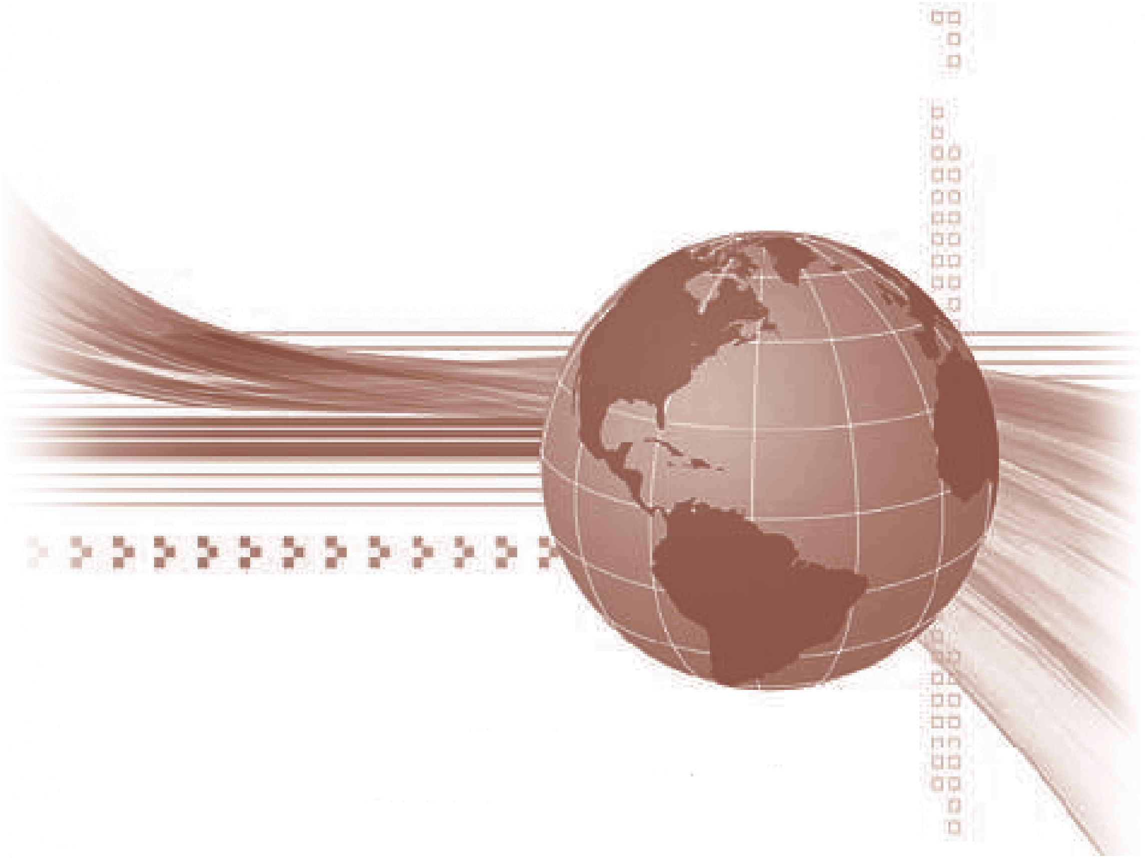
STUDIA UNIVERSITATIS
BABEŞ-BOLYAI



INFORMATICA

1/2010

Anul LV **2010**

STUDIA
UNIVERSITATIS BABEŞ-BOLYAI

INFORMATICA

1

Redacţia: M. Kogălniceanu 1 • 400084 Cluj-Napoca • Tel: 0264.405300

SUMAR – CONTENTS – SOMMAIRE

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LV**, Number 1, 2010

PROFESSOR DUMITRU DUMITRESCU AT HIS SIXTIES

BAZIL PÂRV

Profesor Dumitru Dumitrescu was born on August 22, 1949, in Sineşti, Vâlcea county. In 1972 he graduated the Theoretical Physics programme at Physics Faculty, Babes-Bolyai University, and in 1979 he graduated Mathematics programme at Mathematics-Mechanics Faculty.

Since 1972 he stepped through all didactic positions in our Faculty: Assistant Professor between 1972-1990, Lecturer between 1990-1992, Associate Professor between 1992-1995 and Full Professor since 1995. He obtained his PhD in 1990, under the supervision of Prof. Dr. D.D. Stancu, with the thesis *Iterative Methods for Solving Fuzzy Classification Problems and Fuzzy Relational Equations*. Since 2001 he has been a PhD supervisor himself. Between 1998-2000 he was Researcher and Associate Professor at the Department of Information Technology, University of Pisa, Italy.

As a professor at our Faculty, Dumitru is known for his high level of scientific knowledge, having didactic activities (lectures, seminars, laboratories) in 16 disciplines. He introduced the first courses of Artificial Intelligence in our Computer Science curricula and managed the Artificial Intelligence master programme, now called Intelligent Systems and taught in English. During his Visiting Researcher and Associate Professor position at Pisa University, he taught there a course on Artificial Intelligence.

The main research interests of Professor Dumitrescu are: *Artificial Intelligence* (general aspects), *Neural Computing* (neural networks, connectionist models), *Evolutionary Computing* (population models, multimodal problems), *Data Analysis*, *Pattern Classification and Recognition*, *Intelligent Control*, *Non-standard Logics* and their application to Artificial Inteligence, *Information Theory and Ergodic Systems*, *Mathematical Theory of Fuzzy Systems*, and *Game Theory.*

A synthesis of his scientific results shows the following:

- 14 books and university manuals, including 4 books and book chapters published in international editions;
- 45 papers in ISI journals;
- 74 papers in ISI proceedings;

Received by the editors: December 10, 2009.

- 208 papers in other journals and conference proceedings;
- 30 participations in international conferences;
- 52 participations in national conferences;
- 40 participations in program or scientific committees of international conferences.

The main research contributions of Professor Dumitrescu are:

- an ergodic theory of fuzzy systems;
- a fuzzy information theory;
- a new computational paradigm, *far-from-equilibrium computing*, inspired from far-from-equilibrium self-organization of complex systems;
- a new paradigm in evolutive computing: *genetic chromodynamics*, GC; GC for data mining (data classification, analysis, and interpretation), macroevolution models, industrial applications of GC;
- form classification and recognition and data analysis using evolutive methods;
- fuzzy classification and data mining: mathematical foundation; one-level, multi-level and hierarchical fuzzy classification, applications to data analysis and data mining, supervised classificators and their applications;
- fuzzy logic and mathematical aspects of fuzzy sets related to Artificial Intelligence problems;
- training of neural networks using ambiguous or incomplete data.

As a confirmation of his huge research activity, his papers and ideas are frequently cited in the scientific literature, papers and PhD theses. The new paradigms introduced in the national and international AI research are of a great importance. At national level, he published in Romanian the first books on Neural Networks, Evolutionary Computing, and Fuzzy Models in Data Analysis. His books published in international editions are used by students and researchers, being considered as reference works in the Artificial Intelligence.

Professor Dumitrescu is the Director of the Center for Studies on Complexity and the Head of CIRG - Computational Intelligence Research Group. He won many research grants as principal investigator, and participated in organizing committees of international conferences home and abroad. Also, he is member of many scientific and professional organizations and reviewer of many scientific journals of high quality.

As a recognition of his excellent research activity, Romanian Academy offered Professor Dumitrescu in 2006 the Grigore Moisil Prize. Also, he was rewarded by our university in several years for books published in international editions and for scientific excellence.

We wish Professor Dumitrescu a happy and healthy long life, full of achievements, for the benefit of his students, collaborators, friends and family.

Scientific activity

Books.

(1) D. Dumitrescu, B. Lazzerini, L. Jain, *Fuzzy Sets and Their Applications to Clustering and Training*, CRC Press, Boca Raton, New York, 2000, 665 p.
(2) D. Dumitrescu, B. Lazzerini, L. Jain, A. Dumitrescu, *Evolutionary Computation*, CRC Press, Boca Raton, New York, 2000, 416 p,
(3) B. Iantovics, C. Chira, D. Dumitrescu, *Principiile agenţilor inteligenţi* (*Intelligent Agents Principles*), Casa Cărţii de Ştiinţă, Cluj-Napoca, 2007, ISBN: 978-973-133-035-8.
(4) D. Dumitrescu, *Principiile inteligenţei artificiale* (*Artificial Intelligence Principles*), Editura Albastră, Cluj-Napoca, 1999, 380 p.
(5) D. Dumitrescu, *Principiile matematice ale teoriei clasificării* (*Mathematical Foundations of the Classification Theory*), Editura Academiei, Bucureşti, 1999, 470 p.
(6) D. Dumitrescu, H. Costin, *Reţele neuronale* (*Neural Networks*), Teora, Bucureşti, 1996, 460 p.
(7) D. Dumitrescu, *Inteligenţă artificială* (*Artificial Intelligence*), Babeş-Bolyai University Press, Cluj-Napoca, 1995, 253 p.
(8) D. Dumitrescu, *Teoria clasificării* (*Classification Theory*), Babeş-Bolyai University Press, Cluj-Napoca, 1991, 192 p.
(9) D. Dumitrescu, *Modele conexioniste în inteligenţa artificială* (*Conexionist Models in Artificial Intelligence*), Babeş-Bolyai University Press, Cluj-Napoca, 1995, 335 p.

Book chapters and booklets.

(1) R. Stoean, M. Preuss, C. Stoean, E. El-Darzi, D. Dumitrescu, *An Evolutionary Approximation for the Coefficients of Decision Functions within a Support Vector Machine Learning Strategy*, Book chapter in *Studies in Computational Intelligence*, Springer, Aboul Ella Hassanien and Ajith Abraham (Eds.), Vol. 1, pp. 83-114, ISSN 1860-949X.
(2) D. Dumitrescu, C. Groşan, M. Oltean, *Evolving Continuous Pareto Regions*, Contributed chapter in *Evolutionary Computation Based Multi-Criteria Optimization: Theoretical Advances and Applications*, A. Abraham, L. Jain and R. Goldberg (Eds.), Springer-Verlag, London, 2005.
(3) D. Dumitrescu, *Fuzzy Hierarchical Classification Methods in Analytical Chemistry*, Contributed chapter in *Fuzzy Logic in Chemistry*, D.H. Rouvray (Ed), Academic Press, New York, 1997, pp. 321-356.
(4) C. Chira, C.-M. Pintea, D. Dumitrescu, *A Multi-Agent Stigmergic Model for Complex Optimization Problems*, Contributed chapter in *From Natural Language to Soft Computing, New Paradigms in Artificial Intelligence*, L.A.

Zadeh, D. Tufiş, F.G. Filip, I. Dziţac (Eds.), Publishing House of Romanian Academy, 2008, pp. 51-62.

(5) D. Dumitrescu, *Genetic algorithms*, Contributed chapter in *Nature Inspired Computational Models*, T. Petrila, D. Trif, D. Dumitrescu, I. Mihoc (Eds.), Digital Data Publ., Cluj-Napoca, 2003, pp. 200-338.

(6) D. Dumitrescu, *Fuzzy Sets*, Technical University of Cluj-Napoca, I.P.C.-N. Press, 1993.

(7) D. Dumitrescu, *Fuzzy Trainable Classifiers. Fuzzy Learning Machines*, Technical University of Cluj-Napoca, I.P.C.N. Press, 1993.

Journal and Conference Papers.

(1) D. Dumitrescu, R.I. Lung, T.D. Mihoc, *Evolutionary Equilibria Detection in Non-cooperative Games*, Applications of Evolutionary Computing, EvoWorkshops 2009, Tubingen, Germany, April 15-17, 2009.

(2) D. Dumitrescu, R.I. Lung, T.D. Mihoc, *Evolutionary Equilibria Detection in Non-cooperative Games*, in Applications of Evolutionary Computing, LNCS 5484, Springer, 2009, pp. 253-262, ISBN 978-3-642-01128-3.

(3) D. Dumitrescu, R.I. Lung, T.D. Mihoc, *Generative Relations for Evolutionary Equilibria Detection*, Proc. of the 11th Annual conference on Genetic and evolutionary computation GECCO '09, New York, NY, USA, 8-12 July 2009, pp. 1507-1512. ACM. ISBN 978-1-60558-325-9.

(4) D. Dumitrescu, R.I. Lung, T.D. Mihoc, *Equilibria Detection In Electricity Market Games*, Proc. of the International Conference on Knowledge Engineering, Principles and Techniques, KEPT2009, Cluj-Napoca (Romania), July 2-4, 2009, pp. 111-114.

(5) D. Dumitrescu, R.I. Lung, T.D. Mihoc, *Approximating and Combining Equilibria in Non cooperative Games*, Proc. of the 11th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing Timişoara SYNASC 2009, Romania, 2009.

(6) R.I. Lung, D. Dumitrescu, *Evolutionary swarm cooperative optimization in dynamic environments*, Journal of Natural Computing, 10.1007/ s11047-009-9129-9, 2009.

(7) R.I. Lung, D. Dumitrescu, *Evolutionary Multimodal Optimization for Nash Equilibria Detection*, 3rd International Workshop on Nature Inspired Cooperative Strategies for Optimization, NICSO, Springer Studies in Computational Intelligence, 2009.

(8) D. Iclănzan, D. Dumitrescu, B. Hirsbrunner, *Correlation guided model building*, Proc. of the 11th Annual conference on Genetic and evolutionary computation GECCO '09, New York, NY, USA, 8-12 July 2009, pp. 421-428. ACM. ISBN 978-1-60558-325-9.

(9) L. Szilagyi, D. Iclănzan, S.M. Szilgyi, D. Dumitrescu, B. Hirsbrunner, *A generalized c-means clustering model using optimized via evolutionary computation*, In IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'09, Jeju Island, Korea), 2009, pp. 451-455.

(10) C. Chira, C.-M. Pintea, D. Dumitrescu, *Sensitive Stigmergic Agent Systems - A Hybrid Approach to Combinatorial Optimization*, In Innovations in Hybrid Intelligent Systems, Advances in Soft Computing, Springer, Vol. 44, 2008, pp. 33-39.

(11) P.C. Pop, C-M. Pintea, D. Dumitrescu, *An Ant Colony algorithm for solving the dynamic generalized Vehicle Routing Problem*, An. St. Univ. Ovidius Constanta, 2009 (ISSN: 1224-1784).

(12) C. Chira, A. Gog, D. Dumitrescu, *Asynchronous Collaborative Search using Adaptive Coevolving Subpopulations*, ECOMASS Workshop, Genetic and Evolutionary Computation Conference GECCO'09, July 8-12, 2009, Montreal, Canada, F. Rothlauf (Ed), GECCO Companion, pp. 2575-2582, ACM.

(13) C. Chira, C.-M. Pintea, G. C. Crişan, D. Dumitrescu, *Solving the Linear Ordering Problem using Ant Models*, Genetic and Evolutionary Computation Conference, July 8-12, 2009, Montreal, Canada, F. Rothlauf (Ed), pp. 1803-1804, ACM.

(14) D. Iclănzan, B. Hirsbrunner, M. Courant, D. Dumitrescu, *Cooperation in the context of sustainable search*, In IEEE Congress on Evolutionary Computation (IEEE CEC 2009), Trondheim, Norway, 18-21 May 2009, pp. 1904-1911.

(15) C.-M. Pintea, G.C. Crişan, C. Chira, D. Dumitrescu, *A Hybrid Ant-Based Approach to the Economic Triangulation Problem for Input-Output Tables*, Proc. of the 4th International Workshop on Hybrid Artificial Intelligence Systems (HAIS 2009), Real-world HAIS and Data Uncertainty, Salamanca, Spain, LNCS 5572, Springer, 2009, pp. 376-383.

(16) P.C. Pop, C.-M. Pintea, I. Zelina, D. Dumitrescu, *Solving the Generalized Vehicle Routing Problem with an ACS-based Algorithm*, BICS 2008, American Institute of Physics (AIP) Springer, 2009, vol. 1117, pp. 157-162.

(17) C. Chira, C.-M. Pintea, D. Dumitrescu, *Multi-Population Agent Search: Stigmergy and Heterogeneity*, SYNASC 08, IEEE Computer Society, 2009, pp. 526-531.

(18) A. Gog, C. Chira, D. Dumitrescu, *Distributed Asynchronous Collaborative Search*, Studia Univ. Babeş-Bolyai, Informatica, Special Issue, 2009, pp. 99-102.

(19) Chira, C.-M. Pintea, D. Dumitrescu, *A Step-Back Sensitive Ant Model for Solving Complex Problems*, Studia Univ. Babeş-Bolyai, Informatica, Special Issue, 2009, pp. 103-106.

(20) R. Stoean, M. Preuss, C. Stoean, E. El-Darzi, D. Dumitrescu, *An Evolutionary Resemblant to Support Vector Machines for Classification and Regression*, Journal of the Operational Research Society, Vol. 60, Issue 8 (August 2009), Special Issue: Data Mining and Operational Research: Techniques and Applications, K-M Osei-Bryson, V.J. Rayward-Smith (Guest Eds.), 2009, pp. 1116-1122, ISSN 0160-5682.

(21) C.M. Pintea, C. Chira, D. Dumitrescu, *Results of Ant-Based Models for Solving the Linear Ordering Problem*, Proc. of the 11th International Symposium

on Symbolic and Numeric Algorithms for Scientific Computing SYNASC 2009 Timişoara, Romania, 2009, IEEE Computer Society Press.

(22) C. Chira, C-M. Pintea, D. Dumitrescu, *An Agent-Based Approach to Combinatorial Optimization*, Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844, Vol. III (2008), Suppl. Issue, pp. 212-217.

(23) C-M. Pintea, C. Chira, D.Dumitrescu, *Sensitive Ants: Inducing Diversity in the Colony Nature Inspired Cooperative Strategies for Optimization*, Series Studies in Computational Intelligence, Vol. 236 (N. Krasnogor, B. Melin-Batista, J.A. Moreno-Prez, J.M. Moreno-Vega, D. Pelta Eds.), Springer, 2009 (ISBN: 978-3-642-03210-3).

(24) C. Chira, A. Gog, D. Zaharie, D. Dumitrescu, *Complex Dynamics in a Collaborative Evolutionary Search Model*, Creative Mathematics and Informatics, vol. 17, nr. 3, 2009.

(25) A. Gog, C. Chira. D. Dumitrescu, *Asynchronous Evolutionary Search: Multi-Population Collaboration and Complex Dynamics*, Proc. of IEEE Congress on Evolutionary Computation (CEC 2009), Trondheim, Norway, 2009, pp. 240-246.

(26) D. Iclănzan, D. Dumitrescu, *How can artificial neural networks help making the intractable search spaces tractable.* In 2008 IEEE World Congress on Computational Intelligence (WCCI 2008), Hong-Kong, 01-06 June 2008, pp. 4016-4023. ISBN 978-1-4244-1823-7.

(27) D. Iclănzan, D. Dumitrescu, *Towards memoryless model building.* In GECCO '08: Proc. of the 2008 GECCO conference companion on Genetic and evolutionary computation, Atlanta, GA, USA, 2008, pp. 2147-2152. ACM.

(28) D. Iclănzan, D. Dumitrescu, *Large-scale optimization of non-separable building-block problems.* In PPSN 2008: 10th International Conference on Parallel Problem Solving From Nature, Dortmund, Germany, 13-17 September 2008, pp. 899-908.

(29) D. Iclănzan, D. Dumitrescu, *Going for the big fish: Discovering and combining large neutral and massively multimodal building-blocks with model based macro-mutation.* In GECCO '08: Proc. of the 10th annual conference on Genetic and evolutionary computation, Atlanta, GA, USA, 2008, pp. 423-430. ACM.

(30) C. Chira, D. Dumitrescu, C-M. Pintea, *Heterogeneous Sensitive Ant Model for Combinatorial Optimization*, Genetic and Evolutionary Computation Conference GECCO'08, July 12-16, 2008, Atlanta, Georgia, USA, pp. 163-164. ACM.

(31) C. Chira, A. Gog, D. Dumitrescu, *Exploring Population Geometry and Multi-Agent Systems: A New Approach to Developing Evolutionary Techniques*, Genetic and Evolutionary Computation Conference GECCO'08, July 12-16, 2008, Atlanta, Georgia, USA, pp. 1953-1959. ACM.

(32) R. I. Lung, C. Chira, D. Dumitrescu, *An Agent-Based Collaborative Evolutionary Model for Multimodal Optimization*, Genetic and Evolutionary Computation Conference GECCO'08, July 12-16, 2008, Atlanta, Georgia, USA, 2008, pp. 1969-1975. ACM.

(33) L. Szilgyi, D. Iclănzan, S.M. Szilagyi, D. Dumitrescu, *Gecim: A novel generalized approach to c-means clustering.* In Jos Ruiz-Shulcloper, Walter G. Kropatsch (eds), CIARP, LNCS 5197, Springer, 2008, pp. 235-242.

(34) R. Gorunescu, P.H. Millard, D. Dumitrescu, *Evolutionary Placement Decisions of a Multidisciplinary Panel using Genetic Chromodynamics*, Journal of Enterprise Information Management, Vol. 21, No. 1, 2008, pp. 93-104, ISSN 1741-0398.

(35) C. Stoean, M. Preuss, R. Stoean, D. Dumitrescu, *EA-Powered Basin Number Estimation by Means of Preservation and Exploration*, Parallel Problem Solving from Nature - PPSN X, LNCS 5199, Springer, 2008, pp. 569-578, ISBN 978-3-540-87699-1.

(36) R. Stoean, C. Stoean, D. Dumitrescu, *Investigating Landscape Topology for Subpopulation Differentiation in Multimodal Evolutionary Algorithms. Study on Crowding Genetic Chromodynamics*, Postproc. of 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing - SYNASC 2008, IEEE Press, pp. 551-554.

(37) R. Stoean, C. Stoean, D. Dumitrescu, *Shifting from Radius-Centered Separation to Local Landscape Topology-Based Partition into Subpopulations within Crowding Genetic Chromodynamics*, 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2008.

(38) C. Chira, C.-M. Pintea, D. Dumitrescu, *Stigmergy and Sensitivity in Heterogeneous Agent-Based Models*, Workshop on Natural Computing and Applications, Proc. of the 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2008), September 26-29, Timişoara, Romania, 2008, pp. 13-17.

(39) C. Chira, A. Gog, D. Dumitrescu, *Distribution, collaboration and coevolution in asynchronous search*, Proc. of the International Symposium on Distributed Computing and Artificial Intelligence (DCAI 2008), Salamanca, Spain, Advances in Soft Computing, Vol. 50, 2009, pp. 596-604.

(40) A. Gog, C. Chira, D. Zaharie, D. Dumitrescu, *Analysis of a Distributed Collaborative Evolutionary Algorithm*, Workshop on Natural Computing and Applications, Proc. of the 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2008), Timişoara, Romania, September 26-29, 2008, pp. 25-32.

(41) A. Gog, C. Chira, D. Dumitrescu, *Hybrid Multi-Population Collaborative Asynchronous Search.* Proc. of the 3rd International Workshop on Hybrid Artificial Intelligence Systems (HAIS 2008), Burgos, Spain, September 24-26, Lecture Notes in Artificial Intelligence 5271, Springer, 2008, pp. 148-155.

(42) A. Gog, C. Chira, D. Dumitrescu, D. Zaharie, *Analysis of Some Mating and Collaboration Strategies in Evolutionary Algorithms*, Proc. of SYNASC 2008, IEEE Computer Publ.

(43) B. Reiz, L. Csato, D. Dumitrescu, *Prufer Number Encoding for Genetic Bayesian Network Structure Learning Algorithm.* Proc. of SYNASC 2008, IEEE Computer Publ., pp. 239-242.

(44) L. Dioşan, D. Dumitrescu, *Evolutionary coalition formation in full connected and scale free networks*, Intl J. of Computers, Communications & Control, 3, 2008, pp. 259-264.

(45) L. Dioşan, D. Dumitrescu, *Context-based Networks - Scale-free or not?*, Proc. of the Symposium Colocviul Academic Clujean de Informatică, 2008.

(46) C-M. Pintea, D. Dumitrescu, P.C. Pop, *Combining Heuristics and Modifying Local Information to Guide Ant-based Search*, Carpathian J.Math., 24(1), pp. 94-103, 2008 (ISSN: 1584-2851).

(47) C.-M. Pintea, P.C. Pop, C. Chira, D. Dumitrescu, *A Hybrid Ant-Based System for Gate Assignment Problem.* Proc. of the 3rd International Workshop on Hybrid Artificial Intelligence Systems (HAIS 2008), Burgos, Spain, Lecture Notes in Artificial Intelligence 5271, Springer, 2008, pp. 273-280.

(48) C.-M. Pintea, C. Chira, D. Dumitrescu, P.C. Pop *A Sensitive Metaheuristic for Solving a Large Optimization Problem*, SOFSEM 2008: Theory and Practice of Computer Science, LNCS 4910, Springer, V. Geffert, J. Karhumaki, A. Bertoni, B. Preneel, P. Navrat, M. Bielikova (Eds), pp. 551-559, 2008.

(49) A. Gog, D. Dumitrescu, *Evolving Network Topologies for Cellular Automata.* Studia Univ. Babeş-Bolyai, Informatica, Vol. LIII, No. 1, 2008, pp. 45-52.

(50) C. Chira, D. Dumitrescu, C-M. Pintea, *Sensitive Ant Model for Combinatorial Optimization*, 14th International Conference in Soft Computing MENDEL 2008, June 18-20, M. Radomil (Ed), Brno University of Technology, 2008.

(51) D. Dumitrescu, K. Simon, E. Vig, *Genetic chromodynamics. Data mining and training applications.* Studia Univ. Babeş-Bolyai, Informatica, Special Issue (2007) pp. 145-152.

