

Algoritma 2. Mencari Nilai Rowspan Terbesar, dan Jumlah Baris sebagai Batas Property

```
mBatas (0) = 1
while s = 1 do
    rsMax (0) = 1
    Hitung jum<td>
    For i = 1 to jum<td>
        If rs (i) > 1 then
          If rs (i) > = rsMax (i-1) then rsMax (i) = rs (i)
          else
          If rs (i) < rsMax (i-1) then rsMax (i) = rsMax (i-1)
     Next i ;
mBatas (s) = rsMax (i) + s – 1
if mBatas (s) < mBatas (s-1) then mBatas (s) = mBatas (s-1)
until mBatas (s) ;
rowmax_pro = mBatas (s) ;
```

Keterangan :

- RsMax : nilai rowspan tertinggi
- mBatas : nilai batas row sebagai property
- rowmax_pro : batas baris yang dapat disebut sebagai property
- i : tag <td>
- s : tag <tr>

Dengan menggunakan empat algoritma diatas, maka data hasil ekstraksi dapat di simpan ke *database* dan data yang diekstraksi sudah berdasarkan pada *property*-nya.

Ekstraksi Tabel PDF

Pendekatanyang dikembangkan untuk ekstraksi pada dokumen bentuk PDF salah satunya dilakukan oleh Dejean, et.al. (2006). Pendekatan yang dikembangkan dalam penelitiannya adalah membuat sistem untuk konversi dokumen PDF ke format XML. Langkah yang dilakukan adalah dengan melakukan ekstrak pada format PDF untuk teks, path object, dan external object. Algoritma XY-cut digunakan untuk ekstraksi gambar, dan ekstraksi struktur logik dari dokumen dilakukan dengan deteksi ToC (*Table of Content*).Dibuat *Graphic User Interface* (GUI) untuk mempermudah pengguna melakukan perbaikan dari hasil konversi.

Yildiz, et.al. (2005) Mengembangkan metode dengan menggunakan dua heuristik untuk melakukan ekstraksi tabel bentuk PDF dan menyimpan dalam bentuk XML. Pada tahapan awal dari pendekatannya, digunakan aplikasi yang sudah ada yaitu pdftohtml (dikembangkan oleh Geuorgui Ovtcharov dan Rainer Dorsch) untuk merubah PDF ke XML. Masing-masing potongan teks dari PDF yang diubah ke bentuk XML menggunakan 5 atribut.

Ekstraksi tabel dapat dilakukan dengan metode berikut ini, yaitu :1). *pre-defined layout based* : pendekatan yang berisi beberapa *template* yang mungkin sebagai struktur tabel, 2). *heuristics based*: pada pendekatan ini dibuat suatu kumpulan aturan yang digunakan untuk mengambil keputusan, dan 3). *statistical based* : pendekatan ini menggunakan perhitungan statistika, dimana parameter yang didapat akan digunakan untuk mengambil keputusan.

Ekstraksi tabel dilakukan dari dokumen XML hasil dari aplikasi pdftohtml dengan membuat elemen teks pada posisi yang tepat dari potongan teks di PDF. Langkah dalam pendekatan dalam penelitian Yildiz et.al. adalah sebagai berikut :

1. ubah dokumen PDF dengan aplikasi pdftohtml ke bentuk XML.
2. heruristik 1 : *table recognition* (mengidentifikasikan bentuk tabel)
3. heuristik 2 : *table decomposition* (menguraikan tabel mendekati aslinya, dan juga melakukan identifikasi element *header*, menghitung banyak kolom dan baris, dll)

Ekstraksi pada dokumen PDF dengan melihat struktur logika dilakukan karena : (Liu, et.al., 2008)

1. informasi struktur logika pada dokumen PDF tidak terlihat dengan jelas (eksplisit).
2. ekstraksi teks pada dokumen PDF masih ada kekurangan /ketidaktepatan pada hasil ekstraksi dengan *tool* yang ada saat ini.
3. Beberapa *noise* baru dapat muncul pada beberapa *converter tools* jika dilakukan konversi PDF ke bentuk lain.

