University of Maribor
Faculty of Organizational Sciences

DOCTORAL DISSERTARTION

Automatic ontology generation from web tabular structures

Aleksander Pivk

Thesis Committee:
Vladislav Rajkovič, Mentor
Miroljub Kljajić, Chair
Matjaž Gams, Jozef Stefan Institute

June 2005

Abstract

Efficient automatic information handling has become increasingly important in information society. Most information on the Web is presented in the form of semi-structured or unstructured documents, encoded as a mixture of loosely structured natural language text and template units. The lack of metadata, which would precisely annotate the structure and semantics of documents, and ambiguity of natural language makes automatic computer processing very complex. The Semantic Web aims to overcome this bottleneck.

The central contribution of the dissertation presents a novel method for automatic generation of knowledge models such as ontologies from arbitrary tabular structures found on the Web. An accompanying implementation is reflected in a system named TARTAR (Transforming ARbitrary TAbles into fRames) which is a component of the multi-agent system OntoGeMS (Ontology Generation Multi-agent System). The method is based on a grounded cognitive table model introduced by Hurst. The methodology is stepwise instantiated in four steps. In the first step, a table is transformed into regular matrix form. In the following two steps the table is handled from a structural and functional, and in the last step from a semantic point of view. The outcome of the method is threefold: a knowledge frame, an ontology, and a knowledge base, all encoded in an F-Logic representation language. The frame makes explicit the meaning of cell contents, the functional dimension of the table which is comparable to the relational schema, and the meaning of the table based on its structure. In the ontology the concepts are arranged into a directed acyclic graph, where the arcs represent relations among concepts and also the types of relations. The table content is formalized according to the frame into the knowledge base.

The empirical evaluation is performed from four perspectives. The efficiency of the method is measured according to the portion of correctly transformed tables belonging to two domains, tourist and geopolitical, enabling us to prove the domain independency of the approach. Usability of the approach clearly shows the syntactic and semantic correctness of generated frames that are compared to the manually annotated frames. Approach applicability is shown from two views. By querying the content of tables encoded in the knowledge base, it is shown that returned answers are true and complete in all cases. The querying is enabled by the use of the inference engine OntoBroker. In the last case of the evaluation we make use of the automatically generated ontologies for automatic construction of wrappers. Ontologies generated in this way can substitute hand-crafted heuristics that are used as a foundation for wrapper construction tasks. The benefits are clearly shown in terms of better adaptability, easier extensibility, and domain independence.

The present research work opens a number of potential for further research in information handling and the promotion to the Semantic Web.

Keywords:

ontology
ontology learning/generation
semantic web
tabular strucuture
information extraction
intelligent agent

Bibliography

[1] GoogleSets, http://labs.google.com/sets.

[2] W3C Consortium. HTML Specification 4.01, http://www.w3.org/TR/html4/.

[3] W3C Consortium. Resource Description Framework (RDF): Model and Syntax Specification, www.w3.org/TR/REC-rdf-syntax-19990222/.

[4] W3C Consortium. Document Object Model (DOM), http://www.w3.org/DOM/.

[5] CyberNeko HTML Parser, http://www.apache.org/~andyc/neko/doc/html.

[6] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison Wesley, 1995.

[7] B. Adelberg. 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, Washington, pp. 283-294, 1998.

[8] E. Agirre, O. Ansa, E. Hovy, and D. Martinez. Enriching Very Large Ontologies using the WWW. In *Proceedings of the 1st Workshop on Ontology Learning, ECAI 2000*, Berlin, Germany, 2000.

[9] R. Agrawal, T. Imielinski, and A. Swani. Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 207-216, 1993.

[10] A. Antonacopoulos and J. Hu, eds. *Web Document Analysis: Challenges and Opportunities*. World Scientific, 2004.

[11] G.O. Arocena and A.O. Mendelzon. WebOQL: Restructuring Documents, Databases, and Webs. In *Proceedings of the 14th International Conference on Data Engineering*, Orlando, Florida, pp. 24-33, 1998.