(52) D. Dumitrescu, C. Stoean, R. Stoean, *Genetic Chromodynamics for the Job Shop Scheduling Problem*, Studia Univ. Babeş-Bolyai, Informatica, Special Issue, pp. 153-160, 2007.

(53) D. Iclănzan, D. Dumitrescu, *Overcoming hierarchical difficulty by hill-climbing the building block structure.* In Dirk Thierens et al., editor, GECCO '07: Proc. of the 9th annual conference on Genetic and Evolutionary Computation, London, 7-11 July 2007, volume 2, pp. 1256-1263. ACM Press. ISBN 978-1-59593-697-4.

(54) D. Iclănzan, D. Dumitrescu. *Exact model building in Hierarchical Complex Systems.* Studia Univ. Babeş-Bolyai, Informatica, Special Issue, 2007, pp. 161-168.

(55) L. Dioşan, A. Fanea, D. Dumitrescu, *Genetic algorithms based on Ising machine.* The International Journal of Information Technology and Intelligent Computing Vol. 1, No. 3, 2007, pp. 585-594.

(56) C-M.Pintea, D.Dumitrescu, *Dynamically improving ant system*, Automation Computers Applied Mathematics (ACAM), vol.15, no.1, 2007, pp.7-13, (ISSN: 1221-437X).

(57) C. Chira, D. Dumitrescu, R. Găceanu, *Stigmergic Agent Systems for Solving NP-hard Problems*, Studia Univ. Babeş-Bolyai, Informatica, Special Issue, 2007, pp. 177-184.

(58) C. Chira, C-M. Pintea, D. Dumitrescu, *Sensitive ant systems in combinatorial optimization*, Studia Univ. Babeş-Bolyai, Informatica, Special Issue, 2007, pp. 185-192.

(59) C. Chira, C-M. Pintea, D. Dumitrescu: *Stigmergic Agent Optimization*, Romanian Journal of Information Science and Technology (ROMJIST), Ed. Romanian Academy, Bucharest, Vol. 9, No.3, 2007, pp. 175-183 (ISSN: 1453-8245).

(60) L. Dioşan, D. Dumitrescu, *Evolutionary coalition formation in complex network.* Studia Univ. Babeş-Bolyai, Informatica, Vol. LII, No. 2, 2007, pp. 115-129.

(61) C. Stoean, D. Dumitrescu, *Elitist Generational Genetic Chromodynamics as a Learning Classifier System*, Annals of Univ. of Craiova, Mathematics and Computer Science Series, Vol. 33, pp. 132-140, ISSN 1223-6934, 2007.

(62) R. Stoean, C. Stoean, M. Preuss, D. Dumitrescu, *Evolutionary Detection of Separating Hyperplanes in E-mail Classification*, Acta Cibiniensis, Vol. LV Technical series, Lucian Blaga Univ. Sibiu Press, pp. 41-46, 2007.

(63) R.I. Lung, D. Dumitrescu, *An evolutionary model for solving multiplayer games*, Studia Univ. Babeş-Bolyai, Informatica, Special Issue 2007, 209-214.

(64) C-M. Pintea, D. Dumitrescu, P.C. Pop, *Combining Heuristics and Modifying Local Information to Guide Ant-based Search*, Carpathian J. Math., 2007.

(65) R.I. Lung, D. Dumitrescu, *A new evolutionary model for detecting multiple optima*, Genetic And Evolutionary Computation Conference, H. Lipson (Ed), 2007, pp. 1296-1303, ACM Press, ISBN 978-1-59593-697-4.

(66) R.I. Lung, D. Dumitrescu, *A new collaborative evolutionary-swarm optimization technique*, Genetic And Evolutionary Computation Conference, H. Lipson (Ed), 2007, pp. 2817-2820, ACM Press, ISBN 978-1-59593-697-4.

(67) R.I. Lung, D. Dumitrescu, *Guided hyperplane evolutionary algorithm*, Genetic And Evolutionary Computation Conference, H. Lipson (Ed), 2007, pp. 884-891, ACM Press, ISBN 978-1-59593-697-4.

(68) R.I. Lung, D. Dumitrescu, *A Collaborative Model for Tracking Optima.* In 2007 IEEE Congress on Evolutionary Computation, 2007, pp. 564-567, IEEE Press, ISBN 1-4244-1340-0.

(69) A. Gog, D. Dumitrescu, B. Hirsbrunner, *New Selection Operators based on Genetical Relatedness for Evolutionary Algorithms.* Proc. of Congress on Evolutionary Computation (CEC 2007), Singapore, 2007, pp. 4610-4614.

(70) R. Stoean, M. Preuss, C. Stoean, D. Dumitrescu, *Concerning the Potential of Evolutionary Support Vector Machines*, Proc. of Congress on Evolutionary Computation (CEC 2007), Singapore, 2007, pp. 1436-1443.

(71) C. Chira, D. Dumitrescu, *Agent-Based Management and Optimization System for Distributed Computing*, The Third IASTED International Conference on Computational Intelligence, Banff, Canada, 2007, R. Andonie (Ed), ACTA Press, 2007, pp 28-34, ISBN 978-0-88986-671-3.

(72) C-M. Pintea, P.C. Pop, D. Dumitrescu, *An Ant-based Technique for the Dynamic Generalized Traveling Salesman Problem*, Proc. of the 7-th Int. Conf. on Systems Theory and Scientific Computation, Athens, Greece, (Le M.H. et al. Eds.), 2007, pp. 257-261. (ISSN: 1790-5117 ISBN: 978-960-8457-98-0).

(73) C. Stoean, M. Preuss, R. Stoean, D. Dumitrescu, *Disburdening the Species Conservation Evolutionary Algorithm of Arguing with Radii*, The ACM Genetic and Evolutionary Computation Conference (GECCO 2007), London, UK, 2007, pp. 1420-1427.

(74) A. Gog, D. Dumitrescu, B. Hirsbrunner, *Collaborative Evolutionary Algorithms for Combinatorial Optimization.* Proc. of the Genetic and Evolutionary Computation Conference (GECCO 2007), London, UK, 7-11 July 2007, p. 1511.

(75) D. Iclănzan, D. Dumitrescu. *Overcoming hierarchical difficulty by hill-climbing the building block structure.* In GECCO '07: Proc. of the 9th annual conference on Genetic and Evolutionary Computation, Dirk Thierens et al. (Eds), London, 7-11 July 2007, Vol. 2, pp. 1256-1263. ACM Press.

(76) A. Gog, D. Dumitrescu, B. Hirsbrunner, *Best - Worst Recombination Scheme for Combinatorial Optimization.* Proc. of the International Conference on Genetic and Evolutionary Methods (GEM 2007), Las Vegas, USA, 2007, pp. 115-119.

(77) A. Gog, D. Dumitrescu, B. Hirsbrunner, *Community Detection in Complex Networks using Collaborative Evolutionary Algorithms.* Proc. of European Conference on Artificial Life (ECAL 2007), Lisbon, LNCS 4648, Springer, 2007, pp. 886-894.

(78) R.I. Lung, D. Dumitrescu, *Collaborative Evolutionary Swarm Optimization with a Gauss Chaotic Sequence Generator*, Innovations in Hybrid Intelligent Systems, Springer, E. Corchado, J. Corchado, A. Abraham (Eds.), Springer, 2007, pp. 207-214, ISBN 978-3-540-74971-4.

(79) C. Chira, C-M. Pintea, D. Dumitrescu, *Sensitive Stigmergic Agent Systems - A Hybrid Approach to Combinatorial Optimization*, Advances in Soft Computing, Innovations in Hybrid Intelligent Sytems, E. Corchado, J. M. Corchado, A. Abraham (Eds), Springer, 2007, pp. 33-39.

(80) C. Chira, D. Dumitrescu, *Multi-Agent Cooperative Design Support in Distributed Environments.* In Proc. of the 27th international Conference on Distributed Computing Systems Workshops (June 22 - 29, 2007). ICDCSW. IEEE Computer Society, Washington, DC, p. 76.

(81) L. Dioşan, D. Dumitrescu, *A hybrid genetic algorithm based on the Potts system.* In V. Negru, T. Jebelean, D. Petcu, D. Zaharie Eds., 7th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing SYNASC 2007, pp. 453-456, IEEE Computer Society Press.

(82) D. Iclănzan, D. Dumitrescu, *Overrepresentation in neutral genotype-phenotype mappings and their applications.* In V. Negru, T. Jebelean, D. Petcu, D. Zaharie Eds., 7th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing SYNASC 2007, Timişoara, Romania, 26-29 September 2007, pp. 427-432. IEEE Computer Society Press.

(83) D. Iclănzan, P.I. Fulop, D. Dumitrescu, *Neuro-Hill-Climber: A new approach towards more intelligent search and optimization.* In V. Negru, T. Jebelean, D. Petcu, D. Zaharie Eds., 7th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing SYNASC 2007, Timişoara, Romania, 26-29 September 2007, pp. 441-448. IEEE Computer Society Press.

(84) R. Stoean, C. Stoean, M. Preuss, D. Dumitrescu, *Forecasting Soybean Diseases from Symptoms by Means of Evolutionary Support Vector Machines*, Phytologia Balcanica, Vol. 12, No. 3, 2006, pp. 345-350. ISSN 1310-7771.

(85) C. Chira, D. Dumitrescu, *Multi-Agent Systems and Ontologies for Distributed Collaboration*, Transactions on Information Science and Applications, Vol. 3, Issue 8, 2006, pp. 1452-1460, ISSN 1790-0832.

(86) D. Dumitrescu, C. Stoean, *Genetic Chromodynamics Metaheuristic for Multimodal Optimization*, Transactions on Information Science and Application, Issue 8, Volume 3, 2006, pp. 1444-1452, ISSN 1790-0832.

(87) C. Rotar, D. Dumitrescu, R.I. Lung, *Optimization using an Evolutionary Hyperplane Guided Approach*, Acta Univ. Apulensis 11, 2006, pp. 49-63.

(88) L. Dioşan, D. Dumitrescu, D. David, *Far From Equilibrium Computation and Particle Swarm Optimization*, Acta Univ. Apulensis, 11, 2006, pp. 339-352.

(89) C. Chira, C.-M. Pintea, D. Dumitrescu, *Stigmergic Agent Optimization*, Romanian Journal of Information Sciences and Technology, Vol. 9, No. 3, 2006, pp. 175-183.

(90) D. Dumitrescu, A. Roth, *Evolutionary optimization of Coercive Functionals Defined Euler-Monge Surfaces with Fixed Boundary Curves*, Int.J.Comp., Comm. & Control, CCC Publisher, Supplementary issue, 2006, Vol. 1, pp. 31-40, ISSN 1841-9836.

(91) C-M.Pintea, D. Dumitrescu, *Dynamically improving ant system*, Automation Computers Applied Mathematics (ACAM), Vol. 15, No. 1, 2006, pp. 7-13. (ISSN: 1221-437X).

(92) R.I. Lung, D. Dumitrescu, *Collaborative Optimization in Dynamic Environments*, Int.J.Comp., Comm. & Control, CCC Publisher, Supplementary issue, Vol. 1, 2006, pp. 295-301. ISSN 1841-9836.

(93) C. Stoean, D. Dumitrescu, *Elitist Generational Genetic Chromodynamics as a Learning Classifier System*, Annals of Univ. of Craiova, Mathematics and Computer Science Series, Vol. 33, 2006, pp. 132-140, ISSN 1223-6934.

(94) R. Stoean, D. Dumitrescu, C. Stoean, *Nonlinear Evolutionary Support Vector Machines. Application to Classification*, Studia Univ. Babeş-Bolyai, Informatica, Vol. LI, No. 1, 2006, pp. 3-12.

(95) C. Stoean, R. Stoean, M. Preuss, D. Dumitrescu, *A Cooperative Evolutionary Algorithm for Classification*, Int.J.Comp., Comm. & Control, CCC Publisher, Supplementary issue, Vol. 1, 2006, pp. 417-422, ISSN 1841-9836.

(96) C.-M. Pintea, C. Chira, D. Dumitrescu, *Combining Meta-heuristics to Solve the Rook Problem*, In Proc. International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, V. Negru, D. Petcu, D. Zaharie, A. Abraham, B. Buchberger, A. Cicortaş, D. Gorgan, J. Quinqueton (Eds), 2006, pp 239-243, IEEE Computer Society, ISBN 0-7695-2740-X.

(97) A. Gog, D. Dumitrescu, *A New Recombination Operator for Permutation Based Encoding.* In Proc. of the 2nd IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), UT Pres Publishing House, - ISBN (10) 973-662-233-9, 2006, p. 11-16.

(98) D. Dumitrescu, C. Stoean, *The Genetic Chromodynamics Metaheuristic*, In Transactions on Information Science and Application, Proc. of 5th International Conference on Telecommunications and Informatics (TELE-INFO 2006), ISBN 960-8457-45-9, 2006, M. Demiralp, A. Akan, N. Mastorakis, Ed., p. 92-97.

(99) C.-M. Pintea, C. Chira, D. Dumitrescu, *Agent-based Approaches for the Rook Problem*, Proc. of SYNASC 2006, International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, Lisa O'Conner (Ed.), pp. 239-243, IEEE, ISBN 0-7695-2740-X.

(100) C. Stoean, M. Preuss, D. Dumitrescu, R. Stoean, *Cooperative Evolution of Rules for Classification*, Proc. of SYNASC 2006, International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, Lisa O'Conner (Ed.), pp. 317-322, IEEE, ISBN 0-7695-2740-X.

(101) R. Stoean, M. Preuss, D. Dumitrescu, C. Stoean, *Evolutionary Support Vector Regression Machines*, Proc. of SYNASC 2006, International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, Lisa O'Conner (Ed.), pp. 330-335, IEEE, ISBN 0-7695-2740-X.

(102) R. Stoean, C. Stoean, M. Preuss, E. El-Darzi, D. Dumitrescu, *Evolutionary Support Vector Machines for Diabetes Mellitus Diagnosis*, Proc. 3rd International IEEE Conference on Intelligent Systems - IS 2006, Univ. of Westminster, London, 2006, pp. 182-187, ISBN 1-4244-0196-8.

(103) R. Stoean, C. Stoean, M. Preuss, D. Dumitrescu, *Evolutionary Support Vector Machines for Spam Filtering*, RoEduNet IEEE International Conference, Sibiu, Romania, 2006, pp. 261-266.

(104) C. Chira, D. Dumitrescu, *A Multi-Agent Knowledge Management Architecture*, Proc. of International Conference on Multidisciplinary Information Sciences and Technology, Vol. II, 2006, pp. 184-188.

(105) D. Dumitrescu, K. Simon, *Genetic Chromodynamics: A New Evolutionary Optimization Metaheuristics.* Clustering and Training Applications, BIC-TA, Proc. of BIC-TA, Vol. of Evolutionary Computing Section, D.Dumitrescu, L.Pan (Eds), National Natural Science Foundation of China, 2006, pp. 107-116.

(106) C. Chira, C.-M. Pintea, D. Dumitrescu, *Stigmergic Agents for Solving NP-difficult Problems*, BIC-TA, Proc. of BIC-TA, Vol. of Evolutionary Computing Section, D.Dumitrescu, L.Pan (Eds), National Natural Science Foundation of China, 2006, pp. 62-69.

(107) C. Rotar, D. Dumitrescu, R.I. Lung, *An Evolutionary Hyperplane guided Approach for Multicriteria Optimization*, BIC-TA, Proc. of BIC-TA, Vol. of Evolutionary Computing Section, D.Dumitrescu, L.Pan (Eds), National Natural Science Foundation of China, 2006, pp. 70-79.

(108) A. Gog, D. Dumitrescu, *Adaptive Search in Evolutionary Combinatorial Optimization*, BIC-TA, Proc. of BIC-TA, Vol. of Evolutionary Computing Section, D.Dumitrescu, L.Pan (Eds), National Natural Science Foundation of China, 2006, pp. 123-130.

(109) C. Stoean, D. Dumitrescu, M. Preuss, R. Stoean, *Cooperative Coevolution for Classification*, Bio-Inspired Computing: Theory and Applications, BIC-TA 2006, Proc. of BIC-TA, Vol. of Evolutionary Computing Section, D.Dumitrescu, L.Pan (Eds), National Natural Science Foundation of China, 2006, pp. 289-298.

(110) R. Stoean, D. Dumitrescu, M. Preuss, C. Stoean, *Different Techniques of Multi-class Evolutionary Support Vector Machines*, Bio-Inspired Computing: Theory and Applications, BIC-TA 2006, Proc. of BIC-TA, Vol. of Evolutionary Computing Section, D.Dumitrescu, L.Pan (Eds), National Natural Science Foundation of China, 2006, pp. 299-306.

(111) C. Chira, C.-M. Pintea, D. Dumitrescu, *A Multi-Agent Approach to Distributed Collaboration*, 6th Joint Conference on Mathematics and Computer Science, Volume of Abstracts of of 6th Joint Conference on Mathematics and Computer Science,Pecs Univ, ed. Z. Csornyei, 2006, pp. 26-27.

(112) R.I. Lung, D. Dumitrescu, *Considerations on evolutionary detecting Nash equilibria versus Pareto frontier.* 6th Joint Conference on Mathematics and Computer Science, Pecs, HU, 2006.

(113) C. Chira, D. Dumitrescu, *Development of a Multi-Agent Information Management System*, Proc. of 6th International Conference on Recent Advances in Soft Computing (RASC 2006), K. Sirlantzis Ed., 2006, pp. 19-25.

(114) C. Chira, D. Dumitrescu, *Multi-Agent Systems in Distributed Communication Environments*, Proc. of 5th International Conference on Telecommunications and Informatics (TELE-INFO 2006), 2006, M. Demiralp, A. Akan, N. Mastorakis, Ed., pp. 267-272, ISBN 960-8457-45-9.

(115) D. Dumitrescu, C. Stoean, *Genetic chromodynamics - a novel evolutionary heuristic for multimodal optimization*, First International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006, Current Research in Information Sciences and Technologies. Multidisciplinary approaches to Global Information Systems, Volume 2, V.P. Guerrero-Bote (Ed.), Open Institute of Knowledge, Merida, Spain, October 25 - 28, 2006, pp. 238-243.

(116) C. Chira, D. Dumitrescu, *A Multi-Agent Knowledge Management Architecture*, First International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006, Current Research in Information Sciences and Technologies. Multidisciplinary approaches to Global Information Systems, Volume 2, V.P. Guerrero-Bote (Ed.), Open Institute of Knowledge, Merida, Spain, October 25 - 28, 2006, pp. 184-188.

(117) C-M. Pintea, D. Dumitrescu, *A dynamic approach for improving ant system*, The Tenth International Conference on Applied Mathematics and Computer Science, Volume of Abstracts of The Tenth International Conference on Applied Mathematics and Computer Science, 2006, pp. 54-55.

(118) D. David, L. Dioşan, D. Dumitrescu, *A Far From Equilibrium Computation System*, the 6th IEEE Communications International Conference, Proc. of the 6th IEEE Communications International Conference, 2006, pp. 245-248, ISBN 973-718-479-3.

(119) C-M. Pintea, D. Dumitrescu, P. Petrică, *Reinforcing Ant Colony System for the Generalized Traveling Salesman Problem*, Fifth International Workshop on Ant Colony Optimization and Swarm Intelligence, Universit Libre de Bruxelles, Brussels, 2006.

(120) C.-M. Pintea, D. Dumitrescu, *The Importance of Parameters in Ant System*, In Proc. of the International Conference on Computers, Communications & Control 2006, Agora Univ., Ioan Dziţac, Florin Filip, Mişu Manolescu (Eds.), 2006, pp. 387-392, ISSN 1841-9836.

(121) C. Stoean, M. Preuss, D. Dumitrescu, R. Stoean, *A Cooperative Coevolutionary Algorithm for Multi-class Classification*, Proc. of Workshop on Agents for Complex Systems, 8th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2006), pp. 7-14, ISBN 978-0-7695-2740-X.

(122) R. Stoean, M. Preuss, D. Dumitrescu, C. Stoean, *e - Evolutionary Support Vector Regression*, Proc. of Workshop on Agents for Complex Systems, 8th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2006), pp 21-27, ISBN 978-0-7695-2740-X.

(123) C. Stoean, R. Stoean, M. Preuss, D. Dumitrescu, *Spam Filtering by Means of Cooperative Coevolution*, 4th European Conference on Intelligent Systems and Technologies, ECIT 2006 Advances in Intelligent Systems and Technologies - Selected Papers, H. N. Teodorescu (Ed.), Performantica Press, 2006, pp. 157-159, ISBN 973-730-246-X.

(124) C.-M. Pintea, D. Dumitrescu, *A dynamic approach for improving ant system*, Proc. of Workshop Th. Angheluţă 2006, pp. 54-55.

(125) D. Dumitrescu, K. Simon, *A New Dynamic Evolutionary Technique. Application in Designing RBF Neural Network Topologies. II. Numerical Experiments*, Studia Univ. Babeş-Bolyai, Informatica, Vol. L, 2005, pp. 56-69.

(126) C. Groşan, M. Oltean, D. Dumitrescu, *Adaptive Representation for Multiobjective Optimization*, Journal of Applied Mathematics and Computer Science, Poland, 2005.

(127) A. Joo, D. Dumitrescu, *On Growing Generalized Decision Trees*, Complex Systems, 2005.

(128) A. Joo, D. Dumitrescu, *Generalized Decision Trees Built With Evolutionary Techniques*, Studies in Informatics and Control, 2005, pp. 117-132.

(129) A. Gog, D. Dumitrescu, *A New Search Model for Evolutionary Algorithms*, Acta Univ. Apulensis, 10, 2005, pp. 73-78.

(130) A. Joo, D. Dumitrescu, *Generalized Decision Trees Built With Evolutionary Techniques*, Studies In Informatics and Control, 2005, pp. 117-132.

(131) C.-M. Pintea, D. Dumitrescu, *Improving ant systems using a local updating rule*, IEEE Proc. of SYNASC 2005, International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, D. Zaharie et al. (Eds.), 2005, pp. 295-299, ISBN 0-7695-2453-2.

(132) C. Stoean, M. Preuss, R. Gorunescu, D. Dumitrescu, *New Radii-Based Evolutionary Model for Multi-modal Optimisation*, Proc. of The 2005 IEEE Congress on Evolutionary Computation - CEC 2005, IEEE Press, 2005, pp. 1839-1846.

(133) A. Gog, D. Dumitrescu, *A Model for Parallel Evolutionary Search*, IEEE Proc. of SYNASC 2005, International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, D. Zaharie et al., 2005, pp. 63-67, ISBN 0-7695-2453-2.

(134) D. David, L. Dioşan, D. Dumitrescu, *A New Nature-Inspired Computational Model: Ising Model with Rays*, IEEE Proc. of SYNASC 2005, International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, D. Zaharie et al., 2005, pp. 315-320, ISBN 0-7695-2453-2.

(135) A. Gog, D. Dumitrescu, *A New Search Model for Evolutionary Algorithms*, Proc. of International Conference on Theory and Applications in Mathematics and Informatics - ICTAMI 2005, D. Breaz Ed, pp. 48-49, ISSN 1582-5329.

(136) C. Stoean, M. Preuss, R. Gorunescu, D. Dumitrescu, *Diabetes Diagnosis through the Means of a Multimodal Evolutionary Algorithm*, Proc. of the First East European Conference on Health Care Modelling and Computation - HCMC 2005, Craiova Medicala, F. Gorunescu. E. El-Darzi, M. Gorunescu (Eds.), 2005, pp. 277-289, ISBN 973-7757-67-X.

(137) C. Stoean, D. Dumitrescu, *On Solving the 3-SAT Problem Using an Evolutionary Multimodal Optimization Technique*, Proc. of 5th International Conference on Artificial Intelligence and Digital Communications - AIDC 2005, Universitaria Publishing House, N. Ţăndăreanu (Ed.), 2005, pp. 136-142, ISBN 973-671-014-9.

(138) L. Dioşan, D. Dumitrescu, D. David, *Far From Equilibrium Computation and Particle Swarm Optimization*, Proc. of International Conference on Theory and Applications in Mathematics and Informatics - ICTAMI 2005, D. Breaz Ed, 2005, pp. 42-43, ISSN 1582-5329.

(139) K. Simon, F.J. Szabo, D. Dumitrescu, *New Techniques for Evolutionary Optimization and Clustering*, Proc. of International Conference on Theory and

Applications in Mathematics and Informatics - ICTAMI 2005, D. Breaz Ed, 2005, p. 58, ISSN 1582-5329.

(140) R.I. Lung, D. Dumitrescu, *Evolutionary Multimodal Optimization - new Techniques*, Proc. of International Conference on Theory and Applications in Mathematics and Informatics - ICTAMI 2005, D. Breaz Ed, 2005, p. 60, ISSN 1582-5329.

(141) D. Noje, D. Dumitrescu, *On DMV-algebras and Product MV-algebras*, Proc. of Second Romanian-Hungarian Joint Symposium on Applied Computentional Inteligence, 2005, pp. 467-477.

(142) D. Dumitrescu, D. Noje, *Fuzzy connectives, residuated lattices and BL-algebras*, Proc. of Second Romanian-Hungarian Joint Symposium on Applied Computentional Inteligence, 2005, pp. 183-200.

(143) D. David, L. Dioşan, D. Dumitrescu, *A new Computational Model Based on Ising Machine - Ising model with rays*, Proc. of Workshop, NCA, 7th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2005, pp. 45-51.

(144) C.-M. Pintea, D. Dumitrescu, *Inner-Update System for Traveling Salesman Problem*, Proc. of Workshop, NCA, 7th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2005.

(145) C. Stoean, R. Gorunescu, D. Dumitrescu, *A New Evolutionary Model for the Optimization of Multimodal Functions*, The Anniversary Symposium Celebrating 25 Years of the Seminar Grigore Moisil and 15 Years of the Romanian Society for Fuzzy Systems and A.I., pp. 65-72, ISBN 973-730-070-X, 2005.