Metode yang digunakan pada penelitian Liu. et.al. (2008) adalah sebagai berikut :

1. membuat algoritma untuk mendeteksi *sparse line* berdasarkan pada informasi koordinat.
2. algoritma tersebut dikombinasikan dgn *keyword* untuk mengidentifikasikan tabel pertama kali karena biasanya tabel dalam dokumen diberikan nama seperti : tabel 2.1 xxx ; *Form* 2.1. xx
3. area *sparse line* akan mendeteksi batas dari tabel.
4. untuk mendeteksi *sparse line* digunakan pendekatan *buttom-up* untuk memulai proses ekstraksi dari level karakter pada dokumen PDF
5. untuk konversi karakter/kata ke dalam kata/garis, diadopsi beberapa heuristic berdasarkan jarak antara karakter/*word*.

Penelitian Oro, et.al. (2009) tentang PDF-TREX dengan pendekatan *heuristic-based* untuk pengenalan dan ekstraksi tabel dari dokumen PDF. Pendekatan ini dilakukan dengan dibuatnya dokumen menjadi kisi-kisi 2 dimensional dengan garis cartesian dan mengekstraknya menjadi kumpulan *cell* yang lengkap dari kordinat 2 dimensi tersebut. Algoritma heuristic memiliki 8 tahapan, yaitu :

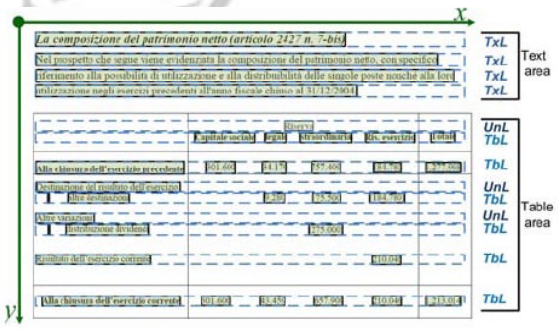
1. *element harvesting*
2. *lines buliding*
3. *segments building & lines tagging*
4. *table area building*
5. *block and row building*
6. *column building*
7. *table building*
8. *extraction*

Algoritma heuristic melakukan pengenalan tabel yang ada didokumen dengan spatial relationship, maka algoritma ini tidak membutuhkan :

1. *linguistic* atau *domain knowledge*
2. *graphical* metadata dan *ruling line*
3. *predefined table layout*

Pendekatan pada penelitian Ramel et.al. (2003) dilakukan untuk mendeteksi dan ekstraksi tabel dengan melakukan analisa *graphic lines*. Tabel sederhana terdiri dari sel-sel matrik, dimana semua sel dalam 1 garis memiliki tinggi yang sama dan semua sel dalam 1 kolom memiliki lebar yang sama. Semua sel tersebut dibatasi oleh *graphic lines*.

Physical layout dan *logical structure* digunakan pada penelitian ini dan dikembangkan DTD (*Document Type Definition*) untuk merepresentasikan tabel ke dalam XML.



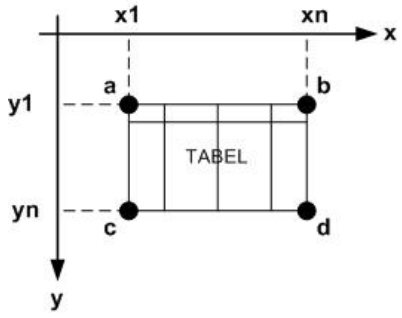
Gambar 3. Elemen, Segmen, Lines, Text dan Area Tabel
Sumber : Oro, et.al. [2009]

Pada gambar 3 adalah proses dalam melakukan ekstraksi mulai dari tahap

element harvesting, lines building, segments building & lines tagging, table area building yang ada pada algoritma heuristic yang dikembangkan oleh Oro, et.al (2009).

Ide Dasar Ekstraksi Tabel PDF pada Penelitian Selanjutnya

Penelitian selanjutnya yang akan dilakukan adalah mengembangkan algoritma untuk melakukan ekstraksi tabel dalam bentuk PDF. Ide dasar pembentukan algoritma awalnya adalah dengan melihat pada penelitian Oro,et.al. (2009) yang membuat dokumen menjadi suatu bidang kartesian, dan dengan tabel dalam bentuk sederhana seperti terlihat pada ilustrasi gambar 4.



