ANALISIS 10 (1), 161-180 (2003)

APLIKASI TEKNIK RANGKAIAN NEURAL DALAM CAPAIAN TEKS

FADHILAH MAT YAMIN
FADZILAH SIRAJ
WAN ROZAINI SHEIKH OSMAN
Fakulti Teknologi Maklumat
Universiti Utara Malaysia

ABSTRAK

Capaian maklumat yang berkesan merupakan keperluan utama pengguna maklumat. Hal ini kerana limpahan maklumat menyebabkan enjin carian maklumat semasa sukar untuk mencapai dan menyaring maklumat yang relevan dengan keperluan pengguna. Oleh itu, kertas kerja ini membincangkan aplikasi kepintaran buatan dalam bidang capaian maklumat. Penggunaan teknik kepintaran buatan dalam capaian maklumat akan menghasilkan sistem capaian maklumat pintar yang dapat membantu dan memudahkan carian maklumat. Selain itu, model bagi sistem capaian maklumat pintar juga turut dibincangkan.

Kata Kunci: *Kepintaran Buatan, Capaian Maklumat Pintar, Enjin Carian.*

ABSTRACT

Effective information retrieval is a major requirement for most users. This is because too much information limits the current search engine capability in filtering the document. Therefore, this paper discussed the applications of Artificial Intelligence in information retrieval system. The application of Artificial Intelligence techniques in information retrieval will produce an intelligent information retrieval system to enhance the retrieval processes. In addition, a model for intelligent information retrieval system is also discussed.

Key Words: *Artificial Intelligence, Intelligent Information Retrieval, Search Engine.*

- Pencarian maklumat lebih interaktif dan pengguna boleh menentukan penyusunan hasil carian berdasarkan kepada penilaian pengguna.

Kertas kerja ini membincangkan penggunaan teknik kepintaran buatan dalam pembangunan sistem capaian maklumat iaitu rangkaian neural, logik kabur, algoritma genetik, agen pintar, sistem pakar, perlombongan data, pembelajaran mesin dan pemprosesan bahasa tabii. Selain itu, model sistem capaian maklumat pintar turut dibincangkan.

TEKNIK KEPINTARAN BUATAN DALAM CAPAIAN MAKLUMAT

Kepintaran Buatan atau *Artificial Intelligence* (AI) merupakan salah satu cabang sains komputer yang dibangunkan berdasarkan kepada ciri-ciri kepandaian manusia. AI boleh didefinisikan sebagai kebolehan untuk memahami, memikir, menangani, kecepatan dalam pembelajaran, berakal, kebolehan untuk memilih dan menyesuai, pandai dan juga pengambilan maklumat (Marzuki, 1994). Pelbagai teknik dalam AI telah dibangunkan dengan mengadaptasikan ciri-ciri kepintaran manusia seperti Rangkaian Neural, Logik Kabur, Algoritma Genetik, Perlombongan Data, Pembelajaran Mesin dan Pemprosesan Bahasa Tabii.

Kajian menunjukkan penggabungan antara teknik capaian maklumat semasa dengan teknik AI dapat menghasilkan keputusan yang lebih baik. Bhandarkar *et al.*, (1989) misalnya menyatakan bahawa capaian dan klasifikasi maklumat menggunakan teknik AI memberikan hasil yang lebih cekap. Bhandarkar *et al.* menilai penggunaan AI dalam proses pengkategorian dan capaian dokumen. Penggunaan teknik AI didapati boleh meningkatkan potensi dalam capaian maklumat. CODER (France and Fox, 1998) pula merupakan salah satu kajian yang mengaplikasikan teknik AI untuk meningkatkan keberkesanan sistem capaian maklumat. Tumpuan diberikan kepada menganalisis perwakilan dokumen yang berbagai-bagai jenis seperti email atau mesej yang bervariasi dari segi gaya, panjang, topik dan struktur.