[12] M. Atkinson, F. Bancilhon, D. Dewitt, K. Dittrich, D. Maier, and S. Zdonik. The Object-oriented Database System Manifesto. In *Proceedings of the International Conference on Deductive and Object-Oriented Databases*, Elsevier Science, pp. 40-57, 1989.

[13] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 544 pages, 1999.

[14] A. Belaïd. Recognition of Table of Contents for Electronic Library Consulting. *International Journal of Document Analysis and Recognition*, 4: 35-45, 2001.

[15] V.R. Benjamins, J. Contreras, O. Corcho, and A. Gómez-Pérez. Six Challenges for the Semantic Web. *AIS SIGSEMIS Bulletin*, 1(1): 24-25, 2004.

[16] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 2001.

[17] J. Biskup and D.W. Embley. Extracting Information from Heterogeneous Information Sources using Ontologically Specified Target Views. *Information Systems*, 28(3): 169-212, 2003.

[18] W.N. Borst. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. PhD thesis, Centre for Telematica and Information Technology, University of Tweenty, Enschede, The Netherlands, 1997.

[19] J.M. Bradshaw, ed. *Software Agents*. AAAI Press, Menlo Park California, 1997.

[20] M. Califf and R. Mooney. Applying ILP-based Techniques to Natural Language Information Extraction: An Experiment in Relational Learning. In *Working Notes of the IJCAI-97 Workshop on Frontiers in Inductive Logic Programming*, 1997.

[21] M.E. Califf and R.J. Mooney. Relational Learning of Pattern-Match Rules for Information Extraction. In *Proceedings of the 16th International Conference on Artificial Intelligence*, Orlando, Florida, pp. 328-334, 1999.

[22] S. Ceri, G. Gottlob, and L. Tanca. *Logic Programming and Databases*. Springer, 1990.

[23] S. Chakrabarti. *Mining the Web: Analysis of Hypertext and Semi Structured Data*. Morgan Kaufmann, 344 pages, 2002.

[24] H. Chen, S. Tsai, and J. Tsai. Mining Tables from Large Scale HTML Texts. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pp. 166-172, 2000.

[25] P. Cimiano, S. Staab, and J. Tane. Automatic Acquisition of Taxonomies from Text: FCA meets NLP. In *Proceedings of the PKDD/ECML'03 International*

Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia, 2003.

[26] P. Cimiano, A. Hotho, and S. Staab. Clustering Concept Hierarchies from Text. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 2004.

[27] P. Cimiano, A. Pivk, L. Schmidt-Thieme, and S. Staab. Learning Taxonomic Relations from Heterogeneous Evidence. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004), Workshop W18 on Ontology learning and population*, Valencia, Spain, pp. 25-30, 2004.

[28] P. Cimiano, L. Schmidt-Thieme, A. Pivk, and S. Staab. Learning Taxonomic Relations from Heterogeneous Evidence. In *Ontology Learning from Text: Methods, Applications and Evaluation*, P. Buitelaar, P. Cimiano, and B. Magnini (editors), IOS Press, pp. to apper, 2005.

[29] F. Ciravegna, A. Dingli, Y. Wilks, and D. Petrelli. Using Adaptive Information Extraction for Effective Human-Centred Document Annotation. In *Text Mining, Theoretical Aspects and Applications*, J. Franke, G. Nakhaeizadeh, and I. Renz (editors), Physica-Verlag, pp. 153-164, 2003.

[30] E.A. Codd. A Relational Model for Large Shared Databanks. *Communication of the ACM*, 13(6): 377-387, 1970.

[31] W.W. Cohen and W. Fan. Learning Page-Independent Heuristics for Extracting Data from Web Pages. *Computer Networks*, 31(11-16): 1641-1652, 1999.

[32] W.W. Cohen, M. Hurst, and L.S. Jensen. A Flexible Learning System for Wrapping Tables and Lists in HTML Documents. In *Proceedings of the 11th World Wide Web Conference*, Honolulu, Hawaii, pp. 232-241, 2002.

[33] W.W. Cohen. Learning and Discovering Structure in Web Pages. *IEEE Data Eng. Bull.*, 26(3): 3-10, 2003.

