Proceeding, Seminar Ilmiah Nasional Komputer dan Sistem Intelijen (KOMMIT 2008)
Auditorium Universitas Gunadarma, Depok, 20-21 Agustus 2008 ISSN : 1411-6286

Keterkinian Solusi Wrapper

[1]Lily Wulandari.E; [2]I Wayan Simri Wicaksana

[1]Program Doktoral TI, Universitas Gunadarma (lily@staff.gunadarma.ac.id)
[2]Pusat Studi Teknologi Sistem Informasi, Universitas Gunadarma
(iwayan@staff.gunadarma.ac.id)

ABSTRAK

Kata 'wrapper' berasal dari komunitas database. Suatu wrapper di dalam konteks ini digunakan sebagai suatu penengah/mediator antara beberapa database dan satu aplikasi [5]. Dengan cara yang serupa, di dalam lingkungan web suatu wrapper mengkonversi informasi dari dokumen HTML ke dalam informasi yang tersusun (seperti XML). Informasi yang tersusun dapat disimpan untuk penggunaan berikutnya, seperti menjawab query-query, atau menghasilkan secara dinamis atas permintaan melalui suatu antar muka Web atau dari satu aplikasi.
Saat ini beberapa perusahaan menggunakan informasi yang tersedia di Web untuk sejumlah tujuan. Namun demikian kebanyakan dari informasi ini hanyalah tersedia dalam bentuk dokumen HTML. Untuk mengatur data Web secara efektif, seorang pengguna perlu untuk mengekstrak informasi terkait, memahami struktur semantiknya, dan mengkonversinya ke format yang diinginkan. Baru-baru ini, beberapa teknik-teknik yang mengizinkan informasi dari Web untuk secara otomatis di-ekstrak telah digambarkan. Kontribusi dari paper ini adalah meninjau ulang teknik-teknik dan tool utama untuk melakukan ekstrak informasi yang tersedia di Web. Secara khusus kami menekankan keuntungan-keuntungan dan kelemahan-kelemahan dari teknik-teknik serta menganalisa dari sudut pandang pengguna.

Kata Kunci : *Ekstraksi Informasi, HTML, Wrapper*

1. PENGANTAR

Wrapper adalah suatu jenis perangkat lunak yang digunakan untuk melakukan attach data bersama-sama dengan komponen-komponen perangkat lunak lain. Seringkali data web yang menarik bukan dalam bentuk sistem database tetapi berupa halaman-halaman HTML, halaman-halaman XML, atau file teks. Sementara penyajian informasi dalam bentuk HTML tidak masalah bagi para pemakai manusia, informasi di dalam bentuk ini tidak dapat dilakukan untuk pengolahan yang otomatis karena akan menghasilkan sejumlah besar informasi yang tidak relevan.

Lebih lanjut, arti semantik dari bidang keilmuan yang berbeda dari suatu dokumen HTML bisa direpresentasikan dalam cara-cara yang berbeda dengan penyajian data yang terstruktur [6]. Data di dalam bentuk-bentuk semantik tidak secara langsung dapat dipakai oleh SQL standard. Karenanya, para pemakai web atau aplikasi-aplikasi memerlukan suatu cara yang cerdas untuk melakukan ekstrak data dari sumber web ini. Salah satu pendekatan yang populer adalah menulis wrapper di sekitar sumber, baik secara manual atau dengan bantuan perangkat lunak, untuk membawa data web ke dalam jangkauan dari tool-tool query yang lebih canggih dan sistem pengintegrasian informasi berbasis mediator.

Teknik-teknik Information Extraction (IE) mengarahkan pada

transformasi informasi dalam bentuk teks yang tidak terstruktur ke dalam informasi yang terstruktur sesuai dengan aturan-aturan ekstraksi yang mengidentifikasi informasi terkait. Aturan-aturan ekstraksi digunakan baik untuk mengenali bagian-bagian suatu dokumen yang berisi data yang relevan, maupun untuk memberi suatu makna semantik kepada informasi yang dikenali, untuk menerjemahkannya ke dalam suatu bentuk terstruktur. Satu kumpulan aturan ekstraksi yang cocok untuk mengekstrak informasi dari suatu Situs web disebut wrapper.