Chen (1995) pula membuat tinjauan terhadap beberapa teknik AI iaitu Rangkaian Neural, Pembelajaran Simbolik dan Algoritma Genetik dalam sistem capaian maklumat. Chen berpendapat bahawa penggunaan AI dalam capaian maklumat dapat membuka lebih

Walau bagaimanapun, kajian yang dilakukan oleh Fabio Crestani dari Universita' di Padova, Itali menunjukkan hasil yang negatif (Crestani, 1993a, Crestani, 1993b, Crestani, 1994) iaitu NN tidak dapat belajar dan mengitlakkan ciri-ciri pengetahuan dalam aplikasi domain berkaitan. Crestani (1995) pula mencadangkan kajian semula ke atas reka bentuk NN dan algoritma pembelajaran yang digunakan. Penyelidik mendapati kegagalan ini bukan sahaja disebabkan oleh jumlah maklumat yang dimasukkan ke dalam sistem terlalu besar tetapi juga corak data input yang mengelirukan Kebanyakkan input data yang sama didapati mempunyai target yang berbeza. Oleh sebab itu, rangkaian terpaksa belajar menyesuaikan corak input dengan target. Keadaan ini menyukarkan NN untuk belajar dan menyebabkan rangkaian terkeliru. Kekangan yang dihadapi oleh Crestani adalah disebabkan oleh input data dan bukannya disebabkan oleh NN atau algoritma pembelajaran yang digunakan. Hal ini adalah kerana corak input yang sama dan target yang berbeza juga boleh mengelirukan manusia.

Kajian oleh Nur Izura *et al.* (1997) mengenengahkan penggunaan rangkaian *Counterpropagation* bagi capaian maklumat dari pangkalan data. Rangkaian ini merupakan gabungan antara dua pendekatan yang berbeza iaitu pendekatan Kohonen dan Grossberg. Kohonen merupakan pendekatan tak selia, manakala Grossberg pula merupakan pendekatan terselia. Kajian tersebut mendapati pendekatan yang digunakan berjaya mengecam data daripada pangkalan data dengan ketepatan yang lebih daripada 70 peratus bergantung kepada peratusan kecacatan input data (*noise*).

Keupayaan NN untuk belajar dan mengklasifikasikan corak tanpa selia juga merupakan satu kelebihan NN berbanding teknik lain. Rangkaian *Self Organizing Map* (SOM) merupakan rangkaian tanpa selia yang popular dan sering digunakan dalam pelbagai aplikasi. Sarjon dan Md. Nor (2001) mengaplikasikan WEBSOM iaitu satu pendekatan capaian maklumat berasaskan SOM bagi menyusun dan mencapai balik dokumen dari koleksi dokumen. WEBSOM berupaya untuk memetakan, memberikan gambaran ringkas tentang koleksi dokumen dan menyediakan kemudahan untuk pencarian interaktif. Dengan menggunakan kaedah ini dokumen dapat disusun berdasarkan kepada kesesuaian kandungan dokumen tersebut. Penyelidikan terhadap penggunaan WEBSOM telah dipelopori oleh WEBSOM Research Group[1] yang diketuai oleh Teuvo Kohonen[2]. Penggunaan dan aplikasi pendekatan tersebut telah dibincangkan dalam pelbagai kertas penyelidikan; antaranya ialah Lagus (2000), Kohonen (1997), Lagus

Agen Pintar

Agen pintar merupakan salah satu teknik AI yang sering mendapat perhatian penyelidik. Jansen (1996) mendefinisikan agen sebagai sebuah perisian komputer yang mengimplementasikan matlamat pengguna. Moukas (1996) pula mendefinisikan agen sebagai program separa-pintar yang membantu pengguna melakukan operasi yang berulangan dan memakan masa yang banyak. Oleh itu, agen boleh ditakrifkan sebagai perisian komputer yang dibangunkan bagi membantu pengguna mencapai matlamat yang sukar dicapai melalui kaedah biasa. Agen pintar pula boleh didefinisikan sebagai agen yang mempunyai ciri-ciri kepintaran. Ia merupakan pendekatan interaktif bagi membantu pengguna menggunakan dan membuat carian maklumat. Pendekatan ini juga akan meningkatkan keupayaan enjin carian semasa (Jansen, 1996). Kejayaan penyelidikan dalam bidang ini boleh dinilai melalui pembangunan beberapa prototaip seperti IfWeb dan CiFi.