[34] W.W. Cohen, P. Ravikumar, and S. Fienberg. A Comparison of String Distance Metrics for Name-matching Tasks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2003), IIWeb Workshop*, 2003.

[35] R.A. Coll, J.H. Coll, and G. Thakur. Graphs and Tables: A Four-factor Experiment. *Communication of the ACM*, 37(4): 76-86, 1994.

[36] J. Cowie and W. Lehnert. Information Extraction. *Communication of the ACM*, 39(1): 80-101, 1996.

[37] V. Crescenzi and G. Mecca. Grammars have Exceptions. *Information Systems*, 23(8): 539-565, 1998.

[38] V. Crescenzi, G. Mecca, and P. Merialdo. RoadRunner: Towards Automatic Data Extraction from Large Websites. In *Proceedings of the 26th International Conference on Very Large Databases (VLDB'01)*, Rome, Italy, pp. 109-118, 2001.

[39] V. Crescenzi and G. Mecca. Automatic Information Extraction from Large Websites. *Journal of the ACM*, 51(5): 731-779, 2004.

[40] M. Dean, D. Connolly, F. Harmelen, J. Hendler, I. Horrocks, D.L. Mcguinness, P.F. Patel-Schneider, and L.A. Stein. OWL Ontology Language 1.0 Reference, http://www.w3.org/owl-ref/.

[41] S. Decker, M. Erdmann, D. Fensel, and R. Studer. Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information. In *Database Semantics: Semantic Issues in Multimedia Systems*, R. Meersman, S. Stevens, and Z. Tari (editors), Kluwer, pp. 351-369, 1999.

[42] A. Deitel, C. Faron, and R. Dieng. Learning Ontologies from RDF Annotations. In *Proceedings of the IJCAI 2001, Workshop on Ontology Learning*, Seattle, USA, 2001.

[43] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J.A. Tomlin, and J.Y. Zien. SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation. In *Proceedings of the 12th International Conference on WWW*, Budapest, Hungary, pp. 178-186, 2003.

[44] A. Doan, P. Domingos, and A. Levy. Learning Source Descriptions for Data Integration. In *Proceedings of the 3rd International Workshop on the Web and Databases*, Dallas, USA, pp. 81-86, 2000.

[45] S. Douglas, M. Hurst, and D. Quinn. Using Natural Language Processing for Identifying and Interpreting Tables in Texts. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pp. 535-546, 1995.

[46] S. Douglas and M. Hurst. Layout and Language: List and Tables in Technical Documents. In *Proceedings of ACL SIGPARSE Workshop on Punctuation in Computational Linguistics*, pp. 19-24, 1996.

[47] L. Eikvil, Information Extraction from World Wide Web: A Survey. *Report No. 945*, 1999.

[48] D. Embley, D. Hurst, D. Lopresti, and G. Nagy. Table Processing Paradigms: A Research Survey. *International Journal on Document Analysis and Recognition*, (submitted), 2004.

[49] D.W. Embley, Y.S. Jiang, and Y.K. Ng. Record-boundary Discovery in Web Documents. In *Proceedings of the ACM SIGMOD International Conference of Management of Data*, Philadelphia, PA, pp. 467-478, 1999.

[50] D.W. Embley, C. Tao, and S.W. Liddle. Automatically Extracting Ontologically Specified Data from HTML Tables with Unknown Structure. In *Proceedings of the 21st International Conference on Conceptual Modeling*, Tampere, Finland, pp. 322-337, 2002.

[51] W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y.K. Ng, and R.D. Smith. Conceptual-model-based Data Extraction from Multiple-record Web Pages. *Data & Knowledge Engineering*, 31(3): 227-251, 1999.

[52] D. Faure and T. Poibeau. First Experiments of Using Semantic Knowledge Learned by ASIUM for Information Extraction Task using INTEX. In *Proceedings of the 14th European Conference on Artificial Intelligence ECAI (ECAI 2000), Workshop on Ontology Learning*, Berlin, Germany, 2000.

