ARSITEKTUR SISTEM EKSTRAKSI INFORMASI ADAPTIF

Kuspriyanto[(1)] dan **Kurnia Muludi** [(2)]

[(1)] STEI ITB Bandung
kuspriyanto@yahoo.com

[(2)] Universitas Lampung
kurnia@unila.ac.id

Abstrak

Informasi yang berlimpah di internet ini tidaklah mudah dicari dan dimanfaatkan. Teknologi Information Extraction *(IE) merupakan salah satu alternatif yang dapat menjawab tantangan ini. Pendekatan konvensional IE menggunakan* Natural Language Processing (NLP) *terkendala masalah portabilitas, skalabilitas dan adaptifitas. Sistem ekstraksi yang ada umumnya tidak dapat mengakomodasi perubahan isi atau perubahan tampilan situs target. Dalam paper ini akan diusulkan arsitektur sistem ekstraksi informasi yang adaptif.*

Kata Kunci*: ekstraksi informasi , sistem adaptif, arsitektur, pembelajaran mesin, ontology.*

1. PENDAHULUAN

Pertumbuhan informasi yang sangat cepat pada World Wide Web menyebabkan jumlah informasi yang tersedia sangat berlimpah sehingga menyebabkan masalah *information overloaded*. Pemilahan dan pemilihan informasi sesuai dengan kebutuhan dari pool sumber informasi yang sangat besar merupakan suatu tantangan yang mesti dihadapi. Query menggunakan *Database-style* merupakan cara yang sangat efektif..

Namun masalahnya Web bukanlah database. Kebanyakan pencarian pada saat ini menggunakan *keyword* dan hasil pencarian me-*retrieve* dokumen, bukan *record*. Dalam hubungan ini *Information Retrieval* (IR) digunakan untuk mencari dokumen yang dibutuhkan berdasarkan *keyword*, tetapi tidak dapat menjawab query. Tujuan dari *Information Extraction* (IE) adalah untuk mencari informasi spesifik yang diinginkan dari suatu teks yang biasanya ditulis dalam *Natural Language* (NL) dan menyimpannya kedalam suatu bentuk yang cocok untuk pencarian dan pemrosesan otomatis.

Secara umum ada beberapa pendekatan dalam teknologi *Information Extraction*. Pada *Information Extraction* yang berbasis *Natural Language Processing* (NLP), *syntax* & *semantic constraints* digunakan untuk mengidentifikasi informasi yang relevan dalam dokumen. Pendekatan dengan cara ini terkendala masalah portabilitas dan skalabilitas [23].

Pemanfaatan teknik-teknik *Machine Learning* (ML) dalam menurunkan *extraction rule* pada *Information Extraction* telah dilakukan banyak peneliti [22]. *Extraction rules* yang telah diperoleh biasanya hanya berlaku untuk suatu domain tertentu dan sangat sensitif terhadap perubahan sumber informasi. Misalnya pada *wrapper* yang dikembangkan untuk ditargetkan kepada suatu halaman web yang tadinya bekerja dengan baik, begitu menghadapi perubahan konten atau *layout* pada halaman target tersebut, *wrapper* menghasilkan informasi yang salah atau bahkan tidak dapat bekerja. Karena konstruksi *wrapper* secara manual sangat tidak efisien dan mahal, maka dibutuhkan suatu mekanisme bagaimana

mengkonstruksi *wrapper* secara otomatis sekaligus mempunyai sifat yang adaptif terhadap perubahan sumber informasi.

2. HASIL-HASIL PENELITIAN TERKAIT

Suatu *Information Extraction System* yang adaptif diantaranya dicirikan oleh sifat-sifat *Accurate*, *Resilient,* dan *Self-repairing* [12]. Sifat *Accurate* menunjukkan kemampuan mengekstrak data dengan benar. Sifat *Resilient* menunjukan system dapat tetap bekerja secara baik meski halaman web berubah. Sifat *Self-repairing* menunjukkan kemampuan system memperbaiki *extraction rules* ketika suatu halaman web berubah.

Dari *requirements* di atas, maka pada Tabel 2 disajikan hasil analisis tingkat adaptifitas beberapa sistem *Information Extraction*.

Tabel 1. Analisis tingkat adaptifitas beberapa sistem *Information Extraction*

Pendekatan		**Accurate**	**Resilient**	**Self-repairing**
Languages	Minerva [7]	Na	T	T
	TSMMIS [13]	Na	T	T
	Web-OQL [2]	Na	T	T
HTML-aware	W4F [20]	Na	T	T
	XWRAP [18]	Na	T	T
	RoadRunner [6]	-	T	T
NLP-based	WHISK [21]	-	T	T
	RAPIER [4]	++	T	T
	SRV [10]	++	T	T
Induction	BWI [9]	+++	T	T
	SoftMealy [14]	++++	T	T
	STALKER [19]	++++	T	T
	$(LP)^2$ [5]	++++	T	T
	DataProg [17]	+++	T	Y
Modeling-based	NoDoSE [1]	na	T	T
	DEByE [16]	na	T	T
Visual Based	[3]	++++	T	T
	[11]	-	T	T
Ontology-based	BYU [8]	++++	Y	T