Agen juga boleh dilatih berdasarkan kepada maklum balas daripada pengguna. Balabanovic dan Shoham (1995) membangunkan agen yang berupaya untuk mencadangkan senarai dokumen kepada pengguna dan belajar daripada maklum balas pengguna. Seterusnya, agen tersebut akan memperbaharui senarai cadangannya apabila pengguna memasuki semula sistem tersebut. Keupayaan agen untuk belajar telah dan cuba diaplikasikan oleh ramai penyelidik lain. Pazzani *et al.* (1995) mengenengahkan pembangunan agen pintar yang mempunyai kebolehan untuk mencari maklumat di WWW dan mencadangkan halaman yang mungkin sesuai dengan pengguna. Agen tersebut belajar dengan menganalisis hiperhubungan yang dicapai oleh pengguna. Berdasarkan kepada hiperhubungan tersebut agen akan memberikan senarai cadangan dokumen yang berkaitan. Melalui kaedah ini juga, agen akan mengenalpasti dan mempelajari profail pengguna dan menentukan maklumat yang mungkin diperlukan oleh pengguna. Azman *et al.* (2001) juga menggunakan profail pengguna sebagai sumber pengetahuan kepada agen yang dibangunkan. Izhar *et al.* (2001) pula mencadangkan penggunaan agen peribadi bagi membantu pengguna membuat carian dengan lebih interaktif. Agen tersebut akan membuat carian berdasarkan kepada keperluan pengguna dan merekodkan setiap capaian pengguna bagi tujuan rujukan akan datang. Contoh agen peribadi ialah Amalthaea (Moukas, 1996).

Di samping itu, terdapat banyak lagi agen pintar yang dibangunkan dengan menggabungkan atau mengaplikasikan pendekatan lain bagi

berpendapat bahawa pemprosesan bahasa tabii sebenarnya kurang berkesan terutama apabila penumpuan yang lebih diberikan untuk mencapai indeks dan perwakilan pertanyaan yang baik. Hal ini kerana penggunaan pendekatan lingualistik biasa tidak dapat meningkatkan ketepatan capaian maklumat. Oleh itu, usaha sedang ditingkatkan bagi mencari strategi yang terbaik untuk meningkatkan penggunaan NLP dalam capaian maklumat. Strzalkowski *et al.* (1996) juga melaporkan bahawa kajian awal mereka menunjukkan peningkatan yang menarik.

NLP merupakan teknologi pelengkap kepada kebanyakkan sistem yang melibatkan interaksi di antara pengguna dengan sistem. Lazimnya, keupayaan komputer untuk memahami dan menterjemah keperluan pengguna agak terhad. Hal ini kerana komputer dicipta hanya untuk memahami isyarat ringkas iaitu 0 atau 1 sahaja. Penggunaan NLP membolehkan pengguna berinteraksi dengan sistem dalam bentuk yang lebih mudah. Askjeeves.com (*http://www.askjeeves.com/*) misalnya, merupakan antara enjin carian maklumat yang mengimplementasi teknik NLP dalam pencarian maklumat. Pengguna dibenarkan memasukkan pertanyaan dalam bentuk teks selain daripada penggunaan kata kunci biasa.

Croft dan Lewis (1987) menggunakan pendekatan NLP bagi memadankan pertanyaan pengguna dengan maklumat yang disimpan. Dalam kajian tersebut, NLP digunakan bagi membentuk "*conceptual case frames*" dan perbandingan dibuat antara model tersebut dengan dokumen yang disimpan. Pelbagai pendekatan telah digunakan bagi mengimplementasi NLP dalam capaian maklumat. Feinstein *et al.* (1997) misalnya menggunakan teknik '*Shallow Parsing*' bagi menyusun dokumen di web berasaskan kepada kombinasi tiga teknik iaitu sintaktik *parsing*, heuristik dan statistik. Fokus kajian mereka ialah untuk menyusun halaman yang bersesuaian dengan kehendak pengguna.

Sistem Pakar

Sistem pakar merupakan salah satu daripada teknik AI yang paling mendapat perhatian terutama dalam bidang penyelesaian masalah. Sistem pakar menggunakan perwakilan pengetahuan berdasarkan pengetahuan pakar atau *domain expert* dalam satu bidang lapangan khusus. Pengetahuan pakar tersebut diwakilkan dalam sistem dan akan digunakan sebagai rujukan apabila pengguna menggunakan sistem tersebut. Antara kajian awal penggunaan sistem pakar dalam capaian maklumat ialah Watters *et al.* (1987). Beliau menggunakan

menggunakan pengetahuan tersebut bagi menyelesaikan masalah yang lain. Berdasarkan kepada maklumat terkini yang diperoleh, komputer akan belajar corak masalah tersebut dan mengemaskini pangkalan pengetahuannya. Bloedorn *et al.* (1996) menggunakan teknik pembelajaran mesin bagi membangunkan sebuah sistem untuk menjana profail pengguna yang mudah difahami dan dapat mengenalpasti keperluan pengguna melalui interaksi yang minimum dengan pengguna. Selain itu, *Reinforcement Learning*, iaitu salah satu daripada teknik pembelajaran mesin telah diimplementasi dalam *Cora.whizbang.com* (McCallum *et al.*, 2000).