Suatu wrapper mempunyai tiga tugas utama: pertama, untuk mendapat kembali suatu dokumen web, kedua, untuk menyaring informasi terkait dari halaman web; dan ketiga, untuk memetakan informasi ke dalam bentuk yang diperlukan oleh setiap aplikasi tertentu [5].

Wrapper dapat digunakan untuk menyajikan suatu antar muka atau interface sederhana ke sumber yang berbeda, sehingga mereka semua menyajikan suatu antar muka yang umum, untuk menambahkan kemampuan akses bagi beragam sumber data, atau untuk pengungkapan antar muka antara berbagai sumber data tersebut.

Dua pendekatan utama untuk perancangan dari tool-tool wrapper telah diusulkan selama tahun terakhir : rancang-bangun pengetahuan dan pendekatan pelatihan otomatis [3]. Di dalam pendekatan rancang-bangun pengetahuan, pembentuk ekstraksi aturan-aturan dirancang oleh seorang ahli domain, menurut latar belakang pengetahuannya tentang karakteristik-karakteristik dokumen. Pada pendekatan yang demikian ketrampilan/skill pengguna memainkan suatu peran yang penting di dalam identifikasi dan ekstraksi informasi terkait. Pendekatan pelatihan yang otomatis sebagai gantinya memanfaatkan teknik-teknik AI untuk mempengaruhi aturan-aturan ekstraksi berawal dari satu kumpulan pola informasi yang ditandai untuk ekstraksi oleh seorang pengguna.

Wrapper Generation

Secara garis besar wrapper dapat digolongkan dengan mempertimbangkan jenis halaman-halaman dimana masing-masing wrapper mampu berhubungan. Tiga jenis halaman web yang berbeda yang terdapat pada [6] yaitu:

- *unstructure pages*: beberapa ketrampilan ilmu bahasa (linguistic) dibutuhkan untuk mempelajari fitur-fitur yang relevan dari halaman-halaman;
- *semi-structure pages* : halaman-halaman yang tidak sesuai dengan suatu bagan yang ditetapkan, dalam pengertian bahwa tidak ada uraian yang terpisah untuk jenis/tipe dari data;
- *structured pages :* informasi tentang struktur tersedia, sedemikian sehingga informasi hanya dapat diekstrak menurut batasan-batasan *syntactic*.

Wrapper generation dapat tercapai dalam tiga cara yang berbeda: *manually*, *semi-automatically* (setengah otomatis) dan *automatically* (secara otomatis). Untuk wrapper yang dihasilkan secara manual suatu pengetahuan yang mendalam tentang struktur dari dokumen yang sedang di-wrap dan kode dari parsers yang cocok untuk suatu struktur dari dokumen tersebut diperlukan. Untuk mempercepat proses ini, di dalam lima tahun terakhir usaha yang besar sudah didedikasikan untuk membangitkan modul-modul template yang dapat berupa semi-otomatis menyesuaikan diri dengan sumber-sumber yang berbeda. Teknik-teknik otomatis untuk wrapper generation biasanya menggunakan teknik-teknik *supervised machine learning* (pembelajaran mesin yang terawasi). Di dalam sistem ini para pemakai diwajibkan untuk memberi nama (label) data yang relevan dari satu kumpulan halaman web yang akan digunakan

sebagai suatu kumpulan pelatihan untuk proses pembelajaran. Fitur dari suatu kumpulan pelatihan dapat dimanfaatkan oleh suatu sistim untuk mengekstrak aturan-aturan (perintah-perintah).

Sesungguhnya, *hand-written wrapper* melakukan lebih baik dibanding hasil yang diperoleh secara otomatis. Kelemahan utama dari *hand-written wrapers* adalah pengembangan yang manual dapat sangat membosankan dan kadang-kadang terlalu sulit untuk dipenuhi.