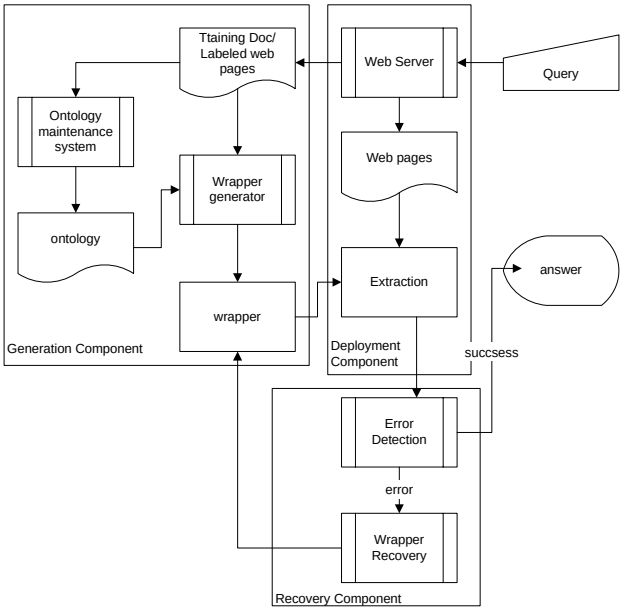
Keterangan: na=tidak tersedia data, -= <70%, + = 71-75%, ++ = 76-80%, +++= 81-85%, ++++= >85%; Y=Ya, T=Tidak

Dari analisa pada Tabel 1 dapat dilihat bahwa sebagian besar pendekatan dalam *Information Extraction* tidak memenuhi kriteria *Adaptive Information System*. Hal tersebut disebabkan system-system tersebut sebagian besar tidak bersifat *resilient* dan tidak *self-repairing*. Dari sisi akurasi SoftMealy [14], STALKER [19], (LP)2 [5], [3] dan BYU [9] menunjukkan hasil yang paling baik. Sebagian besar system bersifat *non-resilient* kecuali BYU [8] karena berbasis ontology yang tidak terpengaruh oleh perubahan halaman web. Hanya DataProg [17] mempunyai sifat *self-repairing* karena mempunyai *wrapper verification system*. System ini dapat memonitor validitas data yang dihasilkan *wrapper* dan dapat secara otomatis melakukan proses perbaikan *wrapper* menggunakan *wrapper re-induction system*

Untuk dapat memenuhi kriteria sistem ekstraksi informasi yang adaptif, maka dalam paper ini akan diusulkan arsitektur alternatif yang sesuai dengan kriteria di atas.

3. DESAIN SISTEM DAN ANALISIS

Arsitektur sistem yang diusulkan disajikan pada Gambar 1.



Gambar 1. Arsitektur sistem ekstraksi informasi adaptif

Secara umum sistem terdiri dari tiga komponen utama yaitu *Generation Component, Recovery Component,* dan *Deployment Component*. *Generation Component* berfungsi untuk menurunkan *extraction rules* yang dikemas dalam *wrapper* menggunakan pembelajaran mesin. Adanya ontology tentang domain target, akan digunakan sebagai *prior-knowledge* bagi algoritma pembelajaran mesin untuk men*generate wrapper* dari dokumen contoh yang tersedia (*annotated document*). Dokumen contoh juga digunakan untuk memperkaya ontology menggunakan *ontology learning* pada *ontology maintenance system*. *Wrapper* kemudian melakukan proses ekstraksi pada dokumen sasaran (halaman web) yang disediakan oleh *deployment component.* Hasil ekstraksi dokumen oleh *wrapper* diperiksa keabsahannya oleh *error detection unit*. Jika tidak ditemui kesalahan, maka hasil ekstraksi ditampilkan atau disimpan dalam database yang telah disiapkan atau dalam format XML. Data ini kemudian dapat digunakan untuk aplikasi selanjutnya. Sebaliknya Jika ditemui kesalahan maka system akan memicu secara otomatis sistem swa-induksi *wrapper* pada *wrapper recovery unit* pada *Recovery Component*. Proses induksi *extraction rules* dilakukan ulang sehingga didapatkan *wrapper* yang sesuai dengan target halaman web tersebut sehingga sifat *self repairing* dapat dipenuhi.

Komponen ontology dibutuhkan untuk membantu dalam proses penurunan wrapper oleh algoritma pembelajaran mesin. Menurut [15] *botleneck* dari pengembangan *adaptive information system* adalah masalah kurang tersedianya *annotated training sample*. Karena itu dibutuhkan suatu pembelajaran mesin yang membutuhkan lebih sedikit *annotated training sample.* Dengan adanya ontology maka jumlah dokumen contoh yang dibutuhkan dapat lebih kecil. Dengan ontology maka sifat *resilience* dapat dipenuhi.

4. KESIMPULAN

Kriteria sifat adaptif sistem ekstraksi informasi dapat diakomodasi melalui desain sistem yang mengimplementasi pembelajaran mesin yang tepat, pemanfaatan ontology dan *Wrapper recovery system*.