METODOLOGI KAJIAN

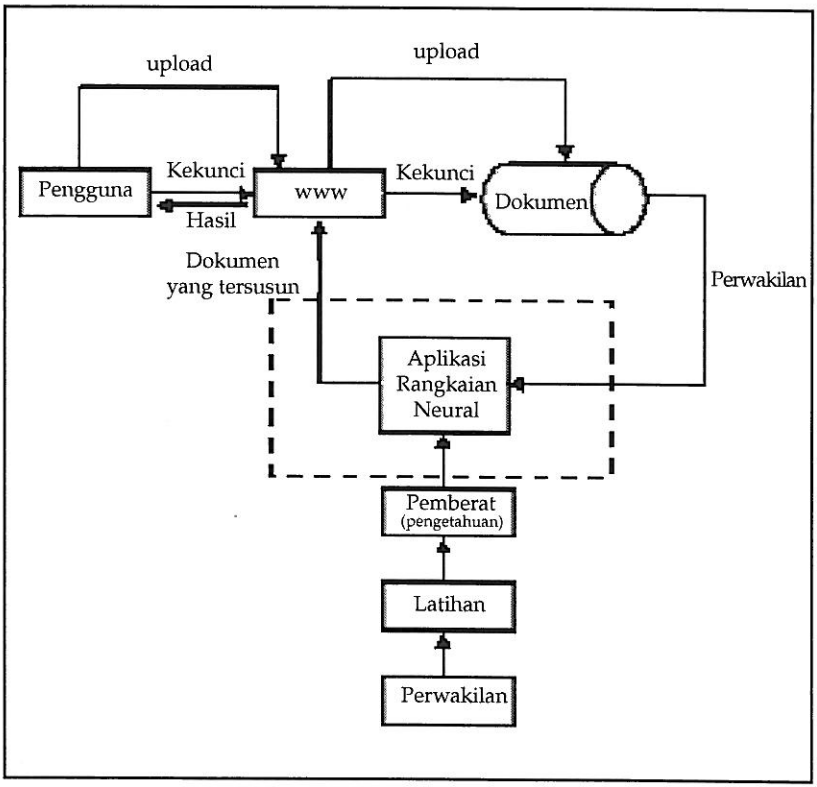
Penyelidikan ini melibatkan empat fasa iaitu pemilihan dan perwakilan attribut, penentuan parameter NN, implementasi pendekatan NN dan pembangunan prototaip. Sehubungan dengan itu, sebanyak lapan attribut telah dipilih iaitu URL, tajuk dokumen, abstrak, kata kunci, pengenalan, kandungan, kesimpulan dan bibliografi (Fadhilah and Fadzilah, 2001). Attribut tersebut diwakilkan ke dalam bentuk binari iaitu 1 atau 0. Bagi tujuan latihan dalam NN, beberapa parameter ditentukan iaitu bilangan unit tersembunyi, benih pemberat yang terlibat, kadar pembelajaran dan momentum. Attribut kemudiannya dimasukkan ke dalam rangkaian dan dilatih bagi menghasilkan pemberat, iaitu pengetahuan yang akan digunakan oleh prototaip sistem capaian maklumat.

MODEL SISTEM

Penyelidikan ini ditumpukan kepada carian maklumat dalam bentuk teks. Oleh itu, beberapa attribut telah dikenal pasti (Rajah 1). Kesemua attribut tersebut merupakan perwakilan unik bagi keseluruhan dokumen. Selain itu, attribut lain seperti nama pengarang, tahun penerbitan, dimana dokumen tersebut diterbitkan dan jenis dokumen juga boleh digunakan untuk menentukan kerelevanan sesuatu dokumen. Hal ini kerana maklumat tersebut mempunyai perkaitan sama ada secara langsung atau tidak langsung dengan dokumen tersebut.