[53] C. Fellbaum. *WordNet, an Electronic Lexical Database*. MIT Press, 1998.

[54] D. Fensel, J. Angele, S. Decker, M. Erdmann, H.-P. Schnurr, S. Staab, R. Studer, and A. Witt. On2broker: Semantic-based Access to Information Sources at the WWW. In *Proceedings of the World Conference on the WWW and Internet (WebNet'99)*, Honolulu, Hawaii, pp. 1366-1371, 1999.

[55] D. Fensel. *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Springer, 2001.

[56] J. Ferber. *Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence*. Addison-Wesley Pub Co., 1999.

[57] D. Freitag. Information Extraction from HTML: Application of a General Machine Learning Approach. In *Proceedings of the 15th International Conference of AAAI-98*, pp. 517-523, 1998.

[58] D. Freitag. Machine Learning for Information Extraction. *Machine Learning*, 39(2/3): 169-202, 2000.

[59] J. Frohn, R. Himmeröder, P. Kandzia, and C. Schlepphorst, How to write F-logic programs in FLORID. A tutorial for the database language F-Logic. Technical report, Version 1.0, Institut für Informatik der Universität Freiburg, 1996.

[60] M. Gams. A Uniform Internet-Communicative Agent. *Electronic Commerce Research*, 1: 69-84, 2001.

[61] M. Gams and A. Pivk. Agents for Semi-automatic Generation of Database Wrappers. In *Proceedings of the International Conference on Information and Knowledge Engineering*, Las Vegas, Nevada, pp. 265-270, 2002.

[62] A.V. Gelder, K.A. Ross, and J.S. Schlipf. The Well-founded Semantics for General Logic Programs. *Journal of the ACM*, 38(3): 620-650, 1991.

[63] A. Gomez-Perez and D. Manzano-Macho, OntoWeb Deliverable 1.5: A Survey of Ontology Learning Methods and Techniques. Universidad Politecnica de Madrid, 2003.

[64] A. Gomez-Perez, M. Fernandez-Lopez, and O. Corcho. *Ontological Engineering: with Examples from the Areas of Knowledge Management, E-commerce, and the Semantic Web*. Springer-Verlag, London, 2004.

[65] T.R. Gruber. A Translation Approach to Portable Ontology Specification. *Knowledge Acquisition*, 5(2): 199-220, 1993.

[66] R.H. Gutmann, A.G. Moukas, and P. Maes. Agents as Mediators in Electronic Commerce. *Electronic Markets*, 8(1): 22-27, 1998.

[67] U. Hahn and S. Schulz. Towards Very Large Terminological Knowledge Bases: A Case Study from Medicine. In *Proceedings of the Canadian Conference on AI 2000*, pp. 176-186, 2000.

[68] R. Hall. *Handbook of Tabular Presentation*. The Ronald Press Company, New York City, NY, 1943.

[69] J. Hammer, H. Garcia-Molina, S. Nestorov, R. Yerneni, M. Breunig, and V. Vassalos. Template-based Wrappers in the TSIMMIS System. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Tucson, Arizona, pp. 532-535, 1997.

[70] S. Handschuh, S. Staab, and F. Ciravegna. S-CREAM - Semi-automatic CREAtion of Metadata. In *Proceedings of EKAW 2002*, pp. 358-372, 2002.

[71] S. Handschuh and S. Staab. *Annotation in the Semantic Web*. IOS Press, 2003.

[72] M.A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France, 1992.

[73] J. Hendler. Making Sense out of Agents. *IEEE Intelligent Systems*: 32-37, 1999.

[74] J. Hendler. Agents and the Semantic Web. *IEEE Intelligent Systems*, 16(2): 30-37, 2001.

[75] I. Horrocks, D. Fensel, F. Harmelen, S. Decker, M. Erdmann, and M. Klein. OIL in a Nutshell. In *Proceedings of the Workshop on Application of Ontologies and PSMs, ECAI 2000*, Berlin, 2000.

[76] I. Horrocks and F. Harmelen. Reference Description of the DAML+OIL (March 2001) Ontology Markup Language, http://www.daml.org/2001/03/reference.html.