Gambar 4. Ilustrasi Dokumen pada Bidang Kartesian

Pada gambar 4. tabel di bentuk oleh 2 garis horizontal sejajar sama panjang (garis ab; garis cd), serta 2 garis vertikal sejajar sama panjang (garis ac; garis bd) dimana ujung garis tersebut bertemu.

Berikut ini adalah langkah awal yang menjadi ide dasar untuk mengenali tabel yang ada dalam dokumen PDF, yaitu dengan mengenali bentuk persegi (a-b-d-c) yang ada pada dokumen yang dapat diasumsikan suatu tabel :

1. dilakukan *scan* pada dokumen PDF.
2. Jika ditemukan teks, maka diabaikan.
3. jika di temukan titik mulai dari titik a(x_1,y_1) sampai dengan b(x_n ,y_1), maka itu adalah garis.
4. selama dilakukan *scan*, jika terdapat kondisi seperti langkah no. 5 dan no. 6, maka dari hasil scan tersebut adalah suatu kotak/persegi yang bisa menjadi suatu tabel.
5. garis horizontal mulai titik a(x_1, y_1) s/d titik b(x_n ,y_1) yang sejajar dengan garis yang dibentuk dari titik c(x_1,y_n) s/d titik d (x_n , y_n)
6. garis vertikal mulai titik a(x_1, y_1) s/d titik c(x_1 , y_n) yang sejajar dengan garis yang dibentuk dari titik b(x_n , y_1) s/d titik d (x_n , y_n)
7. setelah ditemukan adanya kotak/persegi pada dokumen PDF, maka harus dilakukan suatu langkah untuk mengetahui apakah kotak/persegi tersebut adalah tabel, atau hanya suatu batas/kotak dari tampilan gambar/teks pada dokumen.

Menurut Ramel, et.al. (2003), dikatakan bahwa tabel tersusun dari sedikitnya dua kotak/persegi, dan Yildiz, et.al. (2005) juga mengatakan bahwa tabel harus memiliki lebih dari satu kolom. Berdasarkan hal tersebut, didapatkan suatu ide awal untuk menyusun langkah dalam menentukan suatu kotak/persegi adalah tabel atau bukan.

Dalam pendekatan yang akan dikembangkan, diasumsikan bahwa suatu tabel paling sedikit terdiri dari empat kotak/persegi, seperti terlihat pada gambar 5.

p	q
r	s

Gambar 5. Asumsi Bentuk Tabel Minimal dengan 4 Kotak/Persegi

Mengapa tabel dikondisikan mempunyai sedikitnya empat kotak/persegi? Pada penelitian yang akan dilakukan selanjutnya, tabel dari dokumen PDF yang akan diekstrak adalah tabel yang memiliki property, sehingga dari gambar 5 dapat dikatakan kotak p dan kotak q adalah suatu property, sedangkan kotak r dan kotak s adalah isi/data yang merupakan record tabel.

Melihat pada asumsi susun minimal suatu tabel, maka tabel dengan empat kotak/persegi mempunyai tiga garis horizontal yang sama panjang dan sejajar, serta tiga garis vertikal yang sama panjang dan sejajar.

Melalui ide dasar yang dipaparkan pada artikel ini, maka penelitian selanjutnya akan dikembangkan metode ekstraksi tabel bentuk PDF dengan memperhatikan faktor *property*, isi/data *record*, dan adanya *merging*/join baris dan kolom.

KESIMPULAN

Ekstraksi tabel dengan melihat struktur logik, pada penelitian terdahulu telah dikembangkan algoritma untuk melakukan ekstraksi tabel HTML dengan memperhatikan faktor *property* dan *record*, selain itu juga melihat adanya *merge*/join pada baris dan kolom. Data hasil ekstraksi dapat tersimpan berdasarkan pada *property*-nya, sehingga mudah jika akan digunakan pada proses selanjutnya.