Satu taksonomi yang menarik dari tool-tool wrapper generation dapat disediakan berdasarkan pada teknik-teknik AI yang dilibatkan [6]:

- Tool-tool Wrapper generation berdasarkan pada bahasa-bahasa untuk wrapper generation (seperti, TSIMMIS, Minerva dan Web-OQL) yang tujuan utamanya mendukung pemakai di dalam membangun wrappers.
- HTML aware tool (seperti, Lixto [4] dan Road-Runner [9]) adalah tool-tool wrapper generation yang memanfaatkan penyajian-penyajian yang sesuai tentang dokumen HTML yang asli.
- Tool-tool wrapper generation (seperti, RAPIER, SRV, WHISK[7]) yang berhubungan dengan halaman-halaman web yang kebanyakan berisi teks bebas (misal, persewaan dan iklan-iklan pekerjaan, pengumuman-pengumuman). Tool-tool seperti itu didasarkan pada teknik-teknik pemrosesan bahasa alami (NLP) seperti *filtering*, *semantic* dan *syntactic tagging*.
- Tool induksi wrapper yang bekerja dengan batasan-batasan linguistik yang relaxing dan memberikan perhatian yang lebih kepada pengaturan fitur. Contoh-contoh dari tool-tool wrapper generation berikut ini yang mengikuti strategi seperti itu adalah Wien [3], SoftMealy [6] dan STALKER[6].
- Tool-tool berbasis modeling yang mencoba untuk mengidentifikasi pencocokan objek dokumen suatu struktur yang sudah dikenal[6].
- Akhirnya, tool-tool berbasis ontologi yang memanfaatkan ontologi-ontologi yang sudah ada sebelumnya dan mengekstrak informasi yang dipercaya secara langsung pada data [1].

Wrapper generation dari Sudut Pandang Pengguna

Pengelompokan/Penggolongan yang disediakan pada bagian sebelumnya tidak sepenuhnya memuaskan karena ia tidak mempertimbangkan usaha pengguna yang diperlukan untuk mendisain suatu wrapper, dan lebih dari itu tidak mempertimbangkan kategori dari Halaman web yang akan di-wrap. Pada bagian ini, kami memberi satu alternatif taksonomi untuk sistem induksi wrapper, yang berusaha untuk menggolongkan sistem menurut fitur kunci, yang mengizinkan para pemakai untuk membangun wrapper yang baik.

Secara khusus, suatu karakteristik yang perlu dipertimbangkan adalah struktur dari halaman-halaman yang sedang di-wrap. Kami dapat mencirikan dua kategori utama dari sistem:
(1) sistem yang dikhususkan untuk halaman-halaman yang tersusun dan
(2) sistem yang dirancang untuk halaman-halaman dengan suatu struktur yang sedikit lebih sulit.

Sistem seperti ShopBot, Wien, SoftMealy dan STALKER digolongkan ke dalam kelompok yang pertama sementara kelompok kedua memasukkan di dalamnya sistem seperti RAPIER, SRV dan WHISK.

Fitur lain yang relevan untuk para pemakai adalah tingkat dari otomatisasi sistim dari wrapper generation, yaitu usaha yang diperlukan untuk mendisain suatu wrapper. Sesungguhnya, jika suatu bahasa untuk wrapper generation digunakan, interaksi pengguna sangat

tinggi: pengguna tersebut diwajibkan untuk menganalisa dokumen source program, untuk mencari fitur yang menarik, dan untuk menulis beberapa kode yang cocok/sesuai. Suatu sistim induksi wrapper digunakan di sini untuk menghindari pembuatan kode/sintak program karena proses tersebut adalah otomatis atau hampir semiotomatis. Contoh-contoh sistem ini adalah RoadRunner (suatu tool yang secara penuh otomatis) dan BYU [1] (yang memerlukan suatu konstruksi pendahuluan dari suatu ontologi oleh seorang ahli domain).

Karakteristik lain yang relevan adalah pertimbangan kegunaan dari tool-tool. Banyak tool menyediakan antar muka dalam bentuk grafis (GUI): tool-tool HTML-aware, NLP based, dan induksi wrapper sesuai untuk tujuan ini. Beberapa sistem yang menawarkan penuntun-penuntun visual untuk merancang wrapper adalah W4F [5] dan Lixto.