[77] C.-H. Hsu. Generating Finite-State Transducers for Semistructured Data Extraction from the Web. *Information Systems*, 23(8): 521-538, 1998.

[78] J. Hu, R. Kashi, D. Lopresti, G. Nagy, and G. Wilfong. Why table ground-truthing is hard. In *Proceedings of the 6th International Conference on Document Analysis and Recognition*, Seattle, Washington, pp. 129-133, 2001.

[79] J. Hu, R.S. Kashi, D. Lopresti, and G.T. Wilfong. Evaluating the Performance of Table Processing Algorithms. *International Journal on Document Analysis and Recognition*, 4(3): 140-153, 2002.

[80] M. Hurst and S. Douglas. Layout and Language: Preliminary Investigations in Recognizing the Structure of Tables. In *Proceedings of the 4th International Conference on Document Analysis and Recognition*, Ulm, Germany, pp. 1043-1047, 1997.

[81] M. Hurst. Layout and Language: Beyond Simple Text for Information Interaction - Modelling the Table. In *Proceedings of the 2nd International Conference on Multimodal Interfaces*, Hong Kong, 1999.

[82] M. Hurst. *The Interpretation of Tables in Texts*. PhD thesis, University of Edinburgh, 2000.

[83] M. Hurst and T. Nasukawa. Layout and Language: Integrating Spatial and Linguistic Knowledge for Layout Understanding Tasks. In *Proceedings of the 8th International Conference on Computational Linguistics*, Saarbrucken, Germany, 2000.

[84] M. Hurst. Layout and Language: An Efficient Algorithm for Detecting Text Blocks based on Spatial and Linguistic Evidence. In *Proceedings of the Document Recognition and Retrieval VII Conference*, pp. 56-67, 2001.

[85] M. Hurst. Layout and Language: Challenges for Table Understanding on the Web. In *Proceedings of the International Workshop on Web Document Analysis*, pp. 27-30, 2001.

[86] B. Jansen, A. Spink, J. Bateman, and T. Saracevic. Searchers, the subjects they search, and sufficiency: A study of a large sample of EXCITE searchers. In *Proceedings of the World Conference on the WWW, Internet and Intranet (WebNet-98)*, Orlando, Florida, AACE Press, pp. 472-477, 1998.

[87] N.R. Jennings and M.J. Wooldridge. Applying Agent Technology. *Applied Artificial Intelligence*, 9(4): 351-361, 1995.

[88] N.R. Jennings and M.J. Wooldridge, eds. *Agent Technology: Foundations, Applications, and Markets*. Springer Verlag, 1998.

[89] T. Joachims. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML'97)*, Morgan Kaufmann, pp. 143-151, 1997.

[90] J.U. Kietz, A. Maedche, and R.V. R. A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet. In *Proceedings of the12th International Conference on Knowledge Engineering and Knowledge Management, Workshop on Ontologies and Texts*, Juan-les-Pins, France, 2000.

[91] M. Kifer, G. Lausen, and J. Wu. Logical Foundations of Object-oriented and Frame-based Languages. *Journal of the ACM*, 42: 741-843, 1995.

[92] M. Kljajić. *Teorija sistemov*. Moderna organizacija, Kranj, 1994.

[93] M. Klusch. Information Agent Technology for the Internet: A Survey. *Data & Knowledge Engineering, Special Issue on Intelligent Information Integration*, 36(3): 337 - 372, 2001.

[94] C. Knoblock, K. Lerman, S. Minton, and I. Muslea. Accurately and Reliably Extracting Data from the Web: A Machine Learning Approach. In *Intelligent Exploration of the Web*, P. Szczepaniak, J. Segovia, J. Kacprzyk, and L. Zadeh (editors), Springer, pp. 275-287, 2002.

[95] N. Kushmerick, D.S. Weld, and R. Doorenbas. Wrapper Induction for Information Extraction. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 729-735, 1997.

[96] N. Kushmerick. Wrapper Induction: Efficiency and Expressiveness. *Artificial Intelligence Journal*, 118(1-2): 32-43, 2000.