5. FUTURE WORKS

Pada masa mendatang direncanakan akan dilakukan penelitian:

- Penggabungan *inductive learning* dan *analytical learning* dalam meng*generate wrapper*.
- Pengembangan algoritma re-induksi untuk memperbaiki kesalahan wrapper.
- Optimasi jumlah *annotated training sample* untuk menghasilkan *extraction rules*.
- Pengembangan Algoritma *ontology learning* menggunakan *corpus* sebagai input.

6. DAFTAR PUSTAKA

[1] Adelberg, B. NoDoSE - A tool for semi-automatically extracting structured and semistructured data from text documents. In Proceedings of the ACM SIGMOD International Conference on Management of Data (Seattle, WA, 1998), pp. 283-294.

[2] Arocena, G.O. and Mendelzon, A.O. WebOQL: Restructuring documents, databases, and webs. In Proceedings of the 14th International Conference on Data Engineering (Orlando, FL, 1998), pp. 24-33.

[3] Y. Aumann, R. Feldman, Y. Liberzon, B. Rosenfeld, and J. Schler. Visual information extraction. Knowledge and Information Systems, 10(1):1{15, 2006.

[4] Califf, M.E. and Mooney, R.J. Relational Learning of Pattern-Match Rules for Information Extraction. In Proceedings of the Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence (Orlando, FL, 1999), pp. 328-334.

[5] Ciravegna, F. (LP)2, an adaptive algorithm for information extraction from Web-related texts. In IJCAI-2001 Workshop on Adaptive Text Extraction and Mining. Seattle, USA, 2001.

[6] Crescenzi, V., Mecca, G., and Merialdo, P. RoadRunner: Towards automatic data extraction from large Web sites. In Proceedings of the 26th International Conference on Very Large Data Bases (Rome, Italy) , 2001, pp. 109-118.

[7] Crescenzi, V. and Mecca, G. Grammars have exceptions. Information Systems 23, 8 (1998), 539-565.

[8] Embley, D., Campbell, D., Smith, R., and Liddle, S. Ontology-based extraction and structuring of information from data-rich unstructured documents. In *Proceedings of the Conf. on Info. and Knowledge Management* (Nov. 1998), 52–59.

[9] Freitag, D. and Kushmerick, N. Boosted wrapper induction. In AAAI/IAAI, pp. 577–583. 2000.

[10] Freitag, D. Toward general-purpose learning for information extraction. In Christian Boitet and Pete Whitelock, eds., Proc. 36th Annual Meeting of the Association for Computational Linguistics, pp. 404–408. San Francisco, CA, 1998.

[11] Gatterbauer, W., Bohunsky, P., Herzog, M., Krupl, B., and Pollak, B. Towards Domain Independent Information Extraction from Web Tables. WWW 2007, May 8–12, 2007.

[12] Greg, D.G. and Walczak, S. Adaptive Web Information Extraction. Comm. Of the ACM, May 2006, Vol. 49, No. 5.

[13] Hammer, J., McHvoh, J., and Garcia-Molina, H. Semistructured data: The TSIMMIS experience. In Proceedings of the First East-European Symposium on Advances in Databases and Information Systems (St. Petersburg, Russia, 1997), pp. 1-8.

[14] Hsu, C.N. and Dung, M.T. Generating finite-state transducers for semi-structured data extraction from the Web. Information Systems 23, 8 (1998), 521-538.

[15] Kushmerick, N. and Thomas, B. Adaptive information extraction: Core technologies for information agents. In Intelligent Information Agent R&D in Europe: An AgentLink Perspective, 2002.

[16] Laender, A.H.F., Ribeiro-Neto, B., and Da Silva, A.S. DEByE - Data Extraction By Example. Data and Knowledge Engineering 4O, 2, 2002, 121-154.

[17] K. Lerman, S. Minton, and C. A. Knoblock. Wrapper maintenance: A machine learning approach. Journal of Artif. Intell. Research (JAIR), 18:149–181, 2003.

[18] Liu, L., Pu, C., and Han, W. XWRAP: An XML-enabled wrapper construction system for Web information sources. In Proceedings of the 16th International Conference on Data Engineering (San Diego, CA) , 2001, pp. 611-621.

[19] Muslea, I., Minton, S., and Knoblock, C.A. Hierarchical wrapper induction for semistructured information sources. Autonomous Agents and Multi-Agent Systems, 4(1/2):93–114, 2001.

[20] Sahuguet, A. and Azavant, F. Building intelligent Web applications using lightweight wrappers. Data and Knowledge Engineering 36, 3 (2001), 283-316.

[21] Soderland, S. Learning information extraction rules for semi-structured and free text. Machine Learning, 34(1–3):233–272, 1999.

[22] Turmo, J., Ageno, A., and Catala, N. Adaptive Information Extraction. ACM Computing Surveys, Vol. 38, No. 2, Article 4, Publication date: July 2006.

[23] Yildiz, B. and Miksch, S. Motivating Ontology-Driven Information Extraction. Proceeding International on Semantic Web and Digital Libraries (ICSD-2007).