Pada ekstraksi tabel bentuk PDF, penelitian yang telah dilakukan oleh para peneliti adalah dengan menggunakan suatu aplikasi tertentu untuk merubah format PDF ke format yang mempermudah untuk melakukan ekstraksi. Berdasarkan pada tinjauan pustaka mengenai ekstraksi pada PDF, maka penelitian selanjutnya yang akan dilakukan adalah mengembangkan suatu algoritma untuk melakukan ekstraksi pada tabel PDF. Selain itu, dengan adanya ide dasar ekstraksi tabel PDF, akan membantu dalam penelitian selanjutnya.

Penelitian lanjutan yang perlu dilakukan adalah dengan mengembangkan aplikasi real sampai dengan tahapan penggabugan tabel dari berbagai sumber dan direpresentasikan dalam model data terstruktur. Data terstruktur untuk teknologi saat ini akan mengadapsi dari model XML dan RDF.

DAFTAR PUSTAKA

Craven, Timothy, C., 2003, *HTML Tags as Extraction Cues for Web Page Description Construction*, Informing Science Journal, Vol 6

Cohene, T, Khouri, A. 2002. *How to Export a Table from a PDF File into an Excel Spreadsheet.* http://www.library.mcgill.ca/edrs/services/publications/howto/PDFtoxls/PDFtoexcel.html. diunduh pada 21 September 2011

Dejean, Herve. Meunier, Jean-Luc. 2006. *A System for Converting PDF Documents into Structured XML Format*. Document Analysis Systems'06

Gatterbauer. Wolfgang, Bohunsky. Paul, Herzog. Marcus, Krupl. Bernhard, Pollak. Bernhard, 2007, *Towards Domain-Independent Information Extraction from Web Tables*, Proceedings of the 16th International Conference on World Wide Web, Canada, pp.71-80

Gultom, R.A.G, Fitri Sari, R, Budiarjo, B. 2011. *Proposing the new Algorithm and Technique Development for Integrating Web Table Extraction and Building a Mashup*. Journal of Computer Science 7 (2) : 129-142, ISSN 1549-3636

Lin, J, Wong, J, Nichols, J, Cyper, A. 2009. *End-User Programming of Mashups with Vegemite*. Proceedings of the 13th international conference on Intelligent user interfaces pp. 97-106

Liu, Ying.Mitra, Prasenjit.Giles C. Lee. Bai, Kun. 2006. *Automatic Extraction of Table Metadata from Digital Documents*. Proceedings of the 6th ACM/IEEE-CS Joint Conference

Liu, Ying. Mitra, Prasenjit. Giles, C. Lee. 2008. *A Fast Perprocessing Method for Table Boundary Detection : Narrowing Down the Sparse Lines using Solely Coordinate Information*. The Eight IAPR International Workshop on Document Analysis Systems, DAS '08.

Oro, Ermelinda. Ruffolo, Massimo.2009. *PDF-TREX : An Approach for Recognizing and Extracting Tables from PDF Documents*. Proceeding ICDAR '09. 10th International Conference on Document Analysis and Recognition

Purnamasari, Detty. Wicaksana, I Wayan Simri. Ruhama, Syamsi. 2012. *Algoritma untuk Ekstraksi Tabel HTML di Web*. Konferensi Nasional Sistem Informasi (KNSI)

Purnamasari, Detty. Wicaksana, I Wayan Simri. Banowosari, Lintang Yuniar. 2012. *The Approach for Table Extraction in Internet Based on Property and Instance*. International Conference on Soft Computing, Intelligent System, and Information Technology (3rd ICSIIT)

Ramel, J.Y. Crucianu, M. Vincent, N. Faure,C.2003. *Detection, Extraction and Representation of Tables*. Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR)

Ronggobramantyo, D. 2007. *Belajar HTML yang merupakan Dasar dari Pembuatan Website.* http://www.dhimasronggobramantyo.com/artikel/Belajar_HTML_yang_merupakan_dasar_dari_pembuatan_website. diunduh pada 20 September 2011.

Yildiz, Burcu. Kaiser, Katharina. Miksch, Silvia.2005. *pdf2table : A Method to Extract Table Information from PDF Files*. Proceedings of the 2nd Indian International Conference on Artificial Intelligence IICAI05 Pune India

Zanibbi, Richard. Blostein, Dorothea. Cordy, James R. 2003. *A Survey of Table Recognition: Models, Observations, Transformations, and Inferences*. International Journal on Document Analysis and Recognition.