[97] N. Kushmerick and B. Thomas. Adaptive Information Extraction: Core Technologies for Information Agents. In *Intelligent Information Agents R&D in Europe: An Agentlink Perspective, LNCS 2586*, Springer, pp. 79-103, 2003.

[98] A.H.F. Laender, B.A. Ribiero-Neto, and A.S.D. Silva. DEByE - Data Extraction by Examples. *Data & Knowledge Engineering*, 40(2): 121-154, 2002.

[99] K. Lerman, S. Minton, and C. Knoblock. Wrapper Maintenance: A Machine Learning Approach. *Journal of Artificial Intelligence Research*, 18: 149-181, 2003.

[100] A.Y. Levy and D.S. Weld. Intelligent Internet Systems. *Artificial Intelligence*, 118: 1-14, 2000.

[101] L. Liu, C. Pu, and W. Han. An XML-enabled Wrapper Construction System for Web Information Sources. In *Proceedings of the 16th International Conference on Data Engineering*, San Diego, California, pp. 611-621, 2000.

[102] D. Lopresti and G. Nagy. A Tabular Survey of Automated Table Processing. In *Graphics Recognition - Recent Advances*, A. Chhabra and D. Dori (editors), Springer-Verlag, pp. 93-120, 2000.

[103] A. Maedche and S.Staab. Discovering Conceptual Relations from Text. In *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI'2000)*, Berlin, Germany, IOS Press, pp. 21-25, 2000.

[104] A. Maedche. *Ontology Learning for the Semantic Web*. Kluwer Academic Publishers, 2002.

[105] A. Maier, H.-P. Schnurr, and Y. Sure. Ontology-based Information Integration in the Automotive Industry. *Lecture Notes in Computer Science*, 2870: 897-912, 2003.

[106] A. Mccallum, D. Freitag, and F. Pereira. Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proceedings of the ICML 2000*, pp. 591-598, 2000.

[107] M. Missikoff, R. Navigli, and P. Velardi. The Usable Ontology: An Environment for Building and Assessing a Domain Ontology. In *Proceedings of the 1st International Semantic Web Conference, LNCS 2342*, Springer-Verlag, pp. 39-53, 2002.

[108] E. Morin. Automatic Acquisition of Semantic Relations between Terms from Technical Corpora. In *Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE'99)*, Vienna, Austria, 1999.

[109] I. Muslea, S. Minton, and C. Knoblock. Hierarchical Wrapper Induction for Semistructured Information Sources. *Agents and Multi-Agent Systems*, 4(1-2): 93-114, 2001.

[110] I. Muslea, S. Minton, and C. Knoblock. Active + Semi-Supervised Learning = Robust Multi-View Learning. In *Proceedings of the 19th International Conference on Machine Learning ICML-2002*, pp. 435-442, 2002.

[111] I. Muslea, S. Minton, and C. Knoblock. Active Learning with Strong and Weak Views: A Case Study on Wrapper Induction. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2003)*, pp. 415-420, 2003.

[112] R. Navigli, P. Velardi, and A. Gangemi. Ontology Learning and Its Application to Automated Terminology Translation. *IEEE Intelligent Systems*, 18(1), 2003.

[113] R. Neches, R.E. Fikes, T. Finnin, T.R. Gruber, T. Senator, and W.R. Swartout. Enabling Technology for Knowledge Sharing. *AI Magazine*, 12(3): 36-56, 1991.

[114] F. Neri and L. Saitta. Machine Learning for Information Extraction. In *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, LNAI*, M.T. Pazienza (editor), Springer-Verlag, pp. 171-191, 1997.

[115] H.T. Ng, C.Y. Kim, and J.L.T. Koo. Learning to Recognize Tables in Free Text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 443-450, 1999.

[116] H.S. Nwana. Software Agents: An Overview. *Knowledge Engineering Review*, 11(3): 1-40, 1996.

[117] M.T. Pazienza, ed. *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*. Lecture Notes in Artificial Intelligence, Springer-Verlag, 1997.