Pembangunan Prototaip dibahagikan kepada dua peringkat utama iaitu pembangunan enjin carian ringkas dan implementasi rangkaian rambatan balik sebagai agen maklum balas (Rajah 2). Enjin carian

Rajah 2
Seni Bina Sistem



Implementasi rangkaian rambatan balik merupakan aplikasi tambahan utama kepada prototaip enjin carian maklumat ringkas yang dibangunkan pada peringkat awal (Rajah 3). Aplikasi ini bertujuan untuk meningkatkan keupayaan enjin carian tersebut menilai dokumen yang dipulangkan kepada pengguna. Aplikasi ini terbahagi kepada dua bahagian iaitu simulator rangkaian rambatan balik atau *Backpropagation Simulator* (BPSim) dan bahagian aplikasi rangkaian. BPSim dibangunkan secara berasingan dan digunakan untuk belajar corak data bagi menentukan nilai kerelevanan. Hasil daripada latihan tersebut iaitu pemberat (berfungsi sebagai pengetahuan kepada enjin carian) akan disimpan dan digunakan dalam aplikasi rangkaian. Bahagian aplikasi rangkaian dibangunkan dan "dipasang" dalam prototaip enjin carian. Fungsi utama bahagian ini ialah untuk mengira nilai kerelevanan setiap dokumen yang dipulangkan oleh sistem berdasarkan kepada pengetahuan yang telah disimpan (pemberat).

memainkan peranan penting dalam menghasilkan enjin carian pintar. Enjin carian pintar lazimnya mempunyai kelebihan berbanding enjin carian biasa seperti dalam pencarian dan capaian maklumat, pemadanan dokumen dengan kata kunci dan perwakilan semula dokumen.

Model sistem capaian maklumat pintar yang dibincangkan di atas merupakan satu contoh aplikasi teknik AI iaitu NN dalam sistem capaian maklumat. Prototaip bagi model tersebut masih dalam pembangunan dan dijangka akan dapat membantu meningkatkan capaian maklumat yeng relevan.

ENDNOTES

1. WEBSOM Research Group (http://websom.hut.fi/websom) merupakan sebahagian daripada Neural Networks Research Centre (NNRC) dari Helsinki University of Technology (HUT).
2. Teuvo Kohonen merupakan salah seorang penyelidik yang aktif dalam bidang Rangkaian Neural. Beliau adalah individu yang bertanggungjawab memperkenalkan Rangkaian Kohonen.

RUJUKAN

Azman Yasin, Azizi Zakaria, Fadzilah Siraj, Mohamad Shamrie Sainin & Muhammad Yusof. (2000). Capaian maklumat teks menggunakan teknik genetik algoritma. *Prosiding Persidangan Ulang Tahun Ke-30 UKM*, 5-6 September 2000 158-169. Universiti Kebangsaan Malaysia.

Azman Yasin, Mohamed Yusof, Tengku Mohamed Tengku Sembuk & Mohd Shahrul Azman Mohd Noah. (2001). An information filtering agent to learn user's preferences. *Proceedings of Artificial Intelligence Seminar (AIS) 2001 (CD-ROM)*. Universiti Utara Malaysia, Sintok, 1-3 November 2001.

Balabanovic, M. & Shoham, Y. (1995). Learning information retrieval agent: experiments with automated web browsing. *Proceedings of the (AAAI) Spring Symposium on Information Gathering from Heterogenous, Distributed Resources*, 13-18.

Bhandarkar, A., Chandrasekar, R., Ramani, S. & Bhatnagar, A. (1989). Intelligent categorization, archival and retrieval of Information. In J. Siekmann (Ed.). *Lecture Notes in Artificial Intelligence*, 309-320. New York: London.

Bloedorn, E., Mani, I. & MacMillan, R. (1996). Representational issues

(PADD97), 125-136.

France, R. K. & Fox, E.A. (1998). An artificial intelligence environment for information retrieval research. Virginia Polytechnic Ins and State University. http://cs-tr.cs.cornell.edu:80/Dienst/UI...rize/ncstrl.vatech_cs/TR-88-10?abstract, 15 February 2001.

Hayes, D. (2000). Business accelerator launches research service on internet. *TheStar* http://www.kcstar.com/item/pages/business.pat.business/37745a3d.330, 13 Januari 2002.

Haynes, T. (1998). A comparison of random serach versus genetic programming as engine for collective adaptation. In V. Wiliam Porto (Ed.). *Proceedings of the Seventh International Conference on Evolutionary Programming.*

Hui, B. (1998). Applying NLP to IR: why and how. *Nota Kuliah: Intelligence Interface.*

IEI (1999). Neural network web site placement on the altavista search engine. Imagination engines inc. URL: http//www.imagination-engines.com/altavista/pparse.htm. 04 September 2001.

Izhar Che Zainol Rashid, Fadzilah Siraj & Nur Azzah Abu Bakar. (2001). Personalized web agent approach for web content mining. *Artificial Intelligence Seminar (AIS) 2001 (CD-ROM)*, 1-3 November Sintok: Universiti Utara Malaysia.