(146) R. Stoean, D. Dumitrescu, *Evolutionary Support Vector Machines - a New Learning Paradigm The Linear Non-separable Case*, Proc. of Zilele Academice Clujene, 2005, pp. 15-20.

(147) C. Stoean, D. Dumitrescu, *Cloning within Genetic Chromodynamics*, Proc. of Zilele Academice Clujene, 2005, pp. 9-14.

(148) D. Dumitrescu, C. Groşan, M. Oltean, *Adaptive Representation for Multiobjective Optimization*, International Journal of Applied Mathematics and Computer Science, 2004.

(149) A. Gog, D. Dumitrescu, *Parallel Mutation Based Genetic Chromodynamics*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLIX, 2004, pp. 45-54.

(150) D. Dumitrescu, R. Gorunescu, *Evolutionary Clustering Using Adaptive Prototypes*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLIX, 2004, pp. 15-21.

(151) D. Dumitrescu, R.I. Lung, *Roamimg Optimization: a New Evolutionary Technique for Multi-Modal Optimizaton*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLIX, No. 1, 2004, pp. 99-110.

(152) D. Dumitrescu, C. Groşan, V. Varga, *Stochastic Optimization of Querying Distributed Databases III: Evolutionary Method versus Constructive Method*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLIX, No. 1, 2004, pp. 3-14.

(153) D. Dumitrescu, K. Simon, *Evolutionary RBF neural network design*, Analele Univ. de Vest, Timişoara, 51, 2004, pp. 97-110.

(154) D. Dumitrescu, R.I. Lung, *Roamimg Optimization, II*, Analele Univ. de Vest Timişoara, 51, 2004, pp. 155-163.

(155) R. Gorunescu, D. Dumitrescu, *An Evolutionary Approach to Fuzzy Clustering*, Research Notes in Artificial Intelligence and Data Communications, 2004, pp. 51-55.

(156) C. Stoean, R. Gorunescu, M. Preuss, D. Dumitrescu, *An Evolutionary Learning Classifier System Applied to Text Categorization*, Annals of West University of Timisoara, Mathematics and Computer Science Series, Vol. XLII, Special Issue 1, 2004, pp. 265-278.

(157) V. Varga, D. Dumitrescu, C. Groşan, *Solving Stochastic Optimisation in Distributed Databases using Genetic Algorithms*, Advances in Databases and Information Systems, 8th East-European Conference, ADBIS 2004, Budapest, September 2004, LNCS 3255, Springer, pp. 259-274.

(158) M. Oltean, D. Dumitrescu, *Evolving TSP Heuristics using Multi Expression Programming*, Proc. of International Conference on Computational Sciences, ICCS'04, 6-9 June, Krakow, Poland, M. Bubak, G. D. van Albada, P. Sloot, J. Dongarra (Eds.), Vol II, Springer, 2004, pp. 670-673.

(159) M. Oltean, D. Dumitrescu, *A Permutation based Approach for the 2-D Cutting Stock Problem*, Proc. of First International Industrial Conference Bionik 2004, I. Boblan, R. Bannasch (Eds.), 2004, pp. 73-80.

(160) D. Dumitrescu, A. Joo, *Generalized Decision Trees Built With Evolutionary Techniques*, Proc. of International Symposium on Symbolic and Numeric Algorithms for Scientific Computing - SYNASC 2004, pp. 270-279.

(161) D. Dumitrescu, F. Jarai-Szabo, K. Simon, *Link - cell methods for dynamic evolutionary clustering*, Proc. of International Symposium on Symbolic and Numeric Algorithms for Scientific Computing - SYNASC 2004, pp. 480-490.

(162) C. Groşan, D. Dumitrescu, *Quantum Evolutionary Algorithms*, Proc. of International Symposium on Symbolic and Numeric Algorithms for Scientific Computing - SYNASC 2004, pp. 460-469.

(163) C. Stoean, R. Gorunescu, M. Preuss, D. Dumitrescu, *Evolutionary Detection of Rules for Text Categorization. Application to Spam Filtering*, Proc. of Third European Conference on Intelligent Systems and Technologies-ECIT'2004, 2004, pp. 87-95.

(164) C. Stoean, R. Gorunescu, M. Preuss, D. Dumitrescu, *Evolutionary Discovery of Adaptive Rules for Spam Detection*, Proc. of ICAM4, Fourth International Conference on Applied Mathematics, 2004, p. 34.

(165) C. Stoean, R. Gorunescu, M. Preuss, D. Dumitrescu, *An Evolutionary Learning Spam Filter System*, Proc. of International Symposium on Symbolic and Numeric Algorithms for Scientific Computing - SYNASC 2004, pp. 512-522.

(166) A. Gog, D. Dumitrescu, *A new class of evolutionary optimization.* International Conference on Applied Mathematics (ICAM 4) Baia Mare 2004.

(167) D. Dumitrescu, R.I. Lung, *Roamimg Optimization: a New Evolutionary Technique for Multi-Modal Optimization*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLIX, No. 1, 2004, pp. 99-110.

(168) D. Dumitrescu, A. Joo, *MEPDTI2: Evolution of Generalized Decision Trees*, Proc. of International Conference on Computers and Communication (ICCC), 2004, pp. 123-128, ISBN 973-613-542-X.

(169) D. Dumitrescu, K. Simon, *Fitness functions and interaction domain adaptation mechanisms for dynamic evolutionary clustering*, Proc. of International Conference on Computers and Communication (ICCC), 2004, pp. 132-138, ISBN 973-613-542-X.

(170) D. Dumitrescu, K. Simon, M Gaspar, *Adapting parameters for dynamic evolutionary clustering*, Proc. of International Conference on Computers and Communication (ICCC), 2004, pp. 158-162, ISBN 973-613-542-X.

(171) D. Dumitrescu, A. Gog, *A new evolutionary technique for multimodal optimization*, Proc. of International Conference on Computers and Communication (ICCC),2004, pp. 119-123, ISBN 973-613-542-X.

(172) C. Groşan, D. Dumitrescu, A. Lazăr, *Particle Swarm Optimization for solving 0/1 Knapsack Problem*, Proc. of International Conference on Computers and Communication (ICCC), 2004, pp. 172-183, ISBN 973-613-542-X.

(173) D. Dumitrescu, R. Gorunescu, *Evolutionary Adaptive Fuzzy Clustering*, Proc. of Zilele Academice Clujene, 2004, pp. 61-67.

(174) D. Dumitrescu, K. Simon, *Post - procesing Tehniques for Evolutionary Clustering*, Proc. of Zilele Academice Clujene, 2004, pp. 75-82.

(175) M. Gaspar, D. Dumitrescu, *Using Multiple Radiuses for a Prototype in Dynamic Evolutionary Clustering*, Proc. of Zilele Academice Clujene, 2004, pp. 55-60.

(176) C. Groşan, D. Dumitrescu, A. Lazăr, *Genetic algorithms for solving Cryptarithms Problem*, Proc. of Zilele Academice Clujene, 2004, pp. 89-94.

(177) C. Groşan, D. Dumitrescu, *Genetic Algorithms for Solving geometrical place problems.* Lucrările seminarului Didactica Matematicii, Vadu Crişului, 2004.

(178) D. Dumitrescu, C. Groşan, V. Varga, *Stochastic Optimization of Querying Distributed Databases I.:Theory of Four Relations Join*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLVIII, 2003, pp. 79-89.

(179) D. Dumitrescu, C. Groşan, V. Varga, *Stochastic Optimization of Querying Distributed Databases II. Solving Stochastic Optimization*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLVIII, 2003, pp. 17-25.

(180) D. Dumitrescu, R. Gorunescu, *Evolutionary clustering using an incremental technique*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLVIII, 2003, pp. 25-33.

(181) D. Dumitrescu, K. Simon, *A new dynamic evolutionary clustering technique and its application in designing RBF neural*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLVIII, 2003, pp. 45-53.

(182) D. Dumitrescu, A. Joo, *Evolving orthogonal decision tress*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLVIII, 2003, pp. 33-45.

(183) C. Groşan, D. Dumitrescu. *Evolutionary Optimization.* Lucrările seminarului Didactica Matematicii, Vadu-Crişului, 2003.

(184) D. Dumitrescu, R. Gorunescu, *Adaptive prototypes in evolutionary clustering*, 3rd International Conference on Artificial Intelligence and Digital Communications, Craiova, 2003, Research Notes in Computer Science, N. Ţăndăreanu (Ed.), 103, pp. 48-55, Reprograph Press.

(185) R. Gorunescu, D. Dumitrescu, *An Evolutionary Approach to Fuzzy Clustering*, 4th International Conference on Artificial Intelligence and Digital Communications, Craiova, 2004, Research Notes in Computer Science, N. Ţăndareanu (Ed.), 104, pp. 51-55, Reprograph Press.

(186) D. Dumitrescu, K. Simon, *Evolutionary prototype selection*, Proc. of International Conference on Theory and Applications in Mathematics and Informatics - ICTAMI 2003, 2003, pp. 183-191.

(187) C. Groşan, M. Oltean, D. Dumitrescu, *A new evolutionary algorithm for the multiobjective 0-1 knapsack problem*, Proc. of International Conference on Theory and Applications in Mathematics and Informatics - ICTAMI 2003, pp. 233-238.

(188) C. Groşan, M. Oltean, D. Dumitrescu, *A modified PAES algorithm using adaptive representation of solutions*, Proc. of International Conference on Theory and Applications in Mathematics and Informatics - ICTAMI 2003.

(189) C. Groşan, D. Dumitrescu, *A Comparison of Multiobjective evolutionary algorithms*, Proc. of International Conference on Theory and Applications in Mathematics and Informatics - ICTAMI 2003, pp. 101-111.

(190) R.I. Lung, D. Dumitrescu, *A New Evolutionary optimisation tehnique*, Proc. of International Conference on Theory and Applications in Mathematics and Informatics - ICTAMI 2003, pp. 263-271.

(191) R.I. Lung, D. Dumitrescu, *Roaming Optimisation: a New Evolutionary Tehnique*, Proc. of International Symposium on Symbolic and Numeric Algorithms for Scientific Computing - SYNASC 2003, pp. 149-157.

(192) D. Dumitrescu, K. Simon, *Genetic Cromodynamics for Designing RBF Neural Networks*, Proc. of International Symposium on Symbolic and Numeric Algorithms for Scientific Computing - SYNASC 2003, pp. 91-102.

(193) D. Dumitrescu, R. Gorunescu, *Adaptive prototypes in evolutionary clustering*, Internat. AI Conference, Craiova 2003, pp. 21-30.

(194) D. Dumitrescu, K. Simon, *Evolutionary clustering technique for designing RBF neural network*, Proc. of the Symposium Colocviul Academic Clujean de Informatică, Cluj-Napoca, June 2003, pp. 137-143.

(195) D. Dumitrescu, K. Simon, *Reducing complexity of RBF neural network by dynamic evolutionary clustering*, Proc. 11th Conference on Applied Mathematics, CAIM 2003, Oradea, 2003, pp. 83-90.

(196) D. Dumitrescu, R.I. Lung, *Evolutionary multimodal optimisation.* Proc. 11th Conference on Applied Mathematics, CAIM 2003, Oradea, 2003, pp. 73-77.

(197) D. Dumitrescu, C. Florea, *Evolving multiagent systems.* Proc. 11th Conference on Applied Mathematics, CAIM 2003, Oradea, vol 2.

(198) D. Dumitrescu, D. Noje, *Double product MV-algebras versus Product MV-algebras*, Proc 11th Conference on Applied Mathematics, CAIM 2003, Oradea, 2003, pp. 78-82.

(199) C. Groşan, M. Oltean, D. Dumitrescu, *Performance metrics for multiobjective optimization evolutionary algorithms*, Proc. of Conference on Applied and Industrial Mathematics (CAIM), Oradea, 2003.

(200) A. Joo, D. Dumitrescu, *A MEP-Based Technique for Evolving Decision Trees*, Proc. of the Symposium Colocviul Academic Clujean de Informatică, Cluj-Napoca, June 2003, pp. 113-121.

(201) C. Groşan, D. Dumitrescu, *A comparison of multiobjective evolutionary algorithms*, Acta Univ. Apulensis, 4, 2002, pp. 101-111.

(202) C. Groşan, D. Dumitrescu, *A new Evolutionary paradigm for single and multiobjective optimization*, Seminar on Computer Science, Babeş- Bolyai Univ., Cluj-Napoca, 2002.

(203) D. Dumitrescu, C. Florea, P. Pătrânjan, *Evolutionary Reorganization in MAS*; Proc. European Conference Information Technology (ECIT 2002), 2002, pp. 1-5.

(204) D. Dumitrescu, C. Florea, P. Pătrânjan, *A New Evolutionary Model for Multi-Agent Systems*, Proc. of International Symposium on Symbolic and Numeric Algorithms for Scientific Computing - SYNASC 2002, pp. 137-143.

(205) C. Groşan, D. Dumitrescu, *A new Evolutionary paradigm for single and multiobjective optimization*, Seminar on Computer Science, Babeş-Bolyai Univ., Cluj-Napoca, 2002.

(206) D. Dumitrescu, B. Iantovics, C. Florea, *Multi-Agent Systems: a new allocation protocol and evolutionary search for equilibrium*, Proc. Computer Science Conference, Cluj-Napoca, Romania, June, 2002, pp. 14-21.

(207) D. Dumitrescu, M. Chiş, *Evolutionary Hierarchical Clustering for Data Mining*, Proc. of the Symposium Zilele Academice Clujene, Computer Science Section, 14-22 June 2002, Seminar on Computer Science, pp. 12-18.

(208) D. Dumitrescu, C. Groşan, M. Oltean, *Genetic Chromodynamics for Obtaining Continuous Representation of Pareto Regions*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLVI, No. 2, 2001, pp. 15-30.

(209) D. Dumitrescu, C. Groşan, M. Oltean, *A new evolutionary adaptive representation paradigm*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLVI, No. 1, 2001, pp. 19-29.

(210) D. Dumitrescu, C. Groşan, M. Oltean, *Simple Multiobjective Evolutionary Algorithm*, Seminar on Computer Science, Babeş- Bolyai Univ., Cluj-Napoca, 2001, pp. 1-12.

(211) D. Dumitrescu, M. Oltean, *Theorem proving using Resolution*, Proc. of the Joint Conference on Mathematics and Computer Science, Oradea 2001.

(212) D. Dumitrescu, C. Groşan, M. Oltean, *A New Evolutionary Approach for Multiobjective Optimization*, Proc. of the Joint Conference on Mathematics and Computer Science, Oradea 2001.

(213) D. Dumitrescu, C. Groşan, M. Oltean, *Genetic Chromodynamics for multimodal and multiobjective optimization*, Proceeding of the Joint Conference on Mathematics and Computer Science, Oradea 2001.

(214) D. Dumitrescu, B. Lazzerini, F. Marcelloni, *A Fuzzy Hierarchical Classification System for Olfactory Signals*, Pattern Analysis and Applications, Vol. 3 No. 4, 2000, pp. 325-334.

(215) D. Dumitrescu, M. Zsolt, *Genetic chromodynamics for multimodal optimization*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLV, 2000.

(216) D. Dumitrescu, C. Hăloiu, A. Dumitrescu, *Generators of Fuzzy Dynamical Systems*, Fuzzy Sets and Systems, 113, 2000, pp. 447-452.

(217) D. Dumitrescu, C. Groşan, M. Oltean, *A New Evolutionary Approach For Multiobjective Optimization*, Studia Univ. Babeş-Bolyai, Informatica, Volume XLV, No. 1, 2000, pp. 51-68.

(218) D. Dumitrescu, *Genetic Chromodynamics*, Studia Univ. Babeş-Bolyai, Informatica, Volume XLV, No. 1, 2000, pp. 39-50.

(219) C. Groşan, D. Dumitrescu, A. Lazăr, *Cryptarithm problems using evolutionary algorithms*, Lucrările seminarului de Didactica Matematicii, Vadu Crişului, 2000.

(220) D. Dumitrescu, B. Lazzerini, F. Marcelloni, *A fuzzy hierarchical system for olfactory signal detection*, Pattern Analysis and Applications, 2000.

(221) V. Chepoi, D. Dumitrescu, *Fuzzy clustering with structural constraints.* Fuzzy Sets and Systems, 105(1999), pp. 91-97.

(222) D. Dumitrescu, M. Oltean, *An Evolutionary Algorithm for Theorem Proving in Propositional Logic*, Studia Univ. Babeş-Bolyai, Informatica, vol. XLIV, No. 2, 1999, pp. 87-99.

(223) D. Dumitrescu, Z. Murgu, *Genetic Chromodynamics for Multimodal Optimization*, Studia Univ. Babeş-Bolyai, Informatica, vol. XLIV, No. 1, 1999, pp. 21-40.

(224) D. Dumitrescu, H.F. Pop, *Convex decomposition of fuzzy partitions, II.* Fuzzy Sets and Systems, 96(1998), pp. 111-118.

(225) D. Avram, D. Dumitrescu, *On Some Applications Of Evolutionary Computation*, Computer Science Research Seminars, Babeş-Bolyai Univ., 2, 1998, pp. 105-118.

(226) D. Dumitrescu, B. Lazzerini, A. Lehene, *A Genetic Algorithm For Symbolic Data Clustering*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLIII, No. 1, 1998, pp. 61-74.

(227) D. Dumitrescu, D. Avram, B. Lazzerini, *Evolutionary Programming: An Application To Clustering*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLIII, No. 1, 1998, pp. 47-60.

(228) D. Dumitrescu, E. Kovacs, *Cluster Prototype Selection By Genetic Chromodynamics*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLIII, No. 2, 1998, pp. 56-63.

(229) D. Dumitrescu, L. Hui, B. Lazzerini, *Hierarchical Data Structure Detection Using Evolutionary Algorithms*, Studia Univ. Babeş-Bolyai, Informatica, Vol. XLIII, No. 2, 1998, pp. 3-12.

(230) A. Dumitrescu, A. Pop, D. Dumitrescu, *Structural properties of pulsating stars light curves through fuzzy divisive hierarchical clustering*, Astrophysics and Space Science, 250(1997), pp. 205-226.

(231) D. Dumitrescu, A. Dumitrescu, *A unified approach to fuzzy pattern recognition*, European Journal of Operational Research, 96(1997), pp. 471-478.

(232) D. Dumitrescu, *Does AI Really Exists?*, Babeş-Bolyai Univ., Faculty of Mathematics and Computer Science Research Seminars, 2(1997), pp. 23-26.

(233) D. Dumitrescu, C. Dodu, *Comparative Study of Genetic Algorithms in Fuzzy Clustering*, Babeş-Bolyai Univ., Faculty of Mathematics and Computer Science, Research Seminars, 2(1997), pp. 115-126.

(234) D. Dumitrescu, L. Bodrogi, *A New Evolutionary Method and its Application in Clustering*, Babeş-Bolyai Univ., Faculty of Mathematics and Computer Science, Research Seminars, 2(1997), pp. 127-134.

(235) D. Dumitrescu, A. Oneţ, *Genetic chromodynamics for clustering*, Babeş-Bolyai Univ., Faculty of Mathematics and Computer Science, Research Seminars, 2(1997), pp. 107-114.

(236) D. Dumitrescu, I. Stan, *Genetic algorithms in neural networks*, in Artificial Intelligence: Methodology, Systems, Applications, A.M. Ramsay (Ed.), IOS Press, 1996, pp. 134-140.

(237) H.F. Pop, C. Sârbu, O. Horowitz, D. Dumitrescu, *A Fuzzy Classification of the Chemical Elements*, Journal of Chemical Information and Computer Sciences, 36, 1996, pp. 465-482.

(238) D. Dumitrescu, I. Stan, *Genetic algorithms in Cognitive Sciences*, National Conf. on Cognitive Sciences, Cluj-Napoca, 1996.

(239) D. Dumitrescu, H.F. Pop, C. Sârbu, *Fuzzy Hierarchical Cross-Classification of Greek Muds*, Journal of Chemical Information and Computer Sciences, 35, 1995, pp. 851-857.

(240) D. Dumitrescu, H.F. Pop, *Degenerate And Nondegenerate Convex Decomposition of Finite Fuzzy Partitions .1.*, Fuzzy Sets and Systems, 73(1995), pp. 365-376.

(241) D. Dumitrescu, H.F. Pop, *Algorithms for convex decomposition of finite fuzzy partitions*, Fuzzy Systems and A.I., 2(1995), pp. 5-12.

(242) H.F. Pop, D. Dumitrescu, C. Sârbu, *A study of Roman pottery (terra sigillata) using hierarchical fuzzy clustering.* Analitica Chimica Acta, 310, (1995), pp. 269-279.

(243) D. Dumitrescu, H.F. Pop, *Convex decomposition of fuzzy partitions*, I. Fuzzy Sets and Systems, 73(1995), pp. 365-376.

(244) D. Dumitrescu, *Entropy of Fuzzy Dynamical-Systems*, Fuzzy Sets and Systems, 70(1995), pp. 45-57.

(245) D. Dumitrescu, *Fuzzy Conditional Logic*, Fuzzy Sets and Systems, 68(1995), pp. 171-179.

(246) D. Dumitrescu, I. Stan, *Fuzzy invariant criteria and genetic algorithms in clustering*, Univ. Babeş-Bolyai Fac. Math. Res. Seminars Computer Science 5(1995), pp. 87-101.

(247) D. Dumitrescu, M. Berchez, *A new clustering method of data obtained in the selection field*, Bulletin USAMV-CN, 49(1995), pp. 173-177.

(248) D. Dumitrescu, *Fuzzy Training Procedures*, Fuzzy Sets and Systems, 67 (1994), pp. 279-291.

(249) D. Dumitrescu, D. Tătar, *Normal fuzzy formulas and their minimization.* Fuzzy Sets and Systems, 64 (1994), pp. 113-117.

(250) D. Dumitrescu, C. Sârbu, H.F. Pop, *A Fuzzy Divisive Hierarchical-Clustering Algorithm For The Optimal Choice Of Sets Of Solvent Systems*, Analytical Letters, 27(1994), pp. 1031-1054.

(251) D. Dumitrescu, *Fuzzy Measures and the Entropy of Fuzzy Partitions*, Journal of Mathematical Analysis and Applications, 176(1993), pp. 359-373.

(252) C. Sârbu, D. Dumitrescu, H.F. Pop, *Selecting and optimally combining the solvent systems in the cromatography on thin film using the fuzzy sets theory*, The Chemistry Review Vol. 44, No. 5, 1993, pp. 450-459.

(253) D. Dumitrescu, *Fuzzy Training Procedures .1*, Fuzzy Sets and Systems, 56 (1993), pp. 155-169.

(254) D. Dumitrescu, *Entropy of a Fuzzy Process*, Fuzzy Sets and Systems, 55 (1993), pp. 169-177.

(255) D. Dumitrescu, *Fuzzy Partitions with the Connectives T-Infinity*, S-Infinity, International Journal for Fuzzy Sets and Systems, 47 (1992), pp. 193-195.

(256) V. Chepoi, D. Dumitrescu, *Cluster analysis and facility location problems*, Babeş-Bolyai Univ. Fac. Math. Res. Seminar on Computer Science 5 (1992), pp. 139-186.

(257) D. Dumitrescu, *Hierarchical cross-classification*, Economic Computation and Economic Cybernetics Studies and Research, (1991).

(258) D. Dumitrescu, *A fuzzy training algorithm.* Studia Univ. Babeş-Bolyai, Mathematica, Vol. XXXV, No. 3, 1990, pp. 7-11.

(259) D. Dumitrescu, H.F. Pop, *A preliminary bibliography on fuzzy clustering and related fields.* Studia Univ. Babeş-Bolyai, Mathematica, Vol. XXXV, No. 3, 1990, pp. 13-24.

(260) D. Dumitrescu, C. Tămaş, *An algorithm for convex decomposition of fuzzy partitions.* Studia Univ. Babeş-Bolyai, Mathematica, Vol. XXXV, No. 3, 1990, pp. 31-36.

(261) D. Dumitrescu, Gh. Lazarovici, *Fuzzy clustering in Archeology*, vol. Romanian Archometry, 1990.

(262) D. Dumitrescu, *On fuzzy hierarchical clustering.* Univ. of Cluj-Napoca, Fac. Math. Res. Seminars Computer Science 9 (1989), pp. 35-40.

(263) D. Dumitrescu, *Fuzzy hierarchical detection of clusters. Divisive hierarchical classification.* Fuzzy Sets and AI, 1989.

(264) D. Dumitrescu, *Clasificare ierarhică simultană (Hierarchical cross-classification).* Studii şi Cercetări de Calcul Economic şi Cibernetică Economică (Economic Computation and Economic Cybernetics Studies and Research), 1989.

(265) D. Dumitrescu, Gh. Lazarovici, *Clasificarea fuzzy a datelor arheologice*, Seminarul UNESCO Informatica aplicata in domeniile socio-umane, 1989.

(266) D. Dumitrescu, *Hierarchical pattern classification*, Fuzzy Sets and Systems, 28 (1988), pp. 145-162.

(267) D. Dumitrescu, *Divisive hierarchical classification.* Economic Computation and Economic Cybernetics Studies and Research 22(1988), pp. 31-38.

(268) D. Dumitrescu, *Hierarchical classification for linear clusters*, Studia Univ. Babeş-Bolyai, Mathematica, Vol XXXIII, No. 3, 1988, pp. 48-51.

(269) D. Dumitrescu, *A note on a fuzzy logic.* Univ. of Cluj-Napoca, Fac. Math. Res. Seminars 9 (1988), pp. 47-51.

(270) D. Dumitrescu, *A note on fuzzy Information Theory.* Studia Univ. Babeş-Bolyai, Mathematica, Vol XXXIII, No. 2, 1988, pp. 65-69.