2. SISTEM INDUKSI WRAPPER

Penyaringan informasi dikenali sebagai satu aplikasi teknik-teknik pembelajaran mesin standar untuk permasalahan pembagian penggolongan dokumen yang didasarkan pada fitur yang diperoleh dari konteks mereka. Dengan demikian, Induksi wrapper ada yang diawasi dan tidak diawasi. Sistem induksi wrapper dapat juga terdiri atas : *string-based, token-based* dan *HTML-aware.* Sistem induksi wrapper berbasis token dan berbasis string memperlakukan dokumen sebagai suatu urutan dari karakter-karakter atau tanda-tanda. Algoritma-algoritma di balik sistem ini biasanya mencari-cari delimiters, pola-pola delimiter atau bahasa-bahasa delimiter reguler yang mencocokkan data pelatihan dengan baik.

Di dalam bagian ini, kami menguraikan beberapa sistem induksi wrapper. Sebagaimana telah disebutkan di atas, ketika mempertimbangkan wrapper bagan-bagan contoh diperlukan untuk memperoleh definisi wrapper yang paling baik. Dengan demikian kita dapat membedakan antara sistem yang perlu dilabelkan sebagai contoh-contoh (seperti SRV, STALKER,WHISK dan WIEN) dan sistem yang dapat mencari informasi penting secara otonomi (seperti, Shop-Bot dan Wien). Secara khusus, uraian berikut mempertimbangkan keseluruhan struktur dari suatu sistim, jenis dokumen suatu sistim yang dapat berhubungan, dan ketahanan dari wrapper yang dihasilkan. Dalam konteks yang demikian, ketahanan berarti kemampuan suatu wrapper untuk berhubungan dengan variasi-variasi kecil di dalam dokumen, seperti materi atau permutasi-permutasi yang hilang dari materi yang diekstrak. Suatu aspek lebih lanjut yang perlu dipertimbangkan adalah tingkatan struktur dari informasi yang diekstrak. Sesungguhnya, suatu sistim wrapper hanya mampu memisahkan *sentence* (kalimat) di dalam suatu dokumen lebih sedikit ekspresif dibanding suatu sistim yang mampu mengekstrak struktur kompleks (sebagai contoh, daftar record-record).

ShopBot adalah satu agen yang didedikasikan untuk mengekstrak informasi dari halaman-halaman yang berhubungan dengan Web services (sebagai contoh, situs-situs e-commerce). Sistim memasukkan dua langkah utama dengan mengkombinasikan heuristik, pencocokan pola dan teknik-teknik pembelajaran induktif. Di dalam tahap yang pertama, ShopBot menganalisa Situs web target untuk mempelajari strukturnya dan untuk menemukan halaman-halaman berisi informasi terkait. Pada tahap kedua, beberapa heuristik dimanfaatkan untuk menghasilkan suatu uraian yang cocok tentang halaman-halaman yang terpilih. Sekali kumpulan uraian dihasilkan, sistim melaksanakan suatu proses peringkat/*ranking* untuk mengidentifikasi uraian terbaik yang tersedia. Suatu fungsi peringkat yang

sederhana mempertimbangkan banyaknya kemunculan-kemunculan dari suatu uraian item yang diberikan. Keterbatasan ShopBot adalah bahwa ia tidak mampu untuk mengekstrak fitur-fitur label, sehingga suatu proses pemberian label dilaksanakan secara manual.

WIEN (Wrapper Induction ENvironment) adalah tool pertama untuk *inductive wrapper generation*. WIEN beroperasi pada teks-teks yang tersusun yang berisi informasi yang diorganisasikan pada suatu tampilan yang berbentuk tabel. Untuk mempercepat proses pembelajaran, sistim secara otomatis memberi label-label halaman-halaman *training* (pelatihan) dengan memanfaatkan heuristic-heuristik berbagai domain spesifik. Proses generasi wrapper menggunakan suatu teknik pembelajaran yang induktif dari bawah ke atas. Lebih tepat, algoritma mencari spasi dari semua delimiters yang mungkin untuk suatu wrapper yang kompatibel dengan kumpulan pelatihan. Sistim hanya berhubungan dengan dokumen yang memperlihatkan suatu struktur yang tetap (misal, ia tidak berhubungan dengan nilai-nilai yang hilang atau permutasi-permutasi atribut-atribut).