Jansen, J. (1996). Using an intelligent agent to enhance search engine performance. *FirstMonday-Peer-Reviewed Journal on the Internet.* http://www.firstmonday.dk/issues2_3/jansen/, 05 September.

Kaski, S. (1999). Fast winner search for SOM-based monitoring and retrieval of high-dimensional data. *Proceedings of ICANN99, Ninth International Conference on Artificial Neural Networks*, 2, 940-945, IEE. London.

Kaski, S., Honkela, T., Lagus, K. & Kohonen, T. (1996). Creating an order in digital libraries with self-organizing maps. In *Proceedings of WCNN'96, World Congress on Neural Networks*, September 15-18, San Diego, California, 814-817. Lawrence Erlbaum and INNS Press, Mahwah, NJ.

Kohonen, T. (1997). Exploration of very large databases by self-organizing maps. *In Proceedings of ICNN'97, International Conference on Neural Networks*, pages PL1-PL6. IEEE Service Center, Piscataway, NJ

Koskela, M., Laaksonen, J., Laakso, S. & Oja, E. (2000). The PicSOM Retrieval System: Description and Evaluations. In *Proceedings of Challenge of Image Retrieval (CIR 2000).*Brighton,UK.May2000.http://www.cis.hut.fi/picsom/publications.html.

Laakso, S., Laaksonen, J. Koskela, M. & Oja, E. (2001). Self-organizing maps of web link information. In N.Allinson, H.Yin, L.Alliason

McCallum, A.K., Nigam, K., Rennie, J. & Seymore, K. (2000). Automating the construction of internet portals with machine learning. *Information Retrieval Journal*, 3, 127-163. Kluwer Academic Pub.

Moukas, A. (1996). Amalthaea: information discovery and filtering using a multiagent evolving ecosystem. *Proceedings of the Conference on Practical Applications of Agents and Multiagent Technology*. London.

Muniyandi, R.C. (2000). Neural networks: an exploration in document retrieval system. *TENCON Proceedings: Intelligent Systems and Technologies for the New Millennium* 24-27 September, I, 156-161.

Nur Izura Hj. Udzir, Md. Nasir Sulaiman, Ali Mamat, Ramlan Mahmod & Fatimah Ahmad. (1997). Rangkaian neural dalam dapatan pangkalan data. *National Conference on Research and Development in Computer Science and its Applications*, 93-97.

Pannu, A. & Sycara, K. (1996). Learning text filtering preferences. *In 1996 AAAI Symposium on Machine Learning and Information Access*, http://citeseer.nj.nec.com/49897.html.

Pazzani, M., Nguyen, L. & Mantik, S. (1995). Learning from hotlists and coldlists: towards a www information filtering and seeking agent. *In Proceedings of the Seventh International Conference on Tools with Artificial Intelligence*, http://www.ics.uci.edu/~pazzani/Coldlist.pdf.

Rijsbergen, C. J.V. (1979). *Information Retrieval* (2nd). Butterworths: London.

Sarjon Defit & Mohd Nor Md Sap (2001). Organizing and searching collections of textual documents using WEBSOM Method. *Artificial Intelligence Seminar (AIS) 2001 (CD ROM)*, 1-3 November, Sintok: Universiti Utara Malaysia.

Shavlik, J. & Rad, T.E. (1998). Building intelligent agents for web-based tasks: a theory-refinement approach. *Proceedings of the CONALD Workshop on Learning from Text and the Web*, June 1998.

SoloSearch (2000). Mid-West Start-Up Introduces An "Intelligent Search" Manager To Eliminate Internet Information Overload. *Press Release March 28, 2000*. http://www.solosearch.com/press.asp#mid.

Strzalkowski, T., Guthrie, L., Karlgren, J., Leistensnider, J., Lin, F., Perez-Carballo, J., Straszheim, T., Wang, J. & Wilding, J. (1996). Natural language information retrieval: TREC-5 report. In E.M. Voorhees, & Harman D.K. (Eds.). *Proceedings of the Fifth Text Retrieval Conference (TREC-5)* 20-22 November, 291-314.

Theilmann, W. & Rothermel, K. (1998). Domain experts for information retrieval in the world wide web. *Proceeding. 2nd International. Workshop on Cooperative Information Agents (CIA'98)*, Paris, July 4-7, 1998. In M. Klusch, G. Weiß (Eds.). *Lecture Notes in Artificial In-*