(271) D. Dumitrescu, *On a fuzzy logic*, Proc. Coll. Logic and Languages, Braşov, 1988, pp. 111-116.

(272) D. Dumitrescu, *On a Fuzzy Conditional Logic*, Fac. Math. Research Seminars, 9, 1988, pp. 47-51.

(273) D. Dumitrescu, *Preliminarii la o teorie a informaţiei în context fuzzy*, Lucr. Laboratorului de Cercetări Interdisciplinare, Cluj-Napoca, 1988, pp. 29-30.

(274) D. Dumitrescu, L. Kekedy, *Pattern Recognition, New Method of Analytical Data Interpretation .2. Classification of Indigenous Mineral Waters Based on Chemical-Analysis Data*, Revista de Chimie, 8, 1987, pp. 428-431.

(275) L. Kekedy, D. Dumitrescu, *Pattern Recognition, A Modern Method of Explaining Analytical Data .1. General-Principles*, Revista de Chimie, 38, 1987, pp. 339-342.

(276) D. Dumitrescu, *Clasificare ierarhică divizivă (Divisive hierarchical classification).* Studii şi Cercetări de Calcul Economic şi Cibernetică Economică (Economic Computation and Economic Cybernetics Studies and Research) 22 (1987), pp. 23-30.

(277) D. Dumitrescu, *Divisive hierarchical clustering.* Studia Univ. Babeş-Bolyai, Mathematica, Vol XXXII, No. 2, 1987, pp. 24-30.

(278) D. Dumitrescu, *Non metric hierarchical fuzzy classification.* Univ. Babeş-Bolyai, Fac. Math. Res. Seminars Computer Science 5 (1987), pp. 22-27.

(279) D. Dumitrescu, L. Kekedy, *Clasificarea unor ape minerale indigene (Classification of some indigene mineral waters).* Revista de Chimie (The Chemistry Reviews) 38 (1987), pp. 417-420.

(280) Dumitrescu, D., *Principal components of a fuzzy class*, Studia Univ. Babeş-Bolyai, Mathematica, Vol XXXII, No. 1, 1987, pp. 24-28.

(281) C. Lenart, D. Dumitrescu, *Convex decomposition of fuzzy partition.* Univ. Babeş-Bolyai, Fac. Math. Res. Seminars Computer Science 5 (1987), pp. 46-54.

(282) D. Dumitrescu, L. Kekedy, *Classification of mineral waters by pattern recognition processing of chemical composition data*, Studia Univ. Babeş-Bolyai, Chemia, Vol. 2, 1987, pp. 68-73.

(283) D. Dumitrescu, *Clasificare şi sinteză în Inteligenţa Artificială*, Lucrările Sesiunii Ştiinţifice a Centrului de Calcul al Univ. Bucureşti, 1987, pp. 402-406.

(284) D. Dumitrescu, *Clasificare, sinteza conceptelor si reprezentarea cunoasterii*, Colocviul INFO IAŞI, Iaşi, 1987, pp. 305-308.

(285) D. Dumitrescu, *Hierarchical detection of linear clusters substructure.* Univ. Babeş-Bolyai, Fac. Math. Research Seminars Computer Science 2 (1986), pp. 21-28.

(286) D. Dumitrescu, *Numerical methods in fuzzy hierarchical pattern recognition.* Studia Univ. Babeş-Bolyai, Mathematica, Vol. XXXI, No. 4, 1986, pp. 31-36.

(287) D. Dumitrescu, *Logică şi clasificare cu mulţimi nuanţate*, Colocviul Naţional de Limbaje, Logică şi Lingvistică Matematică, Braşov, 1986, pp. 103-107.

(288) D. Dumitrescu, *O nouă clasă de algoritmi de clasificare în recunoaşterea formelor*, Colocviul Informatica şi aplicaţiile sale, Cluj- Napoca, 1986.

(289) D. Dumitrescu, *Asupra unei metode ierarhice în recunoaşterea formelor*, Simpozionul Naţional Teoria Sistemelor, Craiova, 1986, vol. 1, pp. 358-362.

(290) D. Dumitrescu, *Hierarchical classification in pattern recognization*, Colocviul de Informatică INFO IAŞI, Iaşi, 1985, vol. 2, pp. 638-643.

(291) D. Dumitrescu, *Fuzzy hierarchy for linear clusters*, Simpozionul Informatica şi aplicaţiile sale, Cluj-Napoca, 1985, pp. 74-77.

(292) D. Dumitrescu, *Clasificarea ierarchică în sistemele de regăsire a informaţiei*, Simpozionul Informatica şi aplicaţiile sale, Cluj-Napoca, 1985, pp. 83-84.

(293) D. Dumitrescu, M. Barbu, *Fuzzy entropy and processes*, Seminar Functional Equations, Approximation and Convexity, Cluj-Napoca, 1985, pp. 71-74.

(294) D. Dumitrescu, *Hierarchical classification with fuzzy sets.* Univ. of Cluj-Napoca Fac. Math. Res. Seminars Computer Science 5 (1984), pp. 36-55.

(295) D. Dumitrescu, *Recunoaşterea formelor în controlul sistemelor*, Simpozionul Naţional de Teoria Sistemelor, Craiova, 1984, pp. 368-372.

(296) D. Dumitrescu, *Fuzzy hierarchy in pattern recognition*, Coll. On Approximation and Optimization, Cluj-Napoca, 1984, pp. 75-82.

(297) D. Dumitrescu, *On fuzzy partitions*, Seminar on Functional Analysis, Approximation and Convexity, Cluj-Napoca, 1983, pp. 57-60.

(298) D. Dumitrescu, *On fuzzy partitions in cluster analysis*, Seminarul Th. Angheluţă, Cluj-Napoca, 1983, pp. 105-108.

(299) D. Dumitrescu, *Clasificare ierarhică*, Colocviul de Informatică InfoIaşi, Iaşi, 1983, pp. 349-355.

(300) D. Dumitrescu, *A method for cluster analysis*, Symp. Methods, Models and Tech. in Physics and related fields, Cluj-Napoca, 1983, pp. 164-165.

(301) D. Dumitrescu, *Entropie şi măsuri fuzzy*, Simpozionul Naţional Analiza Funcţională şi Aplicaţii, Craiova, 1983, pp. 389-396.

(302) D. Dumitrescu, *Asupra sistemelor nuanţate în recunoaşterea formelor*, Simpozionul Naţional de Teoria Sistemelor, Craiova, 1982, vol. 2, pp. 255-261.

(303) D. Dumitrescu, *Divisive hierarchical clustering*, Studia Univ. Babeş-Bolyai, Mathematica, Vol. XXVI, No. 1, 1981, pp. 24-30.

(304) D. Dumitrescu, *Strategii sociale - un model fuzzy*, Informatica pentru conducere, Cluj-Napoca, 1980.

(305) D. Dumitrescu, *Asupra sintezei în sistemele nuanţate*, Simpozionul Naţional de Teoria Sistemelor, Craiova, 1980, vol. 1, pp. 155-164.

(306) D. Dumitrescu, *Fuzzy correlation.* Studia Univ. Babeş-Bolyai, Mathematica, Vol XXIII, No. 1, 1978, pp. 41-44.

(307) D. Dumitrescu, *On some measures of nonfuzziness*, Studia Univ. Babeş-Bolyai, Mathematica, Vol XXIII, No. 1, 1978, pp. 45-49.

(308) D. Dumitrescu, *Definition of an information energy in fuzzy set theory*, Studia Univ. Babeş-Bolyai, Mathematica, Vol XXII, No. 1, 1977, pp. 57-59.

(309) Z. Gabos, D. Dumitrescu, *Sur la polarisation des bosones de masse de repos zero*, Studia Univ. Babeş-Bolyai, Physica, Vol. XVII, No. 2, 1972, pp. 33-37.

Department of Computer Science, Faculty of Mathematics and Computer Science, Babeş-Bolyai University, 1 M. Kogălniceanu, 400084 Cluj-Napoca, Romania

E-mail address: bparv@cs.ubbcluj.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LV**, Number 1, 2010

HILL-CLIMBING SEARCH IN EVOLUTIONARY MODELS FOR PROTEIN FOLDING SIMULATIONS

CAMELIA CHIRA

Abstract. Evolutionary algorithms and hill-climbing search models are investigated to address the protein structure prediction problem. This is a well-known NP-hard problem representing one of the most important and challenging problems in computational biology. The pull move operation is engaged as the main local search operator in several approaches to protein structure prediction. The considered approaches are: (i) a steepest-ascent hill-climbing search guided by pull move transformations, (ii) an evolutionary model with problem-specific crossover and pull move mutations, and (iii) an evolutionary algorithm based on hill-climbing search operators. Numerical experiments emphasize the advantages of the latter approach for several difficult protein benchmarks.

1. Introduction

Protein folding simulations aim to find minimum-energy protein structures starting from an initially unfolded chain of amino acids. The prediction of protein structures having minimum energies represents an NP-hard problem [1, 3]. The paper addresses this problem in the simplified hydrophobic-polar (HP) lattice model extensively engaged in computational experiments due to its simplicity [8], yet being able to generate significant results.

Several approaches to protein structure prediction based on evolutionary and/or hill-climbing search are investigated. The paper compares the perfromance of a pure hill-climbing search algorithm, a simple evolutionary algorithm and an evolutionary model based on hill-climbing search operators. The common feature of these approaches is the usage of pull move transformations [5] as the main local search operator. Pull move operations result in a single residue being moved diagonally causing the potential transition of

Received by the editors: December 5, 2009.

2010 *Mathematics Subject Classification.* 68T20.

1998 *CR Categories and Descriptors.* I.2.8 Computing Methodologies [**ARTIFICIAL INTELLIGENCE**]:Problem Solving, Control Methods, and Search – *Heuristic methods*.

Key words and phrases. evolutionary algorithms, hill-climbing search, protein structure prediction.

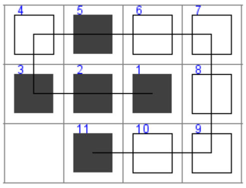


Figure 1. A protein configuration for sequence $SE = HHHPHPPPPPH$ in the square lattice having the energy value of -2. Black squares represent H residues and white squares are P residues.

connecting residues in the same direction in order to maintain a valid protein configuration [5]. Pull moves are engaged in different strategies in each model investigated.

Numerical experiments for bidimensional HP lattice protein sequences are carried out for all the models presented in the paper. Comparative results indicate a better performance of the evolutionary algorithm based on hill-climbing operators.

The paper is organised as follows: the protein structure prediction problem in the HP model is briefly described, pull move transformations are reviewed, the three models discussed in the paper are presented and numerical results and comparisons are given.

2. The Protein Structure Prediction Problem in the HP model

Simplified lattice protein models like the HP model [2] have become important tools for studying proteins being extremely useful in the initial approximation of the protein structure and in the investigation of protein folding dynamics.

In the HP model, a protein structure with n amino acids is viewed as a sequence $S = s_1...s_n$ where each residue $s_i, \forall i$ can be either H (hydrophobic or non-polar) or P (hydrophilic or polar). A valid protein configuration forms a self-avoiding path on a regular lattice with vertices labelled by amino acids. Figure 1 presents a configuration example for protein sequence $SE = HHHPHPPPPPH$ (black squares denote H residues and white squares represent P residues).

Two residues are considered topological neighbors if they are adjacent (either horizontally or vertically) in the lattice and not consecutive in the sequence (for example in Figure 1 the pair of residues labelled 2 and 5 form a H-H topological contact).

In the HP model, the energy associated to a protein conformation takes into account every pair of H residues which are topological neighbors. Every H-H topological contact contributes -1 to the energy function. The aim is to find the protein configuration having minimum energy. This solution will correspond to the protein configuration with the maximal number of H-H topological contacts.

The energy of the protein conformation presented in Figure 1 is -2 (given by H-H contacts $2-5$ and $2-11$).

3. Pull Moves in the HP Square Lattice Model

Pull move transformations have been introduced in [5] as a local search strategy for the bidimensional HP model. Incorporated in a tabu search algorithm, pull moves have been able to detect new lowest energy configurations for large HP sequences having 85 and 100 amino-acids [5].

A pull move operation starts by moving a single residue diagonally to an available location. A valid configuration is maintained by pulling the chain along the same direction (not necessarily until the end of the chain is reached - a valid conformation can potentially be obtained sooner).

A pull move transformation can be applied at a given position i from the considered HP sequence.

Let (x_i, y_i) be the coordinates in the square lattice of residue i at time t. Let L denote a free location diagonally adjacent to (x_i, y_i) and adjacent (either horizontally or vertically) to (x_{i+1}, y_{i+1}). Location C denotes the fourth corner of the square formed by the three locations: L, (x_i, y_i) and (x_{i+1}, y_{i+1}). A pull move is possible if location C is free or equals (x_{i-1}, y_{i-1}). In the latter case, the pull move transformation consists of moving the residue from location (x_i, y_i) to location L. In the case that C is a free location, the first step is to move residue from position i to location L and the residue from position $(i-1)$ to location C. The pull move transformation continues by moving all residues from $(i-2)$ down to 1 two locations up the chain until a valid configuration is reached.

Figure 2 presents an example of a pull move transformation for HP sequence $SE = HHHPHPPPPPH$. The pull move is applied for residue H at position $i = 3$ for which a free location L horizontally adjacent to residue $i+1$ (between residues 4 and 10 in Figure 2.a) is identified. Location C (the location between residues 3 and 11 in Figure 2.a) is free in this example and therefore the pull move will cause moving the residue 3 to location L and

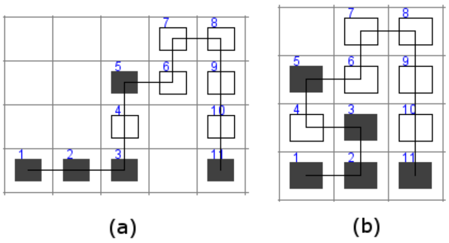


FIGURE 2. Pull move transformation for HP sequence $HHHPHPPPPPH$ at position 3

residue 2 to location C. The remaining residue 1 (only one in this example) is moved up the chain two positions (see Figure 2.b).

Lesh et al [5] prove that the class of pull moves is reversible and complete.

4. Evolutionary and Hill-Climbing Search

This paper investigates the performance of the following three models in solving the protein structure prediction problem: (i) a hill-climbing search algorithm, (ii) a simple evolutionary algorithm based on dynamic crossover and pull moves as mutation, and (iii) an evolutionary model based on hill-climbing search operators.

All three proposed models use the same problem representation (commonly engaged in genetic algorithms for this problem [7, 4]). A protein configuration (problem solution or chromosome) is encoded using an internal coordinates representation. For a protein HP sequence with n residues $S = s_1...s_n$, the chromosome length is $n - 1$ and each position in the chromosome encodes the direction $L(Left)$, $U(Up)$, $R(Right)$ or $D(Down)$ towards the location of the current residue relative to the previous one. For the working example in Figure 1 the chromosome is $LLURRRDDLL$.

The fitness function used corresponds to the energy value of the protein configuration (as given in Section 2).

4.1. **Hill-Climbing Search Model.**

A simple hill-climbing search model based on pull moves is described (see *HC* scheme below). The algorithm applies pull move transformations for a protein configuration each iteration within a steepest ascent hill-climbing procedure.

Hill-climbing search starts by randomly generating one valid configuration for the given HP sequence and setting it as the *current_hilltop*. Pull moves are applied at each position $i, i = 1, ..., n$ (where n is the length of the HP

```
Hill-Climbing Search based on Pull Moves (HC)
Set current_hilltop to a randomly generated configuration rand_c
Set best_c to current_hilltop
Add best_c to hilltop_array
while (maximum number of hc iterations not reached) do
  for $i$=1 to $n$ do
    Generate new configuration $c_i$ by applying a
    pull move transformation at position $i$ in current_hilltop
    if ($c_i$ has better fitness than best_c) then
      Set best_c to $c_i$
    end if
  end for
  if (better configuration best_c found) then
    Set current_hilltop to best_c
  else
    Save best_c in hilltop_array
    Set rand_c to a new randomly generated configuration
    Set current_hilltop and best_c to rand_c
  end if
end while
Return best solution from hilltop_array
```

sequence) resulting in the generation of n new configurations. If any of them has a better fitness value than the *current_hilltop* it replaces the latter one. If no improvement is achieved and the maximum number of hill-climbing iterations has not been reached, the *current_hilltop* is saved in a list of hilltops and then reinitialized with a new randomly generated configuration.

4.2. **Evolutionary Algorithm with Pull Moves.** In the evolutionary approach (see *EA* scheme) to the protein structure prediction problem, a chromosome represents a possible protein configuration for a given HP sequence.

The population size is fixed and offspring are asynchronously inserted in the population replacing the worst parent within the same generation.

For the recombination of genetic material, a one-point dynamic crossover operator is specified. Given two parent chromosomes p_1 and p_2 and a randomly generated cut point χ, two offspring are created as follows. The genes before the crossover point χ are copied from one parent. The second part of the offspring is taken from the other parent in such a way that a valid configuration is maintained. This means that each position j, $j = \chi, .., n-1$ is copied from

Evolutionary Algorithm with Pull Moves (EA)

```
t = 0
Generate $P(t)$ with *pop_size* individuals randomly
**while** (maximum number of generations not reached) **do**
  **for** each individual $i$ in $P(t)$ **do**
    *Apply crossover with probability p_c*
      Select mate $j$ using binary tournament selection
      Generate random cut point $\chi$
      Generate offspring $o$ = crossover($i$, $j$, $\chi$)
      **if** $o$ has better fitness than $i$ or $j$ **then**
        Replace $worst(i, j)$ with $o$ in $P(t)$
        Replace random individual from $P(t)$ with $mutation(o)$
      **end if**
    *Apply pull move mutation with probability p_m*
      Generate random pull move position $k$
      Generate mutated chromosome $m$ by pull move in $i$ at position $k$
      **if** $m$ has better fitness than $i$ **then**
        Replace $i$ with $m$ in $P(t)$
      **end if**
  **end for**
  $t = t + 1$
**end while**
```

the second parent and checked for potential collisions with positions 0 to $j-1$ already copied in the chromosome current substring. If a conflict arises then a random direction leading to a valid position is selected and used in the offspring. The best of the two offspring generated replaces the worst parent if a better fitness was generated.

Pull move transformation is engaged as the mutation operator. For each individual selected for mutation, a random position is generated and a pull move transformation is applied at that position. The new mutated chromosome resulted replaces the parent if it has a better fitness value.

Furthermore, the offspring generated by crossover is transformed using pull move mutation and replaces an individual from the current population at random. This feature facilitates the diversification of genetic material and is also engaged in the evolutionary model with hill-climbing operators (presented in the following subsection).

4.3. **Evolutionary Model based on Hill-Climbing Operators.** In the third model investigated (see *EA-HCO* scheme), a population of configurations

Evolutionary Algorithm with Hill-Climbing Operators (EA-HCO)

```
$t = 0$
Generate $P(t)$ with $pop\_size$ individuals randomly
while (maximum number of generations not reached) do
  Randomly select $p_1$ and $p_2$ from $P(t)$
  while (maximum number of $hc$ iterations not reached) do
    for $k = 1$ to $n$ do
      Generate random cut point $\chi$
      Generate offspring $o_k$ = crossover($i$, $j$, $\chi$)
    end for
    Set $o$ to $best(o_k)$, $k = 1..n$
    if $o$ has better fitness than $p_1$ or $p_2$ then
      Replace $worst(p_1, p_2)$ - also in $P(t)$ - with $o$
      Replace random individual from $P(t)$ with $mutation(o)$
    else
      Set $p_1$ and $p_2$ to new randomly selected individuals from $P(t)$
    end if
  end while
  Hill-climbing pull move mutation for $hc$ iterations
  $t = t + 1$
end while
```

is evolved by hill-climbing crossover and mutation. The evolutionary algorithm uses the same genetic operators described in section 4.2 (dynamic crossover and pull move mutation) with the difference that they are applied now in a steepest ascent hill-climbing manner.

Crossover is engaged for randomly selected pairs of individuals in a hill-climbing mode [6]. The best-fitted offspring replaces the worst parent within the same generation. If no better offspring is identified, both parents are replaced by new randomly selected chromosomes. The process continues until the maximum number of hill-climbing iterations is reached.

Mutation implements a steepest ascent hill-climbing procedure using the pull move operation. This process is able to generate a variable number of new individuals which replace parents within the same generation (if they have a better fitness value).

The hill-climbing pull move mutation step works in similar way with the procedure described in section 4.1 for hill-climbing search except that new

Table 1. Bidimensional HP instances used in experiments

Inst.	Size	Sequence	E^*
S1	20	1H 1P 1H 2P 2H 1P 1H 2P 1H 1P 2H 2P 1H 1P 1H	-9
S2	24	2H 2P 1H 2P 1H 2P 1H 2P 1H 2P 1H 2P 1H 2P 2H	-9
S3	25	2P 1H 2P 2H 4P 2H 4P 2H 4P 2H	-8
S4	36	3P 2H 2P 2H 5P 7H 2P 2H 4P 2H 2P 1H 2P	-14
S5	48	2P 1H 2P 2H 2P 2H 5P 10H 6P 2H 2P 2H 2P 1H 2P 5H	-23
S6	50	2H 1P 1H 1P 1H 1P 1H 1P 4H 1P 1H 3P 1H 3P 1H 4P 1H 3P 1H 3P 1H 1P 4H 1P 1H 1P 1H 1P 1H 1P 1H 1H	-21

individuals required for mutation are not generated anew but they are selected at random from the current population.

The number of individuals undergoing recombination and mutation each generation is dynamic as the hill-climbing operators modify the same structure until no further improvement can be generated and then continue with new individuals. An explicit selection for the next generation is not required as offspring are asynchronously inserted in the population as soon as they are created.

5. Numerical Experiments

The three models presented in the previous section are engaged in a set of numerical experiments for the bidimensional HP protein sequences presented in Table 1 (the known energy denoted by E^* is given for each instance).

The following parameter setting is engaged in the experiments:

- For the hill-climbing search model based on pull moves (refered to as *HC*), the number of *hc* iterations is 10000.
- For the evolutionary algorithm based on pull moves (refered to as *EA*), the population size is 100, the number of generations is 300, the crossover probability is 0.8 and the mutation probability is 0.2.
- For the evolutionary algorithm based on hill-climbing operators (refered to as *EA-HCO*), the population size is 100, the number of generations is 300, the offspring number in crossover hill-climbing is 50 and the number of hill-climbing iterations *hc* for both crossover and mutation is set to 100.

The initial population for the evolutionary algorithms contains randomly generated chromosomes representing valid configurations (each chromosome is

Table 2. Comparison of results achieved by the three investigated models for the HP problem

Inst.	Size	E^*	HC	EA	EA-HCO
S1	20	-9	**-9**	**-9**	**-9**
S2	24	-9	-8	**-9**	**-9**
S3	25	-8	-6	**-8**	**-8**
S4	36	-14	-10	-13	**-14**
S5	48	-23	-17	-20	**-23**
S6	50	-21	-15	-19	**-21**

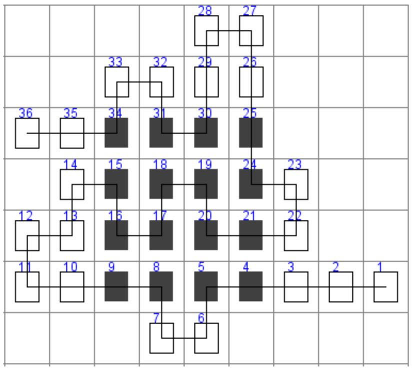


Figure 3. One of the protein configuration detected by *EA-HCO* for sequence S4 = 3P 2H 2P 2H 5P 7H 2P 2H 4P 2H 2P 1H 2P having the best-known energy value of -14

iteratively generated in a random manner until a conformation free of collisions in the HP square lattice model is found).

Table 2 presents comparative results for the HP sequences considered (the results of the best run out of 25 are reported). The known optimum energy E^* for each problem instance and the energy values detected by the three investigated models *HC*, *EA* and *EA-HCO* are given in separate columns.

Evolutionary search based on hill-climbing operators is able to detect optimal solutions for all HP instances considered. Figure 3 shows one of the optimal protein configurations detected by *EA-HCO* for instance $S4$.

Table 3. Percentage of succesful runs and the average generation number producing the best energy value for the *EA* and *EA-HCO* models

Inst.	EA			EA-HCO		
	E	Succ. Runs	Avg. Gen.	E	Succ. Runs	Avg. Gen.
S1	-9	92%	126.74	-9	100%	11.68
S2	-9	76%	153.26	-9	100%	16.20
S3	-8	64%	161.00	-8	100%	27.32
S4	-13	12%	235.00	-14	64%	141.68
S5	-20	4%	277.00	-23	8%	250.50
S6	-19	12%	221.33	-21	56%	182.07

Table 2 indicates that the evolutionary model based on hill-climbing search operators outperforms the other two approaches investigated. It can be observed that all three models are able to detect the best-known solution for the first sequence considered $S1$ having a length of 20. As the size of the protein sequence grows (and therefore the complexity of the search space increases), the power of hill-climbing search and evolutionary search alone gets lower.

Hill-climbing search (model *HC* in table 2) results are far from the optimum for the sequences $S2$ to $S6$ with lengths from 24 to 50. Evolutionary search (model *EA* in table 2) is able to identify optimum solutions for sequences $S1$, $S2$ and $S3$ but fails to guide the search towards the optimum for higher-size sequences. This problem is succesfullly overcome by the same operators applied in a hill-climbing manner in model *EA-HCO* - able to detect optimum energy values for all sequences considered.