[118] D. Pinto, W. Croft, M. Branstein, R. Coleman, M. King, W. Li, and X. Wei. Quasm: A System for Question Answering using Semi-structured Data. In *Proceedings of the Joint Conference on Digital Libraries (JCDL)*, pp. 46-55, 2002.

[119] D. Pinto, A. Mccallum, X. Wei, and W.B. Croft. Table Extraction using Conditional Random Fields. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, ACM Press, pp. 235-242, 2003.

[120] M. Pivec and V. Rajkovič. Obvladovanje znanja z metodami umetne inteligence. *Organizacija*, 32(8/9): 449-453, 1999.

[121] A. Pivk and M. Gams. E-Commerce Intelligent Agents. In *Proceedings of the 3rd International Conference on Telecommunications and Electronic Commerce (ICTEC3)*, Dallas, Texas, pp. 418-429, 2000.

[122] A. Pivk and M. Gams. Intelligent Agents in E-commerce. *Electrotechnical Review*, 67(5): 251-260, 2000.

[123] A. Pivk. *Domensko odvisni agent za avtomatsko generiranje ovojnic*. Magistrsko delo, Univerza v Ljubljani, 2002.

[124] A. Pivk and M. Gams. Domain-dependent Information Gathering Agent. *Expert Systems with Applications*, 23: 207-218, 2002.

[125] A. Pivk. OWL - spletni jezik za opis ontologij. In *Proceedings A of the 6th International Multi-Conference on Information Society (IS'03)*, Ljubljana, Slovenia, pp. 113-116, 2003.

[126] A. Pivk and M. Gams. A Semi-universal E-commerce Agent: Domain-dependant Information Gathering. In *Enterprise Information Systems IV*, Kluwer Academic Publishers, pp. 260-267, 2003.

[127] A. Pivk, P. Cimiano, and Y. Sure. From Tables to Frames. In *Proceedings of the 3rd International Semantic Web Conference (ISWC2004), LNCS 3298*, Hiroshima, Japan, Springer-Verlag, pp. 166-181, 2004.

[128] A. Pivk, P. Cimiano, and Y. Sure. From Tables to Frames. *Web Semantics Journal*: to apper, 2005.

[129] A. Pivk, P. Cimiano, Y. Sure, M. Gams, V. Rajkovič, and R. Studer. Transforming Arbitrary Tables into F-Logic Frames with TARTAR. *Data & Knowledge Engineering*: submitted, 2005.

[130] P. Pyreddy and W.B. Croft. TINTIN: A System for Retrieval in Text Tables. In *Proceedings of the Second ACM International Conference on Digital Libraries*, Philadelphia, Pennsylvania, ACM Press, pp. 193-200, 1997.

[131] V. Rajkovič and B. Zupan. Knowledge Management as a Challenge for Public Administration Reengineering. *Informatics and Management: Selected Topics*: 129-138, 2004.

[132] S. Russel and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Inc., 1995.

[133] A. Sahuguet and F. Azavant. Building Intelligent Web Applications using Lightweight Wrappers. *Data & Knowledge Engineering, Special issue on heterogeneous information resources need semantic access*, 36(3): 283-316, 2001.

[134] M. Sanderson and B. Croft. Deriving Concept Hierarchies from Text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, California, ACM Press, pp. 206-213, 1999.

[135] Y. Shoham. An Overview of Agent-oriented Programming. In *Software Agents*, J.M. Bradshaw (editor), AAAI Press, Menlo Park, California, pp. 271-290, 1997.

[136] Y. Shoham. What we talk about when we talk about software agents. *IEEE Intelligent Systems*, 14(2): 28-31, 1999.

[137] S. Soderland. Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*, 34(1-3): 233-272, 1999.

[138] N. Stojanovic, R. Studer, and L. Stojanovic. An Approach for Step-By-Step Query Refinement in the Ontology-Based Information Retrieval. In *International Conference on Web Intelligence (WI 2004)*, Beijing, China, IEEE Computer Society, pp. 36-43, 2004.