SoftMealy adalah suatu sistim berdasarkan pada non-deterministic finite state automata (NDFA), dan ia sebagian besar dipahami untuk mempengaruhi wrapper untuk halaman-halaman yang semistructured. Secara kontras dibandingkan WIEN, SoftMealy mampu menangani nilai-nilai yang hilang. Sistim menghasilkan penyaringan aturan-aturan atas pertolongan suatu algoritma pembelajaran yang menggunakan halaman-halaman pelatihan yang berlabel yang diwakili sebagai satu penggambaran otomatisasi di semua permutasi data masukan. Status dan transisi-transisi status dari otomatisasi bersesuaian, berturut-turut, untuk menyaring data dan menghasilkan aturan-aturan. SoftMealy lebih ekspresif dibanding WIEN, meski ia mempunyai satu keterbatasan efisiensi karena ia mempertimbangkan masing-masing permutasi yang mungkin dari data masukan.

STALKER adalah suatu sistim untuk wrapper pembelajaran terawasi. Satuan pelatihan disediakan oleh seorang pengguna yang harus menetapkan data yang relevan untuk diekstrak: masing-masing halaman dihubungkan dengan suatu urutan dari token-token (tanda-tanda) yang ditafsirkan sebagai tanda untuk tahap penyaringan. STALKER sesuai untuk menyaring data dari sumber data yang terstruktur secara hierarki, dan ia tidak bersandar pada urutan dari materi (yaitu, ia dapat dengan mudah menangani nilai-nilai yang hilang). Ia lebih ekspresif dibanding WIEN, karena pola belajarnya dapat berisi *wildcards* dan bahkan aturan-aturan yang menyatakan perlawanan. Aturan-aturan ganda dapat bersarang untuk memungkinkan pengenalan struktur kompleks seperti *list of tuples* atau *list of list*. Hasil penyaringan dari wrapper digunakan sebagai masukan untuk wrapper di tingkatan yang lebih dalam. Model nesting ini juga mengizinkan untuk menunjukkan analogi hubungan-hubungan orang tua-anak untuk struktur HTML dari sumber, tetapi struktur nesting harus digambarkan secara manual.

RAPIER (Robust Automated Production of Information Extraction Rules) adalah suatu sistim induksi wrapper untuk halaman-halaman yang *semistructured*. Dibutuhkan sebagai masukan suatu dokumen dan suatu template yang menandakan data yang disaring dari dokumen seperti itu, dan aturan pencocokan pola keluaran-keluaran menurut suatu template yang diberikan. Untuk menghasilkan pola-pola, RAPIER menggunakan baik kemunculan fitur yang semantik maupun yang *syntactic*, berturut-turut, di suatu *tagger* (yang ditugaskan ke masing-masing kata suatu label yang tepat) dan

suatu kamus (yang dirancang untuk berisi informasi tentang class-class semantik di dalam dokumen yang dianalisa). Aturan penyaringan terdiri atas tiga bagian: suatu *pre-compiler*, suatu *compiler* dan suatu *pos-compiler*. Masing-masing bagian adalah suatu pola yang digunakan untuk mengidentifikasi data target dan delimiters untuk data seperti itu. Proses dari pembuatan pola dimulai dengan suatu pola yang spesifik untuk masing-masing contoh dan menyamaratakan pola-pola ini dengan mempertimbangkan masing-masing pasang pola-pola yang diciptakan.

SRV (Sequence Rules with Validation) adalah suatu algoritma pembelajaran hubungan top-down. Ia bekerja di suatu set dari halaman-halaman yang berlabel yang ditentukan, dan menggunakan beberapa fitur (yang dihubungkan dengan tanda-tanda yang muncul di dalam halaman-halaman) untuk menghasilkan aturan penyaringan logika firstorder. Fitur yang dipertimbangkan oleh SRV baik yang sederhana atau relational. Masing-masing fitur yang sederhana memetakan suatu tanda ke suatu nilai kategori (seperti, panjang, jenis model huruf, ortografi), sementara fitur relational menampilkan hubungan *syntactic*/*semantic* antar tanda-tanda. SRV menghasilkan aturan-aturan penyaringan menggunakan semua instance-instance yang mungkin (yaitu., kalimat-kalimat) di dalam dokumen. Masing-masing instance, yang diperkenalkan sebagai satu masukan kepada penggolong, dihubungkan dengan suatu nilai yang menandakan pantas tidaknya sebagai pengisi untuk slot target. Aturan-aturan yang disaring adalah sangat ekspresif, karena mereka dapat menyertakan jenis kata, semantik, class-class dll.