The number of succesful runs (those in which the optimum energy has been detected) out of the 25 runs considered is studied in a further comparison between the *EA* and *EA-HCO* models. Moreover, the generation number producing the best energy value is recorded each run. Table 3 shows the results obtained in the following mode: for each HP sequence, the procentage of successful runs and the average generation number detecting an optimum (or best energy value) are given for the two evolutionary algorithms compared. It should be noted that table 3 considers succsesful runs those in which the best energy was obtained if the optimum was not found. This is the case of *EA* results for sequences $S4$, $S5$ and $S6$.

Table 3 clearly emphasizes the better performance of the *EA-HCO* model compared to *EA*. The percentage of successful runs is higher for each HP instance when the evolutionary algorithm based on hill-climbing search is used.

Furthermore, *EA-HCO* is able to detect the optimal solution in all 25 runs for several protein sequences. The *EA-HCO* model also outperforms *EA* with regard to the average generation in which the best energy configuration is identified. Hill-climbing search operators integrated in an evolutionary model are able to detect optimal solutions in the early stages of the search process. More generations are required as the protein sequence size increases. Even for such sequences, the *EA-HCO* is able to find the optimum solution earlier in the search compared to the stage where the *EA* model finds the best solution (not the optimum as *EA* fails to find optimum solutions for sequences $S4$, $S5$ and $S6$).

Numerical results and comparisons clearly emphasize the benefits of hill-climbing search operators integrated in evolutionary models compared to either hill-climbing or evolutionary search for protein structure prediction.

6. Conclusions and Future Work

Hill-climbing and evolutionary search models are studied for solving the protein structure prediction problem. The results presented emphasize the benefits of integrating hill-climbing search operators in an evolutionary algorithm.

Future work refers to the investigation of *EA-HCO* performance for other protein sequences and the extension of the proposed model to include other search operators.

References

[1] Crescenzi, P., Goldman, D., Papadimitriou, C. H., Piccolboni, A., Yannakakis, M., *On the Complexity of Protein Folding*, Journal of Computational Biology, 50 (1998), 423–466.

[2] Dill, K.A., *Theory for the folding and stability of globular proteins*, Biochemistry, 24 (1985), 6, 1501–1509.

[3] Hart, W., Newman, A., *Protein Structure Prediction with Lattice Models*, Handbook of Computational Molecular Biology, Chapman & Hall CRC Computer and Information Science Series, 2006.

[4] Khimasia, M.M., Coveney, P.V., *Protein structure prediction as a hard optimization problem: the genetic algorithm approach*, Molecular Simulation, 19 1997), 205–226.

[5] Lesh, N., Mitzenmacher, M., Whitesides, S., *A complete and effective move set for simplified protein folding*, in RECOMB '03: Proceedings of the seventh annual international conference on Research in computational molecular biology, ACM, 188–195 (2003).

[6] Lozano, M., Herrera, F., Krasnogor, N., Molina, D., *Real-coded memetic algorithms with crossover hill-climbing*, Evol. Comput., 12 (2004), 3, MIT Press, 273–302.

[7] Unger, R., Moult, J., *Genetic algorithms for protein folding simulations*, J. Molec. Biol., 231 (1993), 75–81.

[8] Zhao, X., *Advances on protein folding simulations based on the lattice HP models with natural computing*, Appl. Soft Comput., 8 (2008), 2, 1029–1040.

Department of Computer Science, Babes-Bolyai University, Kogalniceanu 1, 400084 Cluj-Napoca, Romania

E-mail address: cchira@cs.ubbcluj.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LV**, Number 1, 2010

AN AUTONOMOUS APPROACH TO WHEEL CHANGING PROBLEM

LIVIU ŞTIRB, ZSUZSANNA MARIAN, AND MIHAI OLTEAN

Abstract. We describe a self-repairing robotic car capable of changing its wheels automatically. The robot is constructed from a Lego NXT kit and a Lynx Arm kit which are coordinated from a PC. The difficulty consists in assembling the wheel on its shaft with a high precision which is not possible with the Lego components. This was solved by creating an ensemble made from both wheel and its shaft. The entire ensemble is replaced instead of the wheel alone. We have performed several experiments which show the effectiveness of our approach.

1. Introduction

Changing the broken wheel of a car is an operation that most people try to avoid. When this happens the drivers prefer to call a service company to perform this task. They do that because changing the wheel can generate several problems, some of them being listed below:

- it is time-consuming,
- requires some tools,
- requires some knowledge on how to do it,
- it can make both the hands and clothes very dirty and badly smelling,
- in some cases it can generate injuries to fingers.

For solving these problems we have designed a robotic car which is able to change its wheels autonomously with no interference from humans. The robot consists of 2 main parts: the car itself and an arm placed above the car, which will change the broken wheel. The car is constructed from a Lego NXT kit, while the arm is a standard Lynx arm with 6 degrees of freedom. The difficulty of the problem is due to the low power and low precision of the involved components: it is almost impossible to insert an axel into a motor. Using the Lego and Lynx kits one simply cannot do that.

2010 *Mathematics Subject Classification.* 68T40, 93C85, 70B15.

1998 *CR Categories and Descriptors.* I.2.9 [**Computing Methodologies**]: Artificial Intelligence: Robotics.

Key words and phrases. robot, car, wheel changing, self-repairing, Lego NXT, Lynx arm.

In order to solve this problem we propose a more complex device. Instead of changing the wheel, we change an ensemble composed from wheel and its axel. We also design a mechanism where the wheel and its axel can be inserted without needing a high precision.

The paper is structured as follows: Section 2 surveys some material in the field of self-repairing and self-replicating robots. Section 3 deeply describes the proposed approach. It describes the hardware, the software and all the steps required for changing the wheel. Section 4 discusses the numerical experiments and their outcome. In section 5 we enumerate some of the strong and weak points of our idea. Finally section 6 concludes our paper.

2. Related Work

While there are a lot of articles describing self-replicating, self-reconfiguring, self-constructing and self-assembling robots, there are only a few that present self-repairing robots. One of the best known is presented in [6] and consists of a robotic chair that is capable of falling apart and then reassembling itself again. Building the robot required two years of work. It is made of 14 motors, two gearboxes and many other pieces, it has its *brain* in the seat, and uses a sophisticated algorithm to find its pieces (the four legs and the back of the chair) and eventually re-assembles itself and stands up. The chair, which has first been presented at the IdeaCity conference in Toronto, won the acclaim of the cyberart sophisticates at the 2006 ARS Electronica conference in Linz, Austria [7].

In 2009 researchers from the Pennsylvania University built a modular robot that is capable of self-assembling, after it has been kicked apart [9]. This robot is made of 5 clusters, each cluster being made of 4 modules and a camera, and is capable of moving by itself. Each module has 4 circuit boards and a motor. Some of the modules are connected to the others through screws, while others through magnets. When the robot is kicked, parts connected through magnets separate from each other. Each cluster contains a LED, which displays a custom blinking pattern, and a camera. These two help the clusters to locate and identify each other. Clusters can turn and crawl which helps them to find each other and connect through the magnets. The weak point of this approach is given by the fact that the robot is composed from identical modules which are less likely for real-world approaches.

In [4] a set of modular robot cubes that are capable of self-replication (the capability of constructing a detached, functional copy of itself, which will also be capable of replication) is presented. This robot is made of 10-cm module cubes which have electromagnets that selectively weaken and strengthen the connections, determining where the structure breaks or joins [4]. These cubes are split into halves along the (1,1,1) plan and one half can swivel relative to the other in 120 degrees increments. These cubes are powered through the baseplate and transfer power and data through their faces. This can be considered a disadvantage, because it means that the modules are not functional outside the laboratory. Taking cubes from specific locations a four-module robot was capable to replicate itself in 2.5 minutes, while a three-module robot needed only about a minute [4]. Although this robot is capable of self-replication not self-repair, such systems can be considered self-repairing, because they might be capable of changing a module that is not working properly with a good one. The most important aspect of this approach is the scalability: the replicated robot can contain any number of cubes.

Another interesting self-repairing robot is described in [5]: this robot uses an autonomous and continuous process of self-modeling, thus observing changes in its own morphology and synthesizing new behavior. This process is composed of three algorithmic components, executed continuously by the robot while moving or resting: modeling, testing and prediction. First an arbitrary motor action is performed by the robot and its result is recorded. Using this sensory-actuation casual relationship the model synthesis component creates a set of 15 candidate self-models, and determines the next action, that would give the most information about the robot. After 16 cycles of this, the most accurate model is used to generate the new, desired behavior. The authors present how this approach was tested using a robot with four legs and eight motorized joints, eight joint angle sensors, and two tilt sensors. This robot was allowed to move for a while, after which it suffered some damage (usually a part of one of its legs was removed). Starting from the best self-model from before the accident, the robot was capable of developing a new behavior and it started moving again.

3. Proposed approach

In this section we deeply describe the proposed approach. We start by presenting the involved hardware. Then, we describe the proposed software. Finally, we put all together and we give details on how the robot works.

3.1. **Hardware.** The car is constructed with pieces from the Lego Mindstorm NXT kit. The mechanism that changes the wheel is a Lynx arm with 6-degrees of freedom.

Three servo motors are attached to the Lego brick. Their purpose is:

- The first one is at the back of the robot and is used to move it, using two bigger wheels placed at the back side of the robot. In the current stage of the project these wheels cannot be changed, but we are working on new version where these changes are possible.
- The second motor is used to lift up the robot a little, when the wheel is changed (just like one lifts the car when he changes the wheel). Some

Lego pieces are attached to it, which in normal state (when the robot is moving) do not touch the ground. To lift the car the motor is rotated with 80 degrees and these attached pieces get to touch the ground, moreover lift the front part of the robot a little, so the front wheels do not (or only slightly) touch the ground (see Figure 1).

- The third motor is located at the left side of the robot and is used to move a Lego piece called cross axel, which opens or closes the part that holds the wheel (see Figures 2). If it is open, the left wheel structure can be taken and replaced by another similar one. In order to be able to change both the left and the right wheels at the front of the car, we would need another motor, to work in the same way at the right part of the robot, but that would imply another brick, since the maximum number of motors that can be used with a brick is 3.

The robotic car has a Lynx 6 robotic arm [10] on its top, which is used to change the wheel. This arm has a base, shoulder, elbow, wrist, and a functional gripper. These are important, because the arm has to do delicate movements to handle the wheel.

Although the arm is capable of a variety of movements, inserting a cross axel into a wheel - the most common way to build wheels - needs precision and force the arm is incapable of. This is why we decided to build a special wheel structure that the arm can handle easier. This structure is made of a cross axel, having the wheel at one end and a wedge belt wheel at the other end (see Figure 3). Such a wheel structure can easily be moved by the arm, by gripping the wheel.

It is worth mentioning that in case of real cars, changing the wheel does not require so much precision. You need precision to unscrew the bolts from the wheel, to put the spare wheel to the exact position where it is needed and to screw the bolts again. The problem of working with bolts could be solved by using some special mechanism, which makes them move (screw and unscrew) autonomously and they could somehow be attached to the wheel. However, you would still need an arm to take the flat wheel and put there the spare one.

Finally, the robot has a small platform built of Lego pieces in front of the arm, where the spare wheel is located.

3.2. **Software.** The software of the robot is written in the C++ programming language, using the NXT++ library (available at [11]), an interface that allows controlling Lego Mindstorm robots through a USB or Bluetooth connection. Our choice was a connection through USB.

Moving the motors can be done in two different ways: using the *SetForward* method, which requires a connection to the brick, a motor and a value

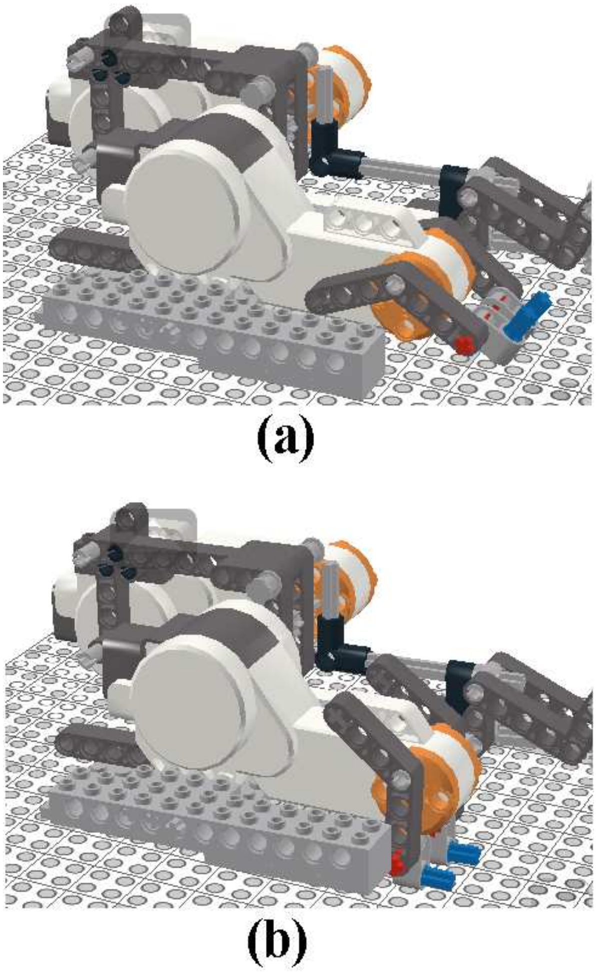


Figure 1. The mechanism for lifting the car. It consists of a motor which has attached a piece whose size is larger than the distance from ground to the chassis. When the car is down (a) it can move. When the car is lifted (b) it cannot be moved, but a robotic arm can replace its wheel.

between 0 and 100, the power. Using this method, the motor starts moving and continues to do so, until the *Stop* method is called for that motor. The second method is using the *GoTo* method, which requires a connection, a motor, the power, the degree to which the motor should be rotated and a Boolean value whether after the rotation braking should be used or not. When

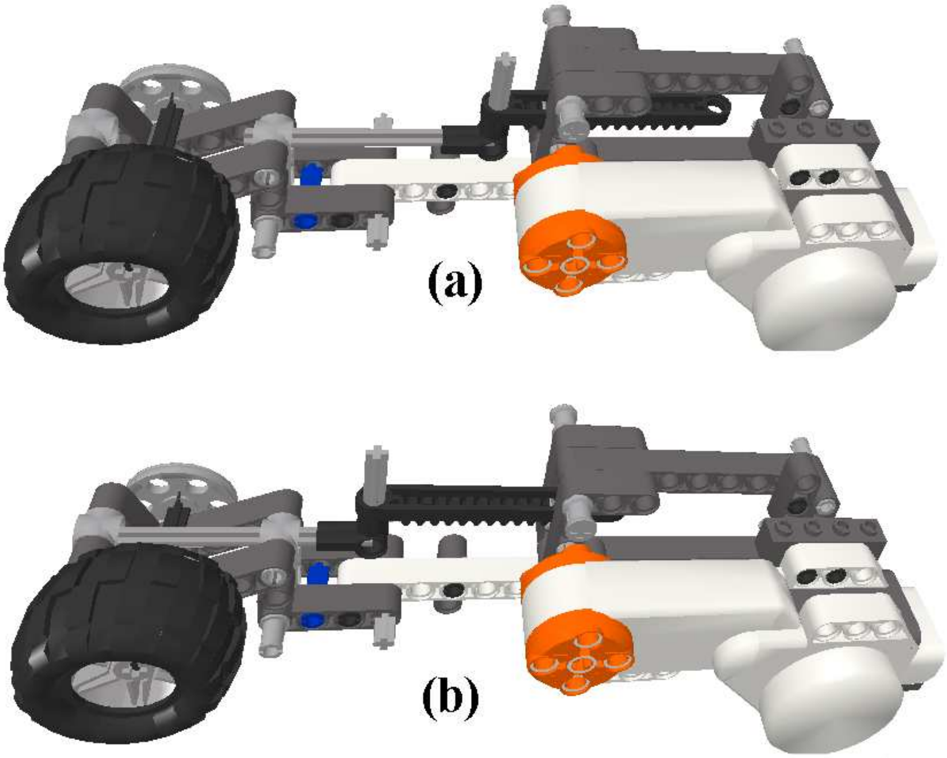


Figure 2. The mechanism for locking a wheel on the car. When the wheel is free (a) a robotic arm can replace it with a spare one. In this case the car cannot move because it risks losing its freed wheel. When the wheel is locked (b) the car can move.

the motor has turned to that degree, it stops, there is no need for the *Stop* method.

To control the Lynx arm we used again some simple methods to send data to and read data from the controller. Data sent to the arm is a string, starting with the "#" sign, followed by the number of the motor to be moved (from 0 to 5), a "P" followed by the pulse width in microseconds (500-2500), this shows how much the motor should move, and, finally, an "S" and the movement speed, which is optional, but here we used it to make slow movements.

3.3. **Detailed description of the procedure.** The robot goes forward for few seconds and then it stops to change the tire.

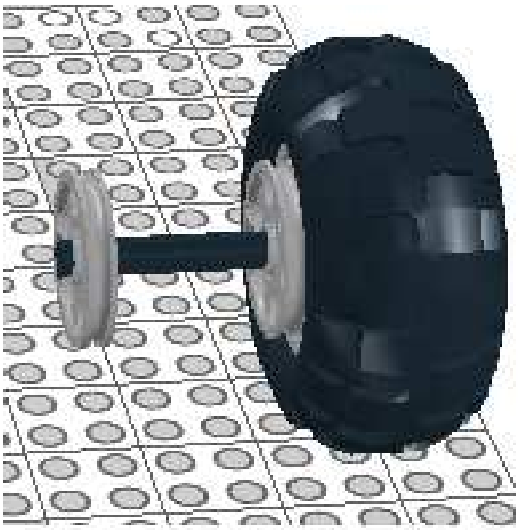


Figure 3. The structure of the wheel.

The first step is to reset the arm into a default position. This operation is not compulsory, but this helps us to simplify the movements of the arm.

Then the following steps are performed:

- The front part of the robot is lifted a little turning the second motor to 80 degrees as described in 3.1.
- This is followed by moving the axel to open the place where the wheel structure is located.
- Now the arm goes to the position where the spare tire is, gripes it, and slowly puts it down to the ground in front of the robot. It is important to move slowly, especially when the arm holds the wheel, because fast movements shake the robot, and the wheel might fall from the gripper. It is also important to put it down slowly, because the wheel, being round, might move from the place where it is placed, which is bad, because the robot has no camera to detect the wheel's position, so it looks for it where it was left. Fortunately, if the arm moves slowly, the wheel usually remains at the position where it was put to.
- The arm goes now and takes the left wheel structure, lifts it, and puts it to the platform where the spare tire was taken from.
- Then it goes back to the position where the spare tire was left, grips it and puts it to the place where it is needed. This is the most complicated movement, it is made of 11 fine movements of the arm to find the exact position.

- After this, the arm goes back to the default position. This position is the same as the one at the beginning of the algorithm.
- The axel is moved back, to close the place where the wheel is, the robot is put back on the ground and it moves forward for few more seconds, to see that it works the same way as at the beginning.

4. Experiments

No special environment is required for performing the experiments. The robot would perform the same on any type of hard floor.

In order to test the effectiveness of the procedure it was run 30 times. Out of these 30 experiments, 21 were successful. The most important reasons for failure are:

- the spare wheel put at the ground by the arm moved from the place where it was left. Thus, when trying to grip it again, the arm did not find it. It is worth mentioning, that if the wheel movement is a small one, than it is still possible that the gripper can get a hold of the wheel (as it can be seen on the movie that is mentioned at 6). This problem could be solved by using a camera to check if the arm is at the right position.
- when the axel was moved to lock it, it hit the axel of the wheel structure, and it was stopped by it. This led to the fall of the wheel structure when the robot started to move again. Although, we do not know why this happened, we think that making the front of the axel sharp would solve this problem.

The average time required for the arm to change the wheel is 2 minutes.

5. Strengths and weaknesses

In this section we describe the weak and strong points of our approach.

5.1. Strengths.

Some advantages of the proposed approach are:

- The wheel is changed autonomously with no interference from the driver,
- The entire procedure is very simple and requires few minutes to perform,
- The algorithm is hard-coded and no complex pattern recognition algorithms [1] are required. Using a webcam and a vision algorithm for detecting the wheels would complicate the problem significantly [2, 3]. Fortunately in our case we do not need one.

5.2. **Weaknesses.** Some of the most important limitations of the system are:

- Cannot changes all its wheels. Changing the right wheel is the same as the case of the left wheel, but this operation requires an extra brick which will coordinate the fourth motor. Changing the other wheels is more complicated because we need a special mechanism which transmits the torque from motor to wheel.
- Cannot changes its wheels while driving. Doing this is simple if we change the wheels not attached to motors. In this case we simply need an extra place where to keep the flat wheel. Also, the mechanism which lifts the car should have a small wheel at its end. However, changing the wheels attached to the motors is very complicated because when the wheel is removed, no more torque is transmitted and the car would stop.
- The car requires an extra arm which is less likely to be placed on the real-world cars. The arm cannot be removed completely, but a smaller (specially designed) one could do the same job as the current one. Another option is to replace the wheel only in a specialized garage, but in this case the car should be able to run many kilometers with the tire flat. The good news is that there is a technology used by the GoodYear company, called RunOnFlat [12], which is based on the concept of reinforced side walls inside the tire, which keep the tire on rim and succeed in carrying the weight of the car for up to 80 km after a puncture.
- The entire procedure is currently run on an external laptop which is connected through cable to the robot. Placing the laptop on the robot is not viable. There are 2 alternatives here: one is to set a wireless communication between the robot and the laptop or to use a gumstix which is very small and can be placed inside the robot.
- A wheel from a real-world car has some brakes attached to it. In this case changing the wheel requires to detach the breaks, which is induces more complexity to the system.

We are working to fix all these limitations in a future version of the robot.

6. Conclusions and further work

Here we have described a self-repairing autonomous car which has the ability to change its wheels automatically.

The further efforts will be spent in the following directions:

- changing the wheel for motorized wheels,
- changing the wheels while driving,
- replacing other components of the car.

As recently the use of autonomous robots to ease the job of people in cars increased - for example in the DARPA Urban challenge [13] autonomous vehicles capable of driving without human help in traffic, performing complex maneuvers such as passing, parking and handling intersection are competing - the idea behind our robot seems to be a promising one, even if there still need to be different refinements (as mentioned at the above section at the disadvantages) to use this robot in real life.

Acknowledgement

A movie presenting an experiment can be seen here:
`http://www.youtube.com/watch?v=f1iHi8Pgb90`

This work was supported from grant IDEI-543/2007 from CNCSIS, Romania.

Mihai Oltean acknowledges the fruitful discussions on similar topics with the undergraduate students from the Intelligent Robots class of Faculty of Mathematics and Computer Science, Babeş-Bolyai University, Cluj-Napoca.

References

[1] Jain AK., Duin RPW., Mao J., Statistical pattern recognition: A review, IEEE Transactions on pattern analysis and machine intelligence, Vol 22, Issue 1, pp. 4-37, 2000
[2] Jain R., Kasturi R., Schunck B.G., Machine Vision, McGraw-Hill, 1995
[3] Forsyth DA., Ponce J., Computer vision: a modern approach, Prentice Hall, 2002
[4] Zykov V., Mytilinaios E., Adams B., Lipson H., Self-reproducing machines, Nature Vol. 435 No. 7038, pp. 163-164, 2005
[5] Bongard J., Zykov V., Lipson H., Resilient Machines Through Continuous Self-Modeling, Science Vol. 314. no. 5802, pp. 1118 - 1121, 2006
[6] Robotic Chair: www.news.cornell.edustoriesOct06robotic.chair.aj.html
[7] Robotic Chair: www.raffaello.nameRoboticSculpture.html
[8] Gumstix: Way Small Computing, www.gumstix.com
[9] Modular Robot: www.nytimes.cominteractive20090727science20090721-modular-graphic.html
[10] Lynx Arm: www.lynxmotion.comCategory.aspx?CategoryID=25
[11] NXT++: nxtpp.clustur.com
[12] RunOnFlat: eu.goodyear.comhome_entiresrunonflat
[13] DARPA Urban Challenge: http:www.darpa.milgrandchallengeindex.asp

Department of Computer Science, Faculty of Mathematics and Computer Science, Babeş-Bolyai University, Kogălniceanu 1, Cluj-Napoca, 400084, Romania

E-mail address: `sdsd0092@scs.ubbcluj.ro`
E-mail address: `mzsi0142@scs.ubbcluj.ro`
E-mail address: `moltean@cs.ubbcluj.ro`

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LV**, Number 1, 2010

ABOUT SELECTING THE "BEST" NASH EQUILIBRIUM

RODICA IOANA LUNG

Abstract. Most games simulating real-wold situations present multiple Nash equilibria. The problem of selecting one equilibrium is tackled using a generative relation for Nash equilibria. The equilibria ascending most strategies from a randomly generated population of strategies can be considered. Numerical examples are used to illustrate the method.

1. Introduction

The problem of selecting one equilibrium of a normal form game has been addressed in the literature in different ways. According to [1] there are three main approaches to deal with multiple Nash equilibria.

One is to introduce an equilibrium selection mechanism that specifies which equilibrium is picked up. Examples include random equilibrium selection, in [4], and the selection of an extremal equilibrium, as in [7].

The second approach is to restrict attention to a particular class of games, such as entry games, and search for an estimator which allows for identification of payoff parameters even if there are multiple equilibria. For example the models in [5, 6] and [3] study situations in which the number of firms is unique even though there may be multiple Nash equilibria. They propose estimators in which the number of firms, rather than the entry decisions of individual agents, is treated as the dependent variable.

A third method [12] Tamer, uses bounds to estimate an entry model. The bounds are derived from the necessary conditions for pure strategy Nash equilibria, which say that the entry decision of one agent must be a best response to the entry decisions of other agents. In [2] Berry and Reiss survey the econometric analysis of discrete games.