[139] R. Studer, V.R. Benjamins, and D. Fensel. Knowledge engineering: Principles and methods. *IEEE Transactions on Data and Knowledge Engineering*, 25(1-2): 161-197, 1998.

[140] A. Tengli, Y. Yang, and N.L. Ma. Learning Table Extraction from Examples. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, Geneva, Switzerland, 2004.

[141] Y.A. Tijerino, D.W. Embley, D.W. Lonsdale, and G. Nagy. Ontology Generation From Tables. In *Proceedings of 4th International Conference on Web Information Systems Engineering (WISE'03)*, Rome, Italy, pp. 242-249, 2003.

[142] Y.A. Tijerino, D.W. Embley, D.W. Lonsdaleand, Y. Ding, and G. Nagy. Towards Ontology Generation from Tables. *World Wide Web Journal*: to appear, 2004.

[143] J.D. Ullman. *Principles of Database and Knowledge-based Systems*. Vol. 2, Computer Science Press, New York, 1989.

[144] M. Vargas-Vera, E. Motta, J. Domingue, M. Lanzoni, A. Stutt, and F. Ciravegna. MnM: Ontology Driven Semi-automatic and Automatic Support for Semantic Markup. In *Proceedings of EKAW 2002, LNCS 2473*, pp. 379-391, 2002.

[145] P. Velardi, P. Fabriani, and M. Missikoff. Using Text Processing Techniques to Automatically enrich a Domain Ontology. In *Proceedings of the ACM International Conference on Formal Ontology in Information Systems*, Ogunquit, Maine, ACM Press, pp. 270 - 284, 2001.

[146] H.L. Wang, S.H. Wu, I.C. Wang, C.L. Sung, W.L. Hsu, and W.K. Shih. Semantic Search on Internet Tabular Information Extraction for Answering Queries. In *Proceedings of the 9th International Conference on Information and Knowledge Management*, Washington DC, pp. 243-249, 2000.

[147] X. Wang. *Tabular Abstraction, Editing and Formatting*. PhD thesis, University of Waterloo, 1996.

[148] Y. Wang, R. Haralick, and I. Phillips. Zone Content Classification and its Performance Evaluation. In *Proceedings of the 6th International Conference on Document Analysis and Recognition (ICDAR01)*, Seattle, Washington, pp. 540-544, 2001.

[149] Y. Wang and J. Hu. Detecting Tables in HTML Documents. In *LNCS*, Vol. 2423, Springer-Verlag, pp. 249-260, 2002.

[150] Y. Wang and J. Hu. A Machine Learning Based Approach for Table Detection on the Web. In *Proceedings of the 11th International Conference on the World Wide Web*, Honolulu, Hawaii, ACM Press, pp. 242-250, 2002.

[151] Y. Wang, I.T. Phillips, R.M. Robert, and M. Haralick. Table Structure Understanding and its Performance Evaluation. *Pattern Recognition Journal*, 37(7): 1479-1497, 2004.

[152] M.J. Wooldridge. *Introduction to MultiAgent Systems*. John Wiley & Sons, 2002.

[153] M. Yoshida, K. Torisawa, and J. Tsujii. Extracting Ontologies from World Wide Web via HTML Tables. In *Proceedings of the Pacific Association for Computational Linguistics (PACLING 2001)*, pp. 332-341, 2001.

[154] M. Yoshida, K. Torisawa, and J. Tsujii. A Method to Integrate Tables of the World Wide Web. In *Proceedings of the International Workshop on Web Document Analysis*, pp. 31-34, 2001.

[155] M. Yoshida. *A Method for Information Extraction from Tables and Lists*. PhD thesis, University of Tokyo, 2003.

[156] M. Yoshida. Extracting Attributes and Their Values from Web Pages. In *Web Document Analysis: Challenges and Opportunities*, A. Antonacopoulos and J. Hu (editors), World Scientific, 2003.

[157] R. Zanibbi, D. Blostein, and J.R. Cordy. A Survey of Table Recognition: Models, Observations, Transformations, and Inferences. *International Journal on Document Analysis and Recognition*, 7(1): 1433-2833, 2004.