Sistim WHISK dapat berhubungan dengan bermacam-macam teks, karena ia memanfaatkan suatu penganalisis *syntactic* dan suatu penggolongan semantik. Diberikan suatu set pelatihan dari halaman-halaman, WHISK menghasilkan ekspresi-ekspresi reguler yang digunakan untuk mengenali konteks dari relevan instance-instance (yaitu., kalimat-kalimat) dan delimiters dari instance-instance seperti itu. Sistim menggunakan suatu algoritma pelindung untuk mempengaruhi ekspresi-ekspresi reguler di suatu top-down fashion: tahap pembelajaran dimulai dengan aturan paling umum dan dilanjutkan dengan semakin mengkhususkan aturan-aturan yang tersedia; proses berhenti ketika semua instance tercakup oleh set dari aturan-aturan. Suatu pembatasan yang utama dari sistim ini adalah kebutuhan akan mempertimbangkan semua permutasi yang mungkin (seperti SoftMealy).

3. KESIMPULAN

eperti yang dapat kita lihat, sistim yang lengkap dan baik adalah WHISK, yang melaksanakan baik pada teks setengah terstruktur maupun tidak terstruktur, dan dapat berhubungan dengan nilai-nilai yang hilang dan permutasi dari field-field. Bagaimanapun, ia memerlukan suatu satuan pelatihan yang “lengkap”, yaitu., satu set contoh-contoh termasuk semua kemunculan dari nilai yang mungkin. STALKER dan SRV tidak memerlukan set pelatihan yang lengkap, tetapi tidak bisa berhubungan dengan plain text; lebih dari itu, mereka tidak melaksanakan penyaringan multi-slot. WIEN mampu melaksanakan penyaringan multislot, tetapi memerlukan halaman-halaman yang sangat terstruktur. STALKER satu-satunya sistim yang mampu berhadapan dengan halaman-halaman yang memperlihatkan suatu struktur yang sangat kompleks; secara khusus, ia mampu berhubungan dengan halaman-halaman yang berisi informasi yang tersusun secara hirarkis ke dalam satu jumlah tingkatan yang tidak tentu.

4. DAFTAR PUSTAKA

1. D.W. Embley et al., Conceptual-model-based data extraction from multiple-record Web pages, *Data and Knowledge Engineering* **31**(3) (1999), 227–251
2. L. Eikvil, Information extraction from World Wide Web – a survey, Technical Report 945, Norweigan computing Center, 1999.
3. N. Kusmerick, Wrapper induction: efficiency and expressiveness, *Artificial Intellegence Journal* **118**(1–2) (2000), 15–68.
4. Robert Baumgartner, Sergio Flesca, and Georg Gottlob, Supervised Wrapper Generation with Lixto, www.vldb.org/conf/2001/P715.pdf , 2001, akses : Maret 2008.
5. Sabine Jabbour and Anne-Marie Vercoustre, Wrapping Web Pages into XML Documents: A Practical Experience and Comparison of Two Tools, http://ausweb.scu.edu.au/aw02/papers/refereed/vercoustre/paper.html**,** akses : Juni 2008
6. Sergio Flesca, Giuseppe Manco, Elio Masciari, Eugenio Rende dan Andrea Tagarelli, “Web wrapper induction: a brief survey”, AI Communications volume 17, 2004, hal. 57–61
7. S. Soderland, Learning information extraction rules for semistructured and free text, *Machine Learning* 34(1–3) (1999), hal 233–272.
8. Thomas M. Breuel, Information Extraction from HTML Documents by Structural Matching, www.csc.liv.ac.uk/~wda2003/Papers/Section_I/Paper_3.pdf, 2003, akses : Maret 2008
9. Valter Crescenzi, Giansalvatore Mecca, dan Paolo Merialdo, RoadRunner: Towards Automatic Data Extraction from Large Web Sites, Proc. VLDB’01 Conf., 2001, hal. 109–118