In this work a new selection method based on a generative relation for Nash equilibria for normal form games is proposed. The Nash ascendancy

Received by the editors: January 10, 2010.

2010 *Mathematics Subject Classification.* 91A10, 91A40.

1998 *CR Categories and Descriptors.* I.2.1 [**Computing methodologies**]: Artificial Intelligence– *Games*.

Key words and phrases. Noncooperative games, Nash ascendancy relation, Equilibria selection.

[8] relation can be used to compute Nash equilibria using Natural Computing methods such as Evolutionary Algorithms. In the case of multiple equilibria it can also be used to differentiate between them by determining which one ascends more strategies from a randomly generated population of strategies.

2. Nash ascendancy relation

A finite strategic game is defined by $\Gamma = ((N, S_i, u_i), i = \overline{1, n})$ where:

- N represents the set of players, $N = \{1, ..., n\}$, n is the number of players;
- for each player $i \in N$, S_i represents the set of actions available to him, $S_i = \{s_{i_1}, s_{i_2}, ..., s_{i_{m_i}}\}$ where m_i represents the number of strategies available to player i and $S = S_1 \times S_2 \times ... \times S_N$ is the set of all possible situations of the game;
- for each player $i \in N$, $u_i : S \to \mathbb{R}$ represents the payoff function.

Denote by (s_{i_j}, s^*_{-i}) the strategy profile obtained from s^* by replacing the strategy of player i with s_{i_j} i.e.

$$(s_{i_j}, s^*_{-i}) = (s^*_1, s^*_2, ..., s^*_{i-1}, s_{i_j}, s^*_{i+1}, ..., s^*_1).$$

The most common concept of solution for a non cooperative game is the concept of Nash equilibrium [9, 10]. A collective strategy $s \in S$ for the game Γ represents a Nash equilibrium if no player has anything to gain by changing only his own strategy.

Several methods to compute NE of a game have been developed. For a review on computing techniques for the NE see [9].

Consider two strategy profiles s^* and s from S. An operator $k : S \times S \to N$ that associates the cardinality of the set

$$k(s^*, s) = |(\{i \in \{1, ..., n\} | u_i(s_i, s^*_{-i}) \geq u_i(s^*), s_i \neq s^*_i\}|$$

to the pair (s^*, s) is introduced.

This set is composed by the players i that would benefit if - given the strategy profile s^* - would change their strategy from s^*_i to s_i, i.e.

$$u_i(s_i, s^*_{-i}) \geq u_i(s^*).$$

Let $x, y \in S$. We say the strategy profile x **Nash ascends** the strategy profile y in and we write $x \prec y$ if the inequality

$$k(x, y) < k(y, x)$$

holds.

Thus a strategy profile x dominates strategy profile y if there are less players that can increase their payoffs by switching their strategy from x_i to

y_i than vice-versa. It can be said that strategy profile x is more stable (closer to equilibrium) then strategy y.

Two strategy profiles $x, y \in S$ may have the following relation:

(1) either x dominates y, $x \prec y$ $(k(x,y) < k(y,x))$
(2) either y dominates x, $y \prec x$ $(k(x,y) > k(y,x))$
(3) or $k(x,y) = k(y,x)$ and x and y are considered indifferent (neither x dominates y nor y dominates x).

The strategy profile $s^* \in S$ is called non-ascended in Nash sense (NAS) if

$$\nexists s \in S, s \neq s^* \text{such that } s \prec s^*.$$

In [8] it is shown that all non-ascended strategies are NE and also all NE are non-ascended strategies. Thus the Nash ascendancy relation can be used to characterize the equilibria of a game.

3. Selection of Nash Equilibria

Using the ascendancy relation an equilibrium can be characterized by the number of strategies it ascends. The equilibrium ascending most strategy profiles may be considered to be the most "popular' equilibrium and thus a selection method is proposed.

In order to approximate the number of strategies ascended by an equilibrium the following method of comparing equilibria is proposed.

A population of strategies of size R is uniformly random generated 100 times. For each equilibrium and each population the number of strategies ascended by the equilibrium and the number of strategies that are indifferent to the equilibrium is computed.

The ratio of that number to R is a number between 0 and 1 representing a measure of ascendancy of that equilibrium. The average of these numbers over the 100 populations is denoted by M_a and can be used to compare different equilibria of a game.

The corresponding standard deviation S_a is also computed. These measures represent simple descriptive statistics tools that indicate the potential of the method.

4. Numerical examples

Several normal form games presenting multiple Nash Equilibria are presented. The size R is considered of 100000 strategies. Each game is presented by its payoff matrix and the Nash equilibria. For each NE, $M_a(NE)$ and $S_a(NE)$ are presented.

P1-P2	1	2
1	1,3	4,2
2	2,1	1,3

Table 1. Payoff table for Game 1

The payoff space illustrated for each game is generated by representing 100000 uniformly generated strategies. This representation is used in multi-objective optimization and offers some extra input in the features of the game.

4.1. **Game 1.** The first game has been chosen for illustration purposes. It is a game with two players each having two strategies. The payoff matrix is presented in Table 1. This game has one mixed NE at $(2/3, 1/3)$ and $(3/4, 1/4)$. Th e payoff space is visualized in Figure 1. All tested strategies are ascended by the NE of the game and $M_a(NE) = 100000$ and $S_a(NE) = 0$.

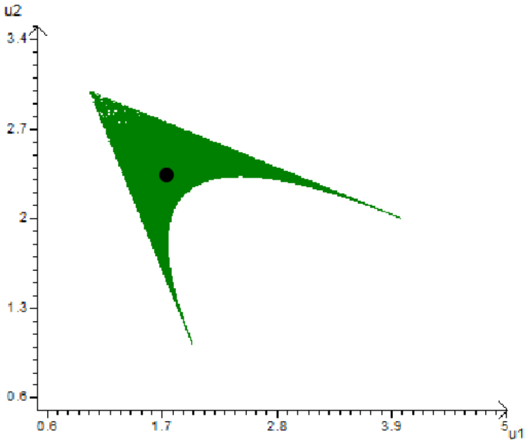


Figure 1. Game1. The circle represents the NE

4.2. **Game2.** This game represents a discrete four step version of the centipede game [11]. It is a two player game with two strategies for each player. The payoff matrix is presented in Table 2.

This game has two NE presented in Table 3, one in pure form and one in mixed form, both having the same payoff illustrated in Figure 2.

According to our experiments, both NE ascend the same number of strategies as $M_a(NE_1) = M_a(NE_2) = 1$ and $S_a(NE_1) = S_a(NE_2) = 0$. this is a specific feature for this game. Whatever the second player will chose, when

P1-P2	1	2
1	3,1	3,1
2	2,6	12,4

Table 2. Payoff table for Game 2

NE	1	2	Payoffs
1	1,0	1,0	(3,1)
2	1,0	0.9, 0.1	(3,1)

Table 3. NEs for Game 2

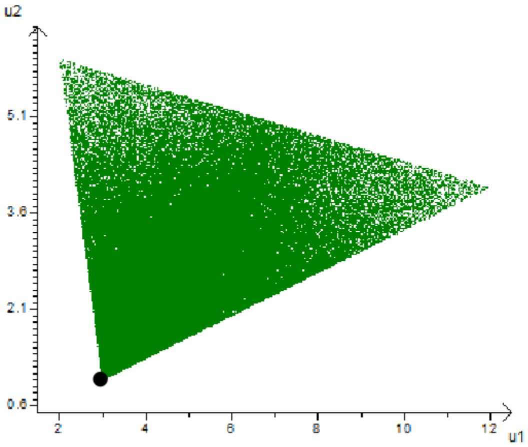


Figure 2. Game2. The circle represents the payoff for the two NE

the first one plays it first strategy the payoff for both is the same and there is no point in choosing between the two NE.

4.3. **Game 3.** The third game has two players each of them having three strategies available and payoffs in Table 4. It presents three Nash equilibria, two pure and one in mixed form as presented in Table 5. The pure Nash Equilibria weakly Pareto dominate the mixed one (Figure 3).

The results of our experiments are also presented in Table 5. These indicate a slightly 'higher popularity" of the mixed NE over the pure ones even though is weakly dominated by both. However, further statistical tools have to be used to determine if the difference between results is significant.

P1-P2	1	2	3
1	5,5	10,8	6,7
2	8,10	8,8	10,8
3	7,6	8,10	5,5

Table 4. Payoff table for Game 3

NE	1	2	Payoffs	M_a	S_a
1	**0.4, 0.6, 0**	**0.4,0.6,0**	**(8,8)**	**0.9904**	**0.0002**
2	1,0,0	0,1,0	(10,8)	0.9809	0.0004
3	0,1,0	1,0,0	(8,10)	0.9810	0.0004

Table 5. NEs for Game 3

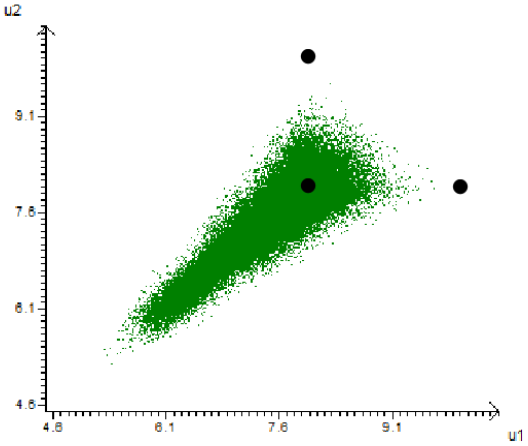


Figure 3. Payoff space for Game3. Circles represent payoffs of the three NE

4.4. **Game 4.** The fourth game is a two player game, each with three strategies. It presents nine NEs, three in pure form and six in mixed form. Payoffs are given in Table 6, the equilibria and results in Table 7 and the payoffs space is illustrated in Figure 4. According to these results, the 'best' choice would be the last NE yielding a payoff $(2, 2)$.

5. Conclusions and further work

An attempt to introduce a new method for selecting between multiple Nash Equilibria of a normal form game is made in this paper.

P1-P2	1	2	3
1	3,1	0,0	0,1
2	1.5,1	2,2	1.5,1
3	0,1	0,0	3,1

TABLE 6. Payoff table for Game 4

NE	P1	P2	Payoffs	M_a	S_a
1	0.5,0.5,0	0.5714, 0.4285,0	(1.71,1)	0.7899	0.0011
2	0.5,0.5,0	0.5,0,0.5	(1.5,1)	0.8925	0.0009
3	0,0.5,0.5	0.5,0,0.5	(1.5,1)	0.8924	0.0009
4	0,0.5,0.5	0, 0.4285,0.5714	(1.71,1)	0.7898	0.0012
5	1,0,0	1,0,0	(3,1)	0.8670	0.0010
6	1,0,0	0.5,0,0.5	(1.5,1)	0.8670	0.0010
7	0,0,1	0.5,0,0.5	(1.5,1)	0.8669	0.0009
8	0,0,1	0,0,1	(3,1)	0.8669	0.0009
9	**0,1,0**	**0,1,0**	**(2,2)**	**0.9620**	**0.0005**

TABLE 7. NEs for Game 4

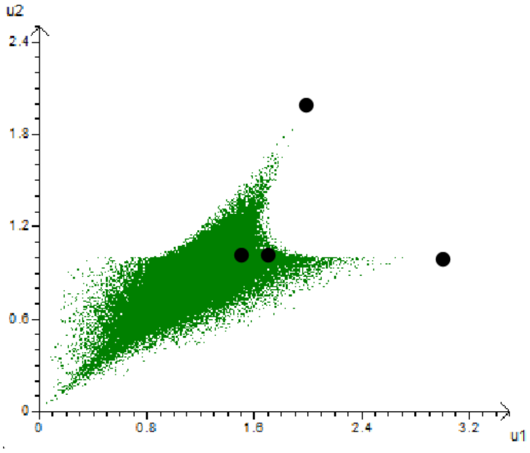


FIGURE 4. Payoff space for Game4. Circles represent payoffs of the NE

An ascendancy relation is used to determine which of the NE is most 'popular' or ascends most strategies from a randomly generated set. Some simple numerical experiments illustrate the use of this method.

Further work includes the use of statistical inference tools in order to evaluate the significance of the results.

References

[1] Patrick Bajari, Han Hong, and Stephen P. Ryan. Identification and estimation of a discrete game of complete information. *Econometrica*, forthcoming paper, 2010.

[2] Steven Berry and Peter Reiss. *Empirical Models of Entry and Market Structure*, volume 3 of *Handbook of Industrial Organization*, chapter 29, pages 1845–1886. Elsevier, June 2007.

[3] Steven T Berry. Estimation of a model of entry in the airline industry. *Econometrica*, 60(4):889–917, July 1992.

[4] Paul A. Bjorn and Quang H. Vuong. Simultaneous equations models for dummy endogenous variables: A game theoretic formulation with an application to labor force participation. Working Papers 537, California Institute of Technology, Division of the Humanities and Social Sciences, July 1984.

[5] Timothy F Bresnahan and Peter C Reiss. Entry in monopoly markets. *Review of Economic Studies*, 57(4):531–53, October 1990.

[6] Timothy F. Bresnahan and Peter C. Reiss. Empirical models of discrete games. *Journal of Econometrics*, 48(1-2):57–81, 1991.

[7] Panle Jia. What happens when wal-mart comes to town: An empirical analysis of the discount retailing industry. *Econometrica*, 76(6):1263–1316, November 2008.

[8] Rodica Ioana Lung and D. Dumitrescu. Computing nash equilibria by means of evolutionary computation. *Int. J. of Computers, Communications & Control*, III(suppl.issue):364–368, 2008.

[9] Richard D. McKelvey and Andrew McLennan. Computation of equilibria in finite games. In H. M. Amman, D. A. Kendrick, and J. Rust, editors, *Handbook of Computational Economics*, volume 1 of *Handbook of Computational Economics*, chapter 2, pages 87–142. Elsevier, 1 1996.

[10] John F. Nash. Non-cooperative games. *Annals of Mathematics*, 54:286–295, 1951.

[11] Robert W. Rosenthal. Games of perfect information, predatory pricing and the chainstore paradox. *Journal of Economic Theory*, 25(1):92–100, August 1981.

[12] Elie Tamer. Incomplete simultaneous discrete response model with multiple equilibria. *Review of Economic Studies*, 70(1):147–165, January 2003.

Babes-Bolyai University, Faculty of Economics and Business Administration
E-mail address: rodica.lung@econ.ubbcluj.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LV**, Number 1, 2010

MODEL ALIGNMENT BY USING THE CONCEPT DEFINITIONS

ADELA SÎRBU[1,2], LAURA DIOŞAN[1,2], ALEXANDRINA ROGOZAN[1],
JEAN-PIERRE PECUCHET[1]

Abstract. The alignment between two dictionaries will certainly improve the performances of the information retrieval process. We develop a custom terminology alignment by an SVM classifier with an optimised kernel trained on a compact, but relevant representation of such definition pairs by several similarity measures and the length of definitions. The aligner was trained on a database of aligned definitions that was semi-automatically created by using the Coma++ tool. The results obtained on the test set show the relevance of our approach.

1. Introduction

One of the goals of the ASICOM project[1] is to improve the fusion between a specialized terminology and a general vocabulary employed by a neophyte user in order to retrieve documents on Internet. These goals could be summarised as follows:

- to design a model alignment in order to help or to guide the automatically transformation of dictionaries;
- to quantify the role of ontologies and/or hierarchies of concepts/dictionaries for the model transformation;
- to align two concepts from their definitions only, or from their definitions and their paths in the hierarchies of concepts, or from their definitions, their paths and their fathers in the hierarchies of concepts.

The goal of the mapping between two hierarchies of concepts (dictionaries) is to align a concept from a dictionary with a concept of another dictionary

Received by the editors: January 10, 2009.

2010 *Mathematics Subject Classification.* 68T05,91E45.

1998 *CR Categories and Descriptors.* code I.2.6 [**Learning**]: – *Concept learning.*

Key words and phrases. Concept alignment, Machine Learning, Binary Classification, Support Vector Machine.

[1] Architecture de Systme d'information Interoprable pour les industries du Commerce, project financed by *Pôle de compétitivité Industries du Commerce* and *Logistique Seine-Normandie, 2009-2011*

by using their definitions, since their names may be different. In other words, the purpose is to identify concepts semantically identical, even though their concept labels are different.

The correspondence may be such that:

- a concept from a dictionary matches even several concepts from another dictionary. In other words, several concepts of the second dictionary can then be labelled with the same concept from the first dictionary. These concepts could then be seen as "the result of" the concept of the first dictionary. We actually deal by a "sort of" relationship.
- a concept from a dictionary matches at most one concept from another dictionary. In this case, the two concepts are so similar or equivalent; therefore we discuss about an "equivalence" relation.

Of course, another relationships could be discovered among the content of two dictionaries.

In order to perform some numerical experiments two dictionary were actually considered: ASICOM_CCL08A (or, shorter, CCL) and Customs_WCO (WCO). The first dictionary contains 2191 of concepts and for each concepts information about id, object class term (or father in the hierarchy), property term, representation term, entry name (in fact, label and path) and explanation are retained. In the case of WCO dictionary, we deal with 264 concepts for which we have retained the id, the entry name (WCO name or label and UNTDED Name or path in the hierarchy of concepts), data model class (actually, the concept father) and the explanation.

In order to perform our analysis, we have to build a database with couples of definition aligned. This base could be realised in a semi-automatically manner or manually, by a specialist area. Because our aim is to design a human-knowledge independent application, we have decided to use a database that was constructed in a semi-automatically manner by using a special tool: Coma++.

The definitions aligned by Coma++ are utilised as train data for a Machine Learning algorithm in order to aligned unseen definitions. In fact, the database constructed by Coma++ help us to enable the self-aligning process, as well as it serves as a repository for automatic alignment. The Coma++ principles could also be used as a complementary technique to discover alignments not seen by the automatic alignment.

This paper is structured as follows: Section 2 gives a short review of different alignment models. Section 3 details the characteristics of the corpora and of the linguistic treatments that have been performed. The alignment model is described and analysed (through several numerical experiments) in the next two sections. Finally, Section 6 concludes the paper.

2. Related work

To our knowledge, only the problem of aligning sentences from parallel bilingual corpora has been intensively studied for automated translation. While much of research has focused on the unsupervised models [1, 4], a number of supervised discriminatory approaches have been recently proposed for automatic alignment [2, 10, 12].

One of the first algorithms used to align parallel corpora proposed by Brown [1] is based solely on the number of words/characters in each sentence. Chen [4] has developed a simple statistical word-to-word translation model. Dynamic programming, at the level of words, performs the search of the best alignment in these models.

Related to the use of linguistic information a more recent work [11] shows the benefit of combining multilevel linguistic representations (these levels refer to morphological, syntactic and semantic analyses). Moreover, data fusion has been exhaustively investigated in the literature, especially in the framework of Information Retrieval [11].

Concerning the supervised methods, Taskar et al. [12] have cast the word alignment as a maximum weighted matching problem where each pair of sentences has associated a score function, which reflects the desirability of the alignment of that pair. The alignment for the sentence pair corresponds to the highest scoring matching under some constraints (for instance, the requirement that matching be one-to-one). Moore [10] has introduced a hybrid and supervised approach that adapts and combines the sentence-length-based methods with the word-correspondence-based methods. Ceausu [2] has proposed another supervised hybrid method that uses a Support Vector Machine (SVM) classifier [14] to distinguish between aligned and non-aligned examples of sentence pairs; each pair has been represented by a set of statistical characteristics (like translation equivalence, word sentence length correlation, character sentence length correlation, word rank correlation, non-word sentence length correlation).

The model we develop in what follows borrows some aspects from Moore and Ceausu's approaches, but it is enriched with several new elements. Our model considers the alignment task as a classification problem (as in Ceausu's case). Although, in our case, the information about definitions is organised based on several similarity measures, while the classification problem is solved by using an SVM algorithm, which involves an optimised kernel function.

3. The corpora and the linguistic processing

3.1. **Coma++.** Coma++ is actually a tool useful for semi-automatically alignment of concepts [8]. It is a schema and ontology matching tool. It utilises a composite approach to combine different match algorithms. Furthermore, it offers a comprehensive infrastructure to solve large real-world matching problems. The graphical interface offers a variety of interactions, allowing the user to influence in the match process in many ways.

It is based on a composite schema matching and a flexible framework for combining matching algorithms. COMA++ supports a comprehensive and extensible library of individual matchers, which can be selected to perform a match operation. Using the GUI, it is easy to construct new, more powerful, matchers by combining existing ones. Moreover, it is possible to specify match strategies as work flows of multiple match steps, allowing to divide and successively solve complex match tasks in multiple stages [8].

Taken into account the alignments performed by Coma++ between definitions from CCL and WCO dictionaries, our aim was to develop a Machine Learning algorithm that will be able to put in correspondence more definitions of two different dictionaries and to improve the performance of alignment compared to Coma++. For this purpose a statistical learning based on SVM [14] is performed from a base of learning achieved after manual correction of alignments produced by Coma++.

The model we propose performs two important steps: represent, in a particular manner, the couples of definitions and than, learn or classify these couples.

Before we present our approach utilised to align the definitions, several details about the preliminary treatments of these definitions are provided.

3.2. **Linguistic processing.** As we already said, in order to automatically perform the alignment, several definitions are considered from two dictionaries: WCO and CCL. The English is the common language for both dictionaries. Each definition is retaining as a vector of words, each element of this vector being enriched only with its lemma and its synonyms.

The literature shows that a purely statistical approach on the plain text provides weak results for automatic text understanding. Several linguistic treatments, such as the labelling at the syntactic level (POS - Parts of speech - tagging) must be performed. Therefore, in order to achieve an efficient automatic classification "aligned" *vs.* "not aligned" of the definition couples, the following (structural and semantic) linguistic processing has been performed:

- segmentation – consists in cutting a sequence of characters so that various characters that form a single word can be brought together. Classic segmentation means cutting the sequences of characters depending on several separation characters such as "space", "tab" or "backspace";

- filtering of the stop words – *stop words* is the name given to words like *in*, *a*, *of*, *the*, *on* that are not representative should not be taken into consideration;
- bringing the words to the canonical form – in order to work directly with the words they must be brought to a canonical form. For instance the words *uses*, *using*, *used* refer to the same thing but under different forms, but if they are compared like this it will be obtained that they are different. There are different ways to solve this problem, and one of them could be to apply a stemming algorithm [13] – is the process of reducing inflected (or sometimes derived) words to their stem, base or root form. The stem does not have to be identical to the morphological root of the word; it is usually sufficient that related words map the same stem, even if this stem is not a valid root in itself.

3.3. **Similarity measures.** To enable a rapid and effective learning of definition alignment, we must avoid the problem associated with a classic representation based on the tf-idf[2] weighting scheme where the bags of words are translated into vectors of large sizes. In our case, the large size of such vectors is equal to the number of words contained by all the definitions chosen from all the dictionaries. In addition, the definitions could be considered as short text and thus, some sparse vectors will correspond to each definition. Therefore, we use several measures of similarity between two structures:

- the *Matching* coefficient [6] – it counts the common elements of the given structures.
- the *Dice* coefficient [7] – it is defined as twice the number of common elements, divided by the total number of elements,
- the *Jaccard* coefficient [9] – it is defined as the number of common elements, divided by the total number of elements,
- the *Overlap* coefficient [5] – it is defined as the number of common elements, divided by the minimum of the element numbers from the given structures,
- the *Cosine* measure – it is defined as the number of common elements, divided by the square of sum between the element number from the first structure and the element number from the second structure.

These statistics are generally used for comparing the similarity and diversity of two sample sets, but they can be adapted to our definition couples and their representation. In order to compute a similarity measure between two definitions, each of them are tokenized (segmentation process), lemmatised and syntactic labelled. In this way, a bag of labelled lemmas is obtained for

[2] term frequency-inverse document frequency

each definition. Then, based on the elements of the corresponding bags, the similarity coefficient of two definitions is computed. The considered definitions can be taken from the same dictionary or from different dictionaries (a general one and a specialised one). Based on the obtained similarities we will decide if the two definitions are aligned or not.

By working only with a representation based on these measures, instead of a classical one, the models we propose are able to map the initial vectors (based on a bag of word approach) into a space of reduced dimension, where the computation effort is smaller. Furthermore, we will see if by this reduction we could loose information.

4. SVM alignment

The alignment is considered as a classification problem where each input is represented by the similarity between two definitions. The label associated to that couple of definitions (aligned or not aligned) represents the output. An SVM algorithm [14] is actually used to perform this classification-alignment.

First of all we represent each definition couple by one of the already presented similarity measures. Although it is very simple to work with such representation, we do not know *a priori* which measure works the best. Therefore, we propose to take into account the complementarities between these similarity measures. All five similarity measures are simultaneously considered, obtaining a compact representation for each couple of two definitions. In addition to the similarity measures, the new representation contains the length of each definition too.

The classification process takes place in two phases that reflect the principles of a learning algorithm. Therefore, each data set[3] has to be divided in two parts: a part for training and a part for testing. The training part is divided again in: a learning sub-set – used by the SVM algorithm in order to learn the model that performs the class separation – and a validation sub-set – used in order to optimise the values of the hyper parameters. The SVM model, which is learnt in this manner, classifies (labels) the unseen definition couples from the test set, which is disjoint to the training one.

In order to classify the definition couples, the SVM algorithm uses one of the above representations and a kernel function. The parameters of the SVM model (the penalty for miss-classification C and the kernel parameters) are optimized on the validation set. A cross-validation framework is utilised in order to avoid the over fitting problems. Thus, we automatically adapt the SVM classifier to the problem, actually the alignment of definitions.

[3] that corresponds to all the couples formed by the definitions from two dictionaries

5. Numerical experiments

5.1. **Construction and analysis of the database.** The training database (in fact, definitions aligned by using Coma++) was provided by Yuhan GUO and Rémy DUPAS, IMS - LAPS - GRAI, from University Bordeaux. They have used a set of matchers composed from Affix, 2-gram 3-gram, Edit Distance, Synonym, Soundex and DataType. The weight setting (the weight corresponding to each matcher) was that default. The threshold for accepting an alignment was set to 0.4 (in fact, the couples with the similarity less than the threshold were ignored). Furthermore, the Coma++ tool allows two types of alignment: single condition and multiple conditions. In the first case, only the definitions of concepts have been used in order to perform the alignments, while in the case of a multiple-conditions alignments, Y. Gao and R. Dupas have taken into consideration the definitions, the paths and the fathers of each concept of the two dictionaries.

After a short analysis of the alignments performed by Coma++ we have obtained the following synthesis:

- the number of alignments produced by Coma++ was:
  - mono-condition: 50 pairs of definitions;
  - multi-condition: 159 pairs of definitions;
- the cardinality of alignments in both cases (mono and multi-condition):
  - one to one: a definition WCO was aligned with a single definition CCL;
  - one to many: a CCL definition was aligned with several definitions WCO;
- the alignments cover single and multi-condition:
  - mono-condition $\cap$ multi-condition = 33 couples (in fact, there are 33 common alignments in mono and multi-condition case);
  - mono-condition – multi-condition = 17 couples (17 alignment couples appear in the mono-condition base and they not appear in the multi-condition base);
  - multi-condition – mono-condition = 126 couples (126 alignment couples appear in the multi-condition base and they not appear in the mono-condition base).

5.2. **Numerical experiments performed by SVM.** A set of experiments are performed by using the SVM-based model and the representation discussed in Section 3.3 (that based on five similarity measures).

The train and test data are composed from aligned and not-aligned couples of definitions from CCL and WCO dictionaries, respectively. The aligned couples are represented by the aligned pairs provided by Coma++, while the not-aligned couples are represented by the definitions from the two dictionaries that were not provided as aligned by Coma++. From all the couples that are formed in this manner, 2/3 of them are considered for training the SVM algorithm and 1/3 of them for testing the aligner.

The C-SVM algorithm, provided by LIBSVM [3], with an RBF kernel is actually used in this experiment. The optimisation of the hyper-parameters is performed by a parallel grid search method. For each combination of these parameters, a 10-fold cross validation[4] is performed during the training phase, the quality of a combination being computed as the average of the accuracy rates estimated for each of the 10 divisions of the data set. Therefore, the best combination is indicated by the best average accuracy rate.

The values of the optimal hyper-parameters and the accuracy rates obtained for 4 different definition processing are presented in Table 1.

Table 1. The performance of the SVM-based aligner.

Pre-processing	RBF Kernel		Accuracy rate	
	γ	C	Mono	Multi
All the words	1/7	10	98.61	98.43
All the words + stemming	1/7	10	98.96	98.07
All the words without stop words	1/7	10	98.26	98.18
All the words without stop words + stemming	1/7	10	98.26	98.07

In order to validate our results, we plan to repeat the experiments by using the model proposed by Ceausu [2], even if a fair comparison between the two models is not possible since the text to be align was different preprocess. We also plan to compare the SVM results with those obtained by other classification algorithm.

6. Conclusions and remarks

In this paper we presented our model for the automatic alignment of definitions taken from two dictionaries (CCL and WCO). The best performances

[4] Cross-validation is a popular technique for estimating the generalization error and there are several interpretations [15]. In k-fold cross-validation, the training data is randomly split into k mutually exclusive subsets (or folds) of approximately equal size. The SVM decision rule is obtained by using $k-1$ subsets on training data and then tested on the subset left out. This procedure is repeated k times and in this manner each subset is used for testing once. Averaging the test error over the k trials gives a better estimate of the expected generalization error.

are obtained by using the SVM algorithm with an RBF kernel and by considering the stemms of all the words of each definition, since the classifier (in fact the hyper-parameters) is better adapted to the alignment task to be solved. However, these conclusions should be validated on some larger corpora.

Further work will be focused on: considering a representation of definitions enriched by semantic and lexical extensions (synonyms, hyponyms, and antonyms) and on developing of an alignment model based on an SVM algorithm with a specialised multiple kernel (this specialisation could be considered in terms of combination of more kernels for text processing (*e.g.* string kernels)).

References

[1] Brown, P. F., Pietra, S. D., Pietra, V. J. D., and Mercer, R. L. The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics 19*, 2 (1994), 263–311.

[2] Ceausu, A., Stefanescu, D., and Tufis, D. Acquis communautaire sentence alignment using Support Vector Machines. In *Proceedings of the 5th LREC Conference* (2006), pp. 2134–2137.

[3] Chang, C.-C., and Lin, C.-J. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] Chen, S. F. Aligning sentences in bilingual corpora using lexical information. In *Meeting of the Association for Computational Linguistics* (1993), ACL, pp. 9–16.

[5] Clemons, T. E., and Bradley, E. L. A nonparametric measure of the overlapping coefficient. *Comput. Stat. Data Anal. 34* (2000), 51–61.

[6] Cormen, T., Leiserson, C., and Rivest, R. *Introduction to Algorithms.* MIT Press, 1990.

[7] Dice, L. Measures of the amount of ecologic association between species. *Ecology 26*, 3 (1945), 297–302.

[8] Do, H.-H., and Rahm, E. Coma++ (combination of schema matching approaches), 2010.

[9] Jaccard, P. The distribution of the flora of the alpine zone. *New Phytologist 11* (1912), 37–50.

[10] Moore, R. Fast and accurate sentence alignment of bilingual corpora. In *AMTA '02* (2002), S. D. Richardson, Ed., Springer, pp. 135–144.

[11] Moreau, F., Claveau, V., and Sébillot, P. Automatic morphological query expansion using analogy-based machine learning. In *ECIR 2007* (2007), G.Amati, C. Carpineto, and G. Romano, Eds., vol. 4425 of *LNCS*, Springer, pp. 222–233.

[12] Taskar, B., Lacoste, S., and Klein, D. A discriminative matching approach to word alignment. In *HLT '05* (2005), Association for Computational Linguistics, pp. 73–80.

[13] van Rijsbergen, C., Robertson, S., and Porter, M. New models in probabilistic information retrieval. Tech. Rep. 5587, British Library, 1980.

[14] Vapnik, V. *The Nature of Statistical Learning Theory.* Springer, 1995.

[15] Wahba, G., Lin, Y., and Zhang, H. GACV for Support Vector Machines. In *Advances in Large Margin Classifiers*, B. Smola and S. SchRolkopf, Eds. MIT Press, Cambridge, MA, 1999.

[16] wei Hsu, C., chung Chang, C., and jen Lin, C. A practical guide to support vector classification, 2003.

[1] LITIS, EA - 4108, INSA, Rouen, France, [2] Computer Science Department, Babeş Bolyai University, Cluj Napoca, Romania

E-mail address: adela_sarbu25@yahoo.com, lauras@cs.ubbcluj.ro

E-mail address: arogozan@insa-rouen.fr, pecuchet@insa-rouen.fr

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LV**, Number 1, 2010

DISCOVERING PATTERNS IN MUSIC. AN FCA GROUNDED APPROACH

ZSUZSANNA MARIAN AND CHRISTIAN SĂCĂREA

Abstract. We apply methods of Formal Concept Analysis in order to discover patterns in modern music. For this, we investigate the *Music Genome*, a project started on the 6th of January 2000 by a group of musicians and music-loving technologists who came together with the idea of creating the most comprehensive analysis of music ever. We also present an environment for processing and representation of music patterns.

1. Introduction

Motto: How sweet the moonlight sleeps upon this bank! Here will we sit, and let the sound of music Creep in our ears; soft stillness and the night Become the touches of sweet harmony.

– Shakespeare (Merchant of Venice, Lorenzo, Act 5, scene 1)

Discovering patterns in music is a usual research topic in music theory (see for instance [4], [5], and [6]). Also, the connection between music and mathematics and physics has been intensively studied along the centuries (see also [2], [3]).

Nevertheless, there are surprisingly few attempts to apply modern knowledge processing and representation methods to the study of musical patterns. The only attempt we know was an experiment conducted by a group of mathematicians and musicians at the Darmstadt University of Technology, Germany. An exhaustive attribute exploration was performed on a set of attributes for classical music in order to obtain a knowledge universe, i.e., a complete set of examples and counterexamples related to the considered set of attributes.

Received by the editors: January 18, 2010.

2010 *Mathematics Subject Classification.* 68P15 Database theory, 03G10 Lattices and related structures.

1998 *CR Categories and Descriptors.* code H.2.1 [**Database Management**]: Logical Design – *Normal Forms*; F.4.m [**Mathematical Logic and Formal Languages**]: Miscellaneous.

Key words and phrases. Conceptual Knowledge Processing, Formal Concept Analysis, Music Genome.

In this paper, we have performed a first study on music patterns for modern music. For this, we have used the so-called music genes, essential attributes for melodies, which are then combined to music. Like the discovery of the human genome, the discovery of the music genome was the result of an intensive work of a group of enthusiastic musicians and music loving technologists. The musical pool under investigation was modern music. While classical music has been exhaustively studied, modern music is actually a huge field for research, insufficiently exploited by modern science.

The French-born composer Edgar Varese (1883–1965) once defined music to be organized sound. But how is music organized and what type of organization is intended?

Classical music theory gives a partial answer to these questions, in terms of tones and harmonies. The Music Genome Project goes one step ahead and investigates the building harmonies of melodies, the so-called genes.

But how are these genes related? How are they related to the music in which they appear, what is the knowledge they encode? How does the logic of these genes look like and how can we discover as much knowledge as possible by studying the Music Genome?

Conceptual Knowledge Processing and its mathematical theory on which this is grounded, the Formal Concept Analysis, appears to open the possibility not only for a satisfactory answer to the questions above, but also for a comprehensive analysis of data and related concepts.

There is a long philosophical tradition in investigating concepts, ordering them in a certain hierarchy of subconcept-superconcept. This tradition has been reflected several times in the development of mathematics, and it can be refound in almost every part of modern mathematics: remember the nineteenth-century attempts to formalize logic (see for example the *Algebra of Logic* of Ernst Schröder). Traditionally, a concept is determined by its extent and its intent (or comprehension). The extent of a concept consists of all objects, individuals or entities which belong to the concept, while the intent consists of all properties which are considered valid for that concept. The hierarchy of concepts is given by the relation of subconcept wrt. certain superconcept, i.e., the extent of a subconcept is part of the extent of the superconcept, while the inverse relation holds for the corresponding intents.

Consider a usual concept with which one usually deal, for example landscape. There is no possibility to list every object of this concept or even to consider in a generally satisfactory way all properties of that what we understand as a landscape. It follows that there is an intrinsic necessity to restrict ourselves to a given set of objects and a clearly defined set of attributes, process which is characteristic for the human thinking, cutting off sections of the reality which will be here later on called formal contexts.

Formal Concept Analysis was born from this vigorous philosophical tradition. At the end of the 1970s, Formal Concept Analysis has crystalized first in an attempt of restructuring lattice theory (see [14]), where R. Wille discusses the status of modern lattice theory considering that "...abstract developments should be brought back to the commonplace in perception, thinking, and action. Thus *restructuring lattice theory* is understood as an attempt to unfold lattice-theoretical concepts, results, and methods in a continuous relationship with their surroundings."

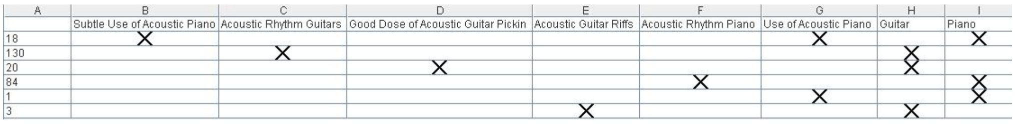
As a mathematical theory, Formal Concept Analysis is based on the formalization of the notion of concept and of the medium from where this concept arises, the formal context. Formally speaking a formal context (mathematically defined in the next section) is a triple consisting of two sets and a binary relation between them. Despite of its simplicity, a formal context encodes in its incidence relation some structural information which can be found in the so-called formal concepts, which can be seen as a kind of closed pieces of the information encoded in the considered formal context.

From now on, Formal Concept Analysis became more as an attempt of reconsidering parts of lattice theory, giving the possibility of investigating a wide range of empirical phenomena using mathematical methods in the study of the data collected about these phenomena. FCA is not only a part of Mathematics, it has deep going results and connections with Computer Science, Medical Science, Social Sciences, Psychology and others.

2. Formal Concept Analysis

2.1. **Context and Concept.** As we have seen before, Formal Concept Analysis is based on a set theoretical model proposing a new paradigm of thinking. A **formal context** $\mathbb{K} := (G, M, I)$ consists of two sets G and M and a binary relation I between G and M. The elements of G are called **objects** (in German *Gegenstände*) and the elements of M are called **attributes** (in German *Merkmale*). The relation I is called the incidence relation of the formal context, and we sometimes write gIm instead of $(g, m) \in I$. If gIm holds, we say that *the object g has the attribute m.*

A small context is usually represented by a cross table, i.e., a rectangular table of crosses and blanks, where the rows are labeled by the objects and the columns are labeled by the attributes. A cross in entry (g, m) indicates gIm. Figure 1 illustrates a formal context. The attributes are the so-called *acoustic instruments music genes*, i.e., Subtle Use of Acoustic Piano, Acoustic Rhythm Guitars, Good Dose of Acoustic Guitar Pickin, Acoustic Guitar Riffs, Acoustic Rhythm Piano, Use of Acoustic Piano, Guitar, Piano. The objects are songs, having in their structure at least one of the considered music genes. Since the list of songs is very large, we have displayed as object names only the number of songs into consideration. The incidence relation is displayed by a series of crosses in the table, indicating when a specific song has a given music gene.



A	B	C	D	E	F	G	H	I
	Subtle Use of Acoustic Piano	Acoustic Rhythm Guitars	Good Dose of Acoustic Guitar Pickin	Acoustic Guitar Riffs	Acoustic Rhythm Piano	Use of Acoustic Piano	Guitar	Piano
18	×					×		×
130		×					×	
20			×				×	
84					×			×
1						×		×
3				×			×	

FIGURE 1. The Acoustic Instruments context.

For a set $A \subseteq G$ of objects we define

$$A' := \{m \in M \mid gIm \text{ for all } g \in A\}$$

the set of all attributes common to the objects in A. Dually, for a set $B \subseteq M$ of attributes we define

$$B' := \{g \in G \mid gIm \text{ for all } m \in B\}$$

the set of all objects which have all attributes in B.

A **formal concept** of the context $\mathbb{K} := (G, M, I)$ is a pair (A, B) where $A \subseteq G$, $B \subseteq M$, $A' = B$, and $B' = A$. We call A the **extent** and B the **intent** of the concept (A, B). The set of all concepts of the context (G, M, I) is denoted by $\mathfrak{B}(G, M, I)$.

2.2. **Many-valued contexts.** As we have seen from the definition of a formal context, an object could have or not some attribute, i.e., the attributes were one-valued. The more general situation is that where attributes have values. We call them *many-valued attributes*, in contrast to the *one-valued attributes* considered so far.

Definition 1. A **many-valued context** (G, M, W, I) consists of sets G, M and W and a ternary relation I between G, M and W (i.e., $I \subseteq G \times M \times W$) for which the following holds:

$$(g, m, w) \in I \text{ and } (g, m, v) \in I \text{ always imply } m = v.$$

The elements of G are called **objects**, those of M (**many-valued**) **attributes** and those of W **attribute values**.

Like the one-valued contexts treated so far, many-valued contexts can be represented by tables, the rows of which are labelled by the objects and the columns labelled by the attributes. The entry in row g and column m represents then the attribute value $m(g)$. If the attribute m does not have a value for the object g, there will be no entry.

In order to assign concepts to a many-valued context we first have to transform it into a formal context, according to some rules, rules which are called **scaling the many-valued context**. The concepts of this *derived* one-valued context are then interpreted as the concepts of the many-valued context. Since data captured in a many-valued context can be very complex, it is obviously that the interpretation process through which a conceptual structure is assigned, called *conceptual scaling* (see [GW 99]) is not uniquely determined. Changing the scale we can get new insights about the information of a given many-valued context. In the process of scaling, first of all each attribute of a many-valued context is interpreted by means of a context. This context is called *conceptual scale.*

Definition 2. A **scale** for the attribute m of a many-valued context is a (formal) context $\mathbb{S}_m := (G_m, M_m, I_m)$ with $m(G) \subseteq G_m$. The objects of a scale are called **scale values**, the attributes are called **scale attributes**.

As we can see, there is no formal difference between a formal context and a scale. But according to the tradition of the Darmstadt school of Formal Concept Analysis, the term *scale* will be used only for contexts which have a clear conceptual structure and which bear meaning.

There are several ways in which a scale for a given attribute $m \in M$ can be chosen. The same holds for combining the scales in order to obtain a formal context from the given many-valued one. The simplest case is called **plain scaling**, and it consists of putting together the individual scales without connecting them.

Definition 3. If (G, M, W, I) is a many-valued context and $\mathbb{S}_m$, with $m \in M$, are scales, then the **derived context with respect to plain scaling** is the context (G, N, J) with

$$N := \bigcup_{m \in M} \{m\} \times M_m,$$

and

$$gJ(m,n) :\Leftrightarrow m(g) = w \; and \; wI_m n.$$

3. The Music Genome

The Music Genome Project was started on the 6th of January 2000 by a group of musicians and music-loving technologists who came together with the idea of creating the most comprehensive analysis of music ever, according to Tim Westergren, one of the founders of this project [7]. They have decided to analyze the structure of a song, defining different attributes - ”genes” - for them and identify similar songs that the listener could be interested in. Nolan Grasser, an actual musicologist whose doctoral thesis dealt with the close analysis of Renaissance composition, helped Westergren to create the lexicon that could transform his genome idea into something a computer could evaluate [8]. During the analysis, every song is first broken down into the large-scale aspects of the music: melody, harmony, rhythm, form, sound and in many cases the lyrics. Each of these categories might have 10, 30, 50 elements. This is how a melody ends up being described by about 400 different genes.

This analysis of the songs is made by a team of analysts, musicians who have studied music for at least 4 years and have passed a music theory exam and completed 40 hours of training. This training was developed with the help of Nolan and is designed to make sure that the analysts are consistent with subjective matters, for example: how ”emotionally intense” on a scale from 1 to 5 is the solo part in a given song [8]. Even with the training, about 10% of the songs are analyzed a second time by a senior analyst and any difference in the opinion over a genes is flagged and reviewed [9].

Due to the time requirements of the analysis (20 to 30 minutes for a 4 minute long song), 5 years and 30 music theory experts were needed to built a database that can be useful. In May 2006, this database contained over 400000 analyzed songs from more than 20000 contemporary artists [10] which was increased to 700000 songs from more than 80000 artists till October 2009 with the estimation that about 10000 new song are added monthly to the database [8].

For this database, an interface was created and made available online under the name Pandora Internet Radio. The Pandora player chooses a succession of songs from its database after the user creates a channel. This is done by entering the name of a song or an artist and hitting the create button. Then Pandora sorts a selection based on the characteristics of the song or artist [11]. This selection can be refined by pushing the buttons ”Thumb up” and ”Thumb down” on this interface, thus showing if the melody that is being played is liked or disliked. Moreover, it has a link saying: ”Why did you play this song?” which takes the listener to a page where some of the genetic elements of the song are presented.

This approach of using the genes of a melody to find similar ones is a novel one, compared to the traditional recommendation systems of the type: ”customers who liked this item also liked that one”, and ignoring the crowd, saying that the taste of your cool friends, your peers, the traditional music critics, big-label talent scouts and the latest influential music blog are all equally irrelevant, because that is cultural information not musical one. To minimize the cultural influence during song listening, Westergren initially wanted to hide the artist of the songs until the listener asked to see it, but in the end he abandoned this idea [8].

Unfortunately, due to licensing problems, the Pandora Internet Radio is available only in the United States. Still, the search part (for artists and/or songs) of the site is available everywhere, and for many songs some of the "genes" are presented too, probably the same ones that a United States user sees when presses the "Why did you play this song?" link. The number of genes varies from song to song, usually there are between 4-20.

For this project, only the genes available at the Pandora site were used (there are approximative 150 such genes describing alternative rock melodies), since more specific information is considered proprietary information and cannot be publically released.

Since Pandora has a huge number of songs in its database, we only investigate songs of type "Alternative Rock" and from albums that appeared between 2007 and 2009. The list of all alternative rock artists is available at Wikipedia [12], this is where the name of artists were taken, searched with Pandora, and for the melodies from albums between 2007 and 2009 the available genes were recorded into a formal context together with the title, artist, album and year the album appeared. Despite the limitations in style and time, there still are a huge number of songs, so, due to lack of time, the context is not complete yet. It contains 451 songs, from 69 albums from 47 artists, which means the beginning of the list of artists whose name starts with "C" in the list at Wikipedia. As specified before, we had to collect the data for this analysis over the Internet, since more information is considered proprietary information and could not be released for our research.

This data resulted in a context with 451 objects (the songs) and 173 attributes (the genes). Some of these attributes are many-valued (for example most attributes related to the vocal characteristic of a song are many valued: they can have the value male of female) while some are one-valued. To scale this context and to create different concept lattices, the Elba editor of the ToscanaJ [13] application was used. To do so, the context was put in an Access database table, to which Elba can connect. Because of the dimension of the data table, the visualization of the entire concept lattice is not very practical, since it has a very complex structure.

We have decided to focus on subsets of attributes, selecting subcontexts and investigating the corresponding conceptual hierarchies. Similar attributes have been grouped together, creating scales and lattices for these. This gave 23 different scales such as: Acoustic Instruments, Electric Instruments, Lyrics, Vocals, Roots, Influences and others. Unfortunately there are 31 attributes which does not appear in either of these scales, but there are attributes which appear in more than one (for example the attribute Acoustic Piano belongs to the concept lattices Acoustic Instruments and Piano).

4. Investigating the Music Genome

As stated above, we have started the analysis of the Music Genome using specific methods of Conceptual Knowledge Processing. For this, data was gathered in tables, named contexts, then several subcontexts were considered, covering topics of interest for this study. Conceptual hierarchies have been build, and a detailed discussion of the results has been performed in order to highlight the conceptual connections between genes and melodies.

The attribute logic has been investigated by means of computing the stem base for attribute implications for each of the subcontexts. Also, association rules have been mined.

In the following, we illustrate these methods on 3 different subcontexts.

4.1. **Acoustic Instruments Context.** There are 5 attributes that belong to the acoustic context: Acoustic Rhythm Piano, Use of Acoustic Piano, Acoustic Guitar Riffs, Acoustic Rhythm Guitars, Good Dose of Acoustic Guitar Pickin'. Only one of these attributes, the Use of Acoustic Piano is multivalued, it appears in the context either simply "Use of Acoustic Piano" (1 object) or "Subtle Use of Acoustic Piano" (18 objects). Since it is a multivalued attribute we need a scale to transform it. An Ordinal Scale has been chosen, since Subtle Use of Acoustic Piano is still Use of Acoustic Piano. Moreover, two more attributes have added, Piano and Guitar to show the two main categories the acoustic instruments belong to and so to highlight the background knowledge on acoustic instruments (see [1]).

The reduced context that contains only these attributes has 187 objects, and its concept lattice can be seen on Figure 2. The nodes on the figure represent concepts, the higher a node is, the more general the concept is. Normally, all nodes connected directly to the uppermost node should be placed on the same level, and so on, but this is impossible because of the high number of nodes and the length of the labels. Yet, it is clearly visible that the concept with label "Use of Acoustic Piano" is more general than the one with label "Subtle Use of Acoustic Piano", because the latter is below the former in the lattice. The numbers associated to the nodes represent the number of objects in the context that have the given attribute, and not the total number of objects that have that attribute or one of its more specific ones. This is why the above mentioned "Subtle Use of Acoustic Piano" has the number 18 attached and the "Use of Acoustic Piano" has only the number 1. The uppermost node represents the attributes that all objects have (in our case there is no such object), while the lower one represents the objects that have all the attributes (in our case this is also empty).

Figure 3 displays the stem base for the 8 attributes of the Acoustic Instruments subcontext.

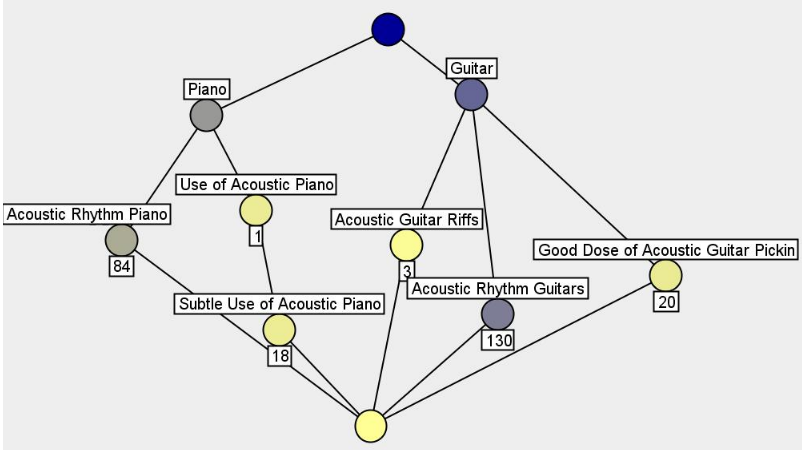
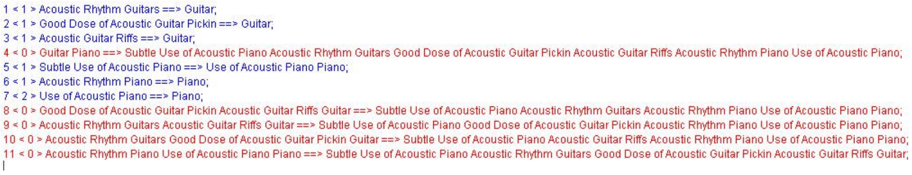


Figure 2. Concept lattice for attributes belonging to the Acoustic Instruments subcontext.

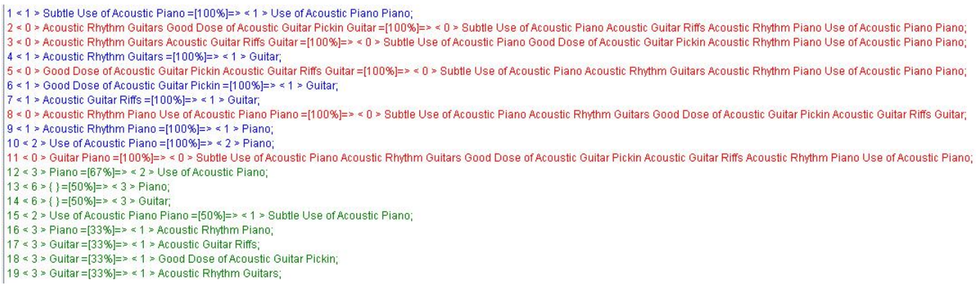


```
1 < 1 > Acoustic Rhythm Guitars ==> Guitar;
2 < 1 > Good Dose of Acoustic Guitar Pickin ==> Guitar;
3 < 1 > Acoustic Guitar Riffs ==> Guitar;
4 < 0 > Guitar Piano ==> Subtle Use of Acoustic Piano Acoustic Rhythm Guitars Good Dose of Acoustic Guitar Pickin Acoustic Guitar Riffs Acoustic Rhythm Piano Use of Acoustic Piano;
5 < 1 > Subtle Use of Acoustic Piano ==> Use of Acoustic Piano Piano;
6 < 1 > Acoustic Rhythm Piano ==> Piano;
7 < 2 > Use of Acoustic Piano ==> Piano;
8 < 0 > Good Dose of Acoustic Guitar Pickin Acoustic Guitar Riffs Guitar ==> Subtle Use of Acoustic Piano Acoustic Rhythm Guitars Acoustic Rhythm Piano Use of Acoustic Piano Piano;
9 < 0 > Acoustic Rhythm Guitars Acoustic Guitar Riffs Guitar ==> Subtle Use of Acoustic Piano Good Dose of Acoustic Guitar Pickin Acoustic Rhythm Piano Use of Acoustic Piano Piano;
10 < 0 > Acoustic Rhythm Guitars Good Dose of Acoustic Guitar Pickin Guitar ==> Subtle Use of Acoustic Piano Acoustic Guitar Riffs Acoustic Rhythm Piano Use of Acoustic Piano Piano;
11 < 0 > Acoustic Rhythm Piano Use of Acoustic Piano Piano ==> Subtle Use of Acoustic Piano Acoustic Rhythm Guitars Good Dose of Acoustic Guitar Pickin Acoustic Guitar Riffs Guitar;
```

Figure 3. List of implications for the attributes in the Acoustic instruments subcontext.

Remark 1. This base contains 11 implications 6 written in blue and 5 in red. The blue ones represent implications that have objects in the context that support that rule premise. These 6 implications are obvious, they follow directly from the way the scales for this context were defined. For example the fifth implication (Subtle Use of Acoustic Piano → Use of Acoustic Piano, Piano) follows from the ordinal scale we used for scaling the many-valued attribute. The red color marks implications which have no objects in the context that support the implication. This usually means that the set of objects from the premise does not occur together in the context. Such implications usually contains all the attributes. For example the first red implication is: Guitar, Piano → Subtle Use of Acoustic Piano, Acoustic Rhythm Guitars, Good Dose of Acoustic Guitar Pickin, Acoustic Guitar Riffs, Acoustic Rhythm Piano, Use of Acoustic Piano; so, the first two attributes imply all the rest. This obviously is an implication without objects because there is no instrument which would be Guitar and Piano in the same time.

Figure 4 displays a list of association rules with minimal support 0 and confidence 30%.



```
1 < 1 > Subtle Use of Acoustic Piano =[100%]=> < 1 > Use of Acoustic Piano Piano;
2 < 0 > Acoustic Rhythm Guitars Good Dose of Acoustic Guitar Pickin Guitar =[100%]=> < 0 > Subtle Use of Acoustic Piano Acoustic Guitar Riffs Acoustic Rhythm Piano Use of Acoustic Piano Piano;
3 < 0 > Acoustic Rhythm Guitars Acoustic Guitar Riffs Guitar =[100%]=> < 0 > Subtle Use of Acoustic Piano Good Dose of Acoustic Guitar Pickin Acoustic Rhythm Piano Use of Acoustic Piano Piano;
4 < 1 > Acoustic Rhythm Guitars =[100%]=> < 1 > Guitar;
5 < 0 > Good Dose of Acoustic Guitar Pickin Acoustic Guitar Riffs Guitar =[100%]=> < 0 > Subtle Use of Acoustic Piano Acoustic Rhythm Guitars Acoustic Rhythm Piano Use of Acoustic Piano Piano;
6 < 1 > Good Dose of Acoustic Guitar Pickin =[100%]=> < 1 > Guitar;
7 < 1 > Acoustic Guitar Riffs =[100%]=> < 1 > Guitar;
8 < 0 > Acoustic Rhythm Piano Use of Acoustic Piano Piano =[100%]=> < 0 > Subtle Use of Acoustic Piano Acoustic Rhythm Guitars Good Dose of Acoustic Guitar Pickin Acoustic Guitar Riffs Guitar;
9 < 1 > Acoustic Rhythm Piano =[100%]=> < 1 > Piano;
10 < 2 > Use of Acoustic Piano =[100%]=> < 2 > Piano;
11 < 0 > Guitar Piano =[100%]=> < 0 > Subtle Use of Acoustic Piano Acoustic Rhythm Guitars Good Dose of Acoustic Guitar Pickin Acoustic Guitar Riffs Acoustic Rhythm Piano Use of Acoustic Piano;
12 < 3 > Piano =[67%]=> < 2 > Use of Acoustic Piano;
13 < 6 > { } =[50%]=> < 3 > Piano;
14 < 6 > { } =[50%]=> < 3 > Guitar;
15 < 2 > Use of Acoustic Piano Piano =[50%]=> < 1 > Subtle Use of Acoustic Piano;
16 < 3 > Piano =[33%]=> < 1 > Acoustic Rhythm Piano;
17 < 3 > Guitar =[33%]=> < 1 > Acoustic Guitar Riffs;
18 < 3 > Guitar =[33%]=> < 1 > Good Dose of Acoustic Guitar Pickin;
19 < 3 > Guitar =[33%]=> < 1 > Acoustic Rhythm Guitars;
```

FIGURE 4. List of association rules for the attributes in the Acoustic instruments subcontext.

Remark 2. There are 19 association rules out of which the first 11 correspond to the 11 implications from the other Figure, which is normal, since every implication is an association rule with a 100% support. The red color shows again association rules with no objects, blue color the association rules which are also implications (they have 100% support) while the green rules represents the so called *not strict rules* which occur only for some percent of the objects (their support is less than 100%). The support of every rule is immediately after the premise of the rule, between square brackets. Rules number 13 and 14 have as premise {}, which means that it holds for every object from the context. These show that there is a 50% chance that an object has the Piano or Guitar attribute. For example, the last three implications show that if an song has the attribute Guitar there is a 33% chance that is has either Acoustic Guitar Riffs or Good Dose of Acoustic Guitar Pickin or Acoustic Rhythm Guitars. This is true, because the attribute Guitar appears only if one of these three attribute is present. On the other hand, probably the combinations of the three attributes have a too small support.

4.2. **Electric Instruments Context.** The Electric Instruments context, although similar to the acoustic one, is a little more complex, since from the original context 7 attributes have been considered in this group: Electric Guitars, Electric Guitar Effects, Electric Guitar Riffs, Electric Guitar Solo, Electric Guitar Wall-o-sound, Electric Rhythm Guitars and Electric Pianos. The multi-valued attributes are the following: Electric Guitar Riffs (with values Electric Guitar Riffs and Dirty Electric Guitar Riffs), Electric Rhythm Guitars (with values Electric Rhythm Guitars and Heavy Electric Rhythm Guitars) and Electric Guitar Solo (with values of Electric Guitar Solo and Dirty Electric Guitar Solo). In all three cases an Ordinal scale was used. Besides this, all objects with attributes Electric Rhythm Guitars and Heavy Electric Rhythm Guitars were also assigned the attribute Electric Guitars because it seemed logic to do so. Similarly to the case with the acoustic instruments two more attributes (Guitar and Piano) were introduced in order to express a certain amount of background knowledge, but in this case there is only one attribute that falls in the Piano category (Electric Pianos) this is why that node has two labels. This sub-context contains 266 objects and the corresponding concept lattice can be seen on Figure 5.

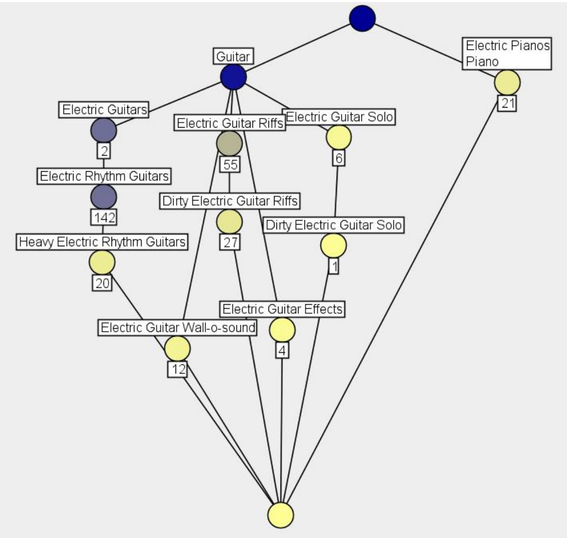
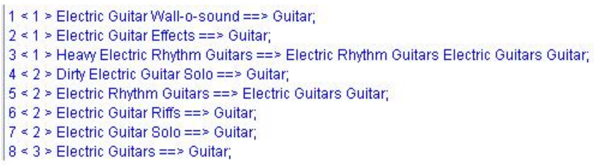


Figure 5. Concept lattice for attributes belonging to the Electric Instruments subcontext.

Figure 6 displays the stem base for the 11 attributes of the Electric Instruments subcontext.

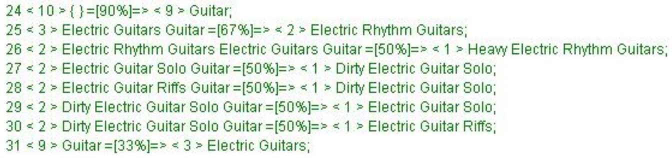


```
1 < 1 > Electric Guitar Wall-o-sound ==> Guitar;
2 < 1 > Electric Guitar Effects ==> Guitar;
3 < 1 > Heavy Electric Rhythm Guitars ==> Electric Rhythm Guitars Electric Guitars Guitar;
4 < 2 > Dirty Electric Guitar Solo ==> Guitar;
5 < 2 > Electric Rhythm Guitars ==> Electric Guitars Guitar;
6 < 2 > Electric Guitar Riffs ==> Guitar;
7 < 2 > Electric Guitar Solo ==> Guitar;
8 < 3 > Electric Guitars ==> Guitar;
```

Figure 6. List of implications for the attributes in the Electric instruments subcontext.

Remark 3. In this subcontext there are 8 implications, all of them are blue, so all of them have objects in the context which support them. All of these implications follow from the way the scales were defined and all of them have as conclusion the attribute Guitar. The third and the fifth implication also has in the conclusion the attributes Electric Rhythm Guitars, Electric Guitars and Electric Guitars, respectively. This is again obvious because an object with the attribute Heavy Electric Rhythm Guitars obviously has also the attributes Electric Rhythm Guitars and Guitar.

Figure 7 displays a list of association rules with minimal support 0 and confidence 30%.



```
24 < 10 > { } =[90%]=> < 9 > Guitar;
25 < 3 > Electric Guitars Guitar =[67%]=> < 2 > Electric Rhythm Guitars;
26 < 2 > Electric Rhythm Guitars Electric Guitars Guitar =[50%]=> < 1 > Heavy Electric Rhythm Guitars;
27 < 2 > Electric Guitar Solo Guitar =[50%]=> < 1 > Dirty Electric Guitar Solo;
28 < 2 > Electric Guitar Riffs Guitar =[50%]=> < 1 > Dirty Electric Guitar Solo;
29 < 2 > Dirty Electric Guitar Solo Guitar =[50%]=> < 1 > Electric Guitar Solo;
30 < 2 > Dirty Electric Guitar Solo Guitar =[50%]=> < 1 > Electric Guitar Riffs;
31 < 9 > Guitar =[33%]=> < 3 > Electric Guitars;
```

Figure 7. List of association rules for the attributes in the Electric instruments subcontext.

Remark 4. The list of association rules contains 8 rules, but the rules from the figure are only part of the list of rules generated by ConExp, which can also be seen by the numbers in front of them: they start from 24, which means there are at least 24 other association rules. A part of the rules not shown here are the implications from the figure above and probably there are also rules with support less than 30%. The first rule says that in 90% of the cases a song has the guitar attribute. This is based on the fact that out of the 10 attributes in this subcontext, 9 are related to guitars and only one to piano.

4.3. **Piano Context.** The third context is the Piano context which contains 6 attributes (Acoustic Rhythm Piano, Acoustic Piano, Subtle Pianos, Electric Pianos, Mellow Piano Timbre, Prominent Rhythm Piano Part) out of which only the Acoustic Piano attribute is multi-valued with values Acoustic Piano and Subtle Acoustic Piano. The scaling was done again using the Ordinal scale, since Subtle Acoustic Piano is still Acoustic Piano. Some of the attributes appeared already in the other two contexts, which is normal, because one attribute can belong to several subcontexts. Similar to the previous two cases, two attributes were also introduced: Acoustic and Electric, to denote the two main categories where a piano can belong to. Still, in this case there are 3 attributes that belong to neither of these categories, the Subtle Pianos, Mellow Piano Timbre and Prominent Rhythm Piano Part does not say what kind the piano is of. Another possibility would have been to include these attributes into both categories. The subcontext contains 124 objects and its concept lattice is presented on Figure 8.

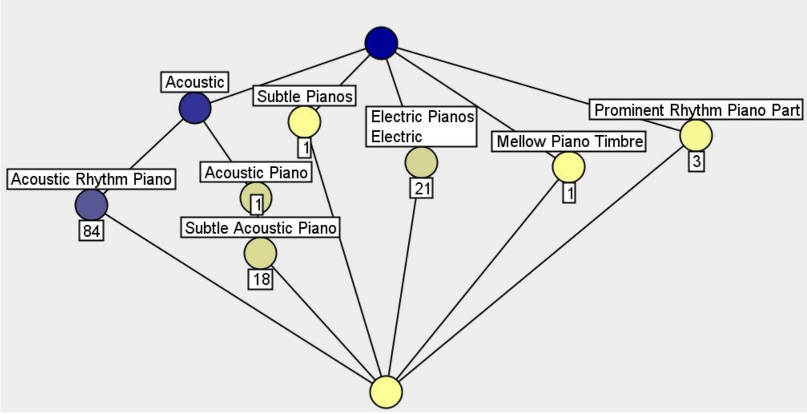


FIGURE 8. Concept lattice for attributes belonging to the Piano subcontext.

5. Towards an FCA Based Music Environment

As it has been stated before, this is a first attempt to investigate the structure of the Music Genome using Conceptual Knowledge Processing methods. Lack of time, the huge dataset and the necessity to optimize the software tool, but also limited space for this presentation, convinced us to restrict ourselves to some relevant topics.

There is more work to be done. We would like to analyze the entire set of attributes, to perform attribute exploration and to compute the stem base over the entire attribute set. Also mining all association rules could display interesting conclusions about how modern music is structured.

Several questions remain unanswered, for example how can we define a similarity measure in order to cluster similar songs?

An FCA based music environment is a research program in this field. The user should be able to choose freely the genomic subset of his convenience, to navigate through conceptual hierarchies, browsing in an intelligent and always conceptual driven way the songs, experiencing music in an entire new way.

References

[1] Carpineto, C., Romano, G.: *Concept Data Analysis, Theory and Applications*, Wiley and Sons, 2004.

[2] Hindemith, Paul, *Traditional Harmony* (second edition), Associated Music Publishers, Inc. New York (1944).

[3] Helmholtz, Hermann L. F., *On the Sensations of Tone*, Fourth (and last) German edition (1877) translated by Alexander J. Ellis as the second English edition (1885), revised, corrected annotated and with additional appendix by the translator, Dover Publications (1954).

[4] Mazzola, Guerino, *Gruppen und Kategorien in der Musik*, Heldermann, Berlin 1985

[5] Mazzola, Guerino, *Geometrie der Toene*, Birkhaeuser, Basel 1990

[6] Mazzola, Guerino, *The Topos of Music-Geometric Logic of Concepts, Theory, and Performance*, Birkhaeuser, Boston-Basel, 1999.

[7] *The Music Genome Project on Pandora:* http:www.pandora.commgp.shtml

[8] Rob Walker, *The Song Decoders*, The New York Times Magazin, October 14, 2009 (available at: http:www.nytimes.com20091018magazine18Pandora-t.html)

[9] Steven Barrie-Anthony (Times Staff Writer), *That Song Sounds Familiar*, Los Angles Times Calendarlive.com, February 3, 2006 (available at: http:msl1.mit.edufurdlogdocslatimes2006-02-03_latimes_pandora_music.pdf)

[10] John Joyce, *Pandora and the Music Genome Project*, Scientific Computing (available at: http:www.scientificcomputing.compandora-and-the-music-genome-project.aspx)

[11] Michael Castelluccio, *The Music Genome Project*, Financial Times, December 2006

[12] List of Alternative Rock Artists: http:en.wikipedia.orgwikiList_of_alternative_rock_artists

[13] ToscanaJ: http:toscanaj.sourceforge.net

[14] Wille, R., *Restructuring lattice theory: an approach based on hierarchies of concepts.* In I. Rival (ed.) *Ordered sets*, Reidel, Dordrecht, Boston 1982, 445–470.

[15] Wille, R.: *Methods of Conceptual Knowledge Processing*, ICFCA 2006, LNAI 3874, Springer, pp. 1-29, 2006.

Faculty of Mathematics and Computer Science, Babes-Bolyai University, Cluj-Napoca, Romania

E-mail address: mzsi0142@scs.ubbcluj.ro

E-mail address: csacarea@math.ubbcluj.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LV**, Number 1, 2010

CONCEPTUAL GRAPH DRIVEN DESIGN FOR CONCEPTUAL DIGITAL DOSSIERS

FLORINA MUNTENESCU, CHRISTIAN SĂCĂREA, AND VIORICA VARGA

Abstract. Conceptual digital dossiers have been proposed as a new paradigm for digital content management. This paper is the first in a series of works towards a complete functional conceptual digital dossier. We have focussed ourselves on the underlying knowledge base, using Conceptual Graphs for representing judgements. The paradigm is that of Conceptual Knowledge Processing. We discuss the design of a digital dossier using conceptual graphs, grounded on Contextual Logic and Formal Concept Analysis.

1. Introduction

In the Information Age, we are all facing a common problem: a very large amount of information in almost every part of our life. Retrieving information, extracting valuable knowledge, and knowledge management are not only a research topic, but more and more a necessity. Digital content is either fragmented over the Web or stored in archives fragmented in several directories on personal computers. This situation restricts by time and space the presentation of diverse information in an interconnected way, usable for research, entertainment or education. As a consequence, theory and practice developed rapidly, many applications and projects are created to build usable collections of information and knowledge, networking people, experience and competence with the aim of a consistent approach towards knowledge management for digital content. Digital content is usually understood as digital information, which may take the form of text, such as documents, multimedia files, such as audio or video files, or any other file type which follows a content life cycle

Received by the editors: January 17, 2009.

2010 *Mathematics Subject Classification.* 68P15, 03G10.

1998 *CR Categories and Descriptors.* code H.2.1 [**Database Management**]: Logical Design – *Normal Forms*; F.4.m [**Mathematical Logic and Formal Languages**]: Miscellaneous.

Key words and phrases. Conceptual Knowledge Processing, Digital Dossiers, Contextual Logic, Database Theory, Conceptual Graphs.

which requires management. Thus, content management is defined to be a set of processes and technologies that support the evolutionary life cycle of digital information. We propose the use of a digital dossier, as a digital archive containing information about topics of interest, serving as an information source for different categories of users, from experts to students or just visitors of our digital archive.

2. Digital Dossier

A digital dossier is defined to be a collection of information that is available in some particular area of interest (e.g. in Arts) presented in an easily accessible way ([8]). The digital dossier should fulfill a list of requirements making it suitable for use both for experts or regular users. For example, [19] states that a digital dossier should be inter twinkled, that means that all information objects in the virtual environment that are already related in the existing information, must also be related in the virtual environment by hyper linking.

Since a digital dossier is a structured collection of digital documents, the use of state of the art methods to discover, process and represent the captured knowledge lies at hand. The methods we suggest are those of Conceptual Knowledge Processing, as they were stated in [18]. But knowledge consists not only of a list of concepts, acquiring knowledge is a complex process where the subsequent logic of data is stepwise unveiled. Formalizing this conceptual logic is a complex task and we follow the outlines stated in [17]. Conceptual Graphs are understood as graphically structured judgements. The Contextual Judgement Logic mathematizes Sowa's conceptual graphs in order to represent information by mathematical structures called concept graphs and semantically interpreted as power context families. According to [17], the *conceptual content* of a concept graph is viewed as the information directly represented by the graph together with the information deducible from the direct information by object and concept implications coded in the power context family.

This approach strongly suggests the use of Sowa's conceptual graphs in order to represent the logical structure of a digital dossier, to represent relationships of the underlying data and to perform queries (see [9]).

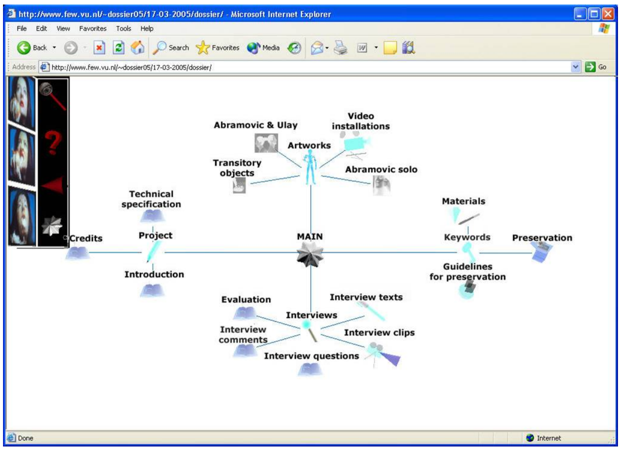
This paper presents a conceptual driven design for digital dossiers based on the theory of Conceptual Graphs as part of Contextual Judgement Logic, as a first step towards a new paradigm for digital documents management, *Conceptual Digital Dossiers.*

3. Conceptual Digital Dossiers

A digital dossier is a digital archive where convenient digital documents are stored. Since visualization and retrieval of both documents and knowledge is important, this archive must be enhanced by a visualization environment, a search engine and processing tools. The concept Digital Dossier was introduced by A. Eliens and his team from the Vrije Universiteit Amsterdam, Holland, in order to describe a case study for creating a digital archive for the Serbian-Dutch contemporary artist Marina Abramovic [4].

The digital dossier presents itself as a digital archive in 3D space, containing information about the artworks of the performance artist Marina Abramovic by presenting media content and relational structures. This digital dossier introduces innovative features with respect to navigation and presentation in 3D environments. For navigation, a graph has been designed, linking multimedia elements in a structured hierarchy. The navigation is dynamic, animation being also implemented, in order to improve visualization of the entire digital content. The selected information determines the presented hierarchy and presented parent - child relationships in this graph. Every node of the graph is an information object, the edges are relationships between different information objects. The presentation of media content is made over a set of three windows.

Example 1. A screenshot of the digital dossier of the contemporary artist Marina Abramovic.



http://www.few.vu.nl/~dossier05/17-03-2005/dossier/ - Microsoft Internet Explorer

Address http://www.few.vu.nl/~dossier05/17-03-2005/dossier/

Done

Figure 1. abramovic dossier

The term conceptual means an integrated view of the knowledge gathered in a digital dossier. Hence, we propose a new approach towards storage of digital information. More than a simple digital archive, a conceptual digital dossier is a knowledge atlas, including data, information, knowledge, route maps for navigating between knowledge clusters, enhanced with a logical structure provided by a coherent formal mathematical theory, Formal Concept Analysis. This logical structure enables a process of extracting, acquiring and knowledge processing which is conceptual driven and human centered.

Visualization of digital content is one of the major contributions of the team of A. Eliens, and so is also the concept of a digital dossier, which seems to be most appropriate to describe this approach. Nevertheless, if the content is complex, one need a more standardized approach in order to retrieve information, to mine data, and to acquire knowledge. This approach is grounded on Conceptual Knowledge Processing, and especially for this paper on Contextual Judgement Logic.

The steps towards building a conceptual digital dossier have been stated in [8]:

1. Collect digital content;
2. Organize digital information in data sheets, i.e., formal contexts;
3. Build conceptual scales;
4. Scale many-valued contexts;
5. Perform an attribute exploration, if necessary;
6. Compute stem base;
7. Compute conceptual hierarchies;
8. Create knowledge base;
9. Create a proper ontology;
10. Compute 3D cone tree hierarchy;
11. Implement visualization tools;
12. Implement conceptual search engine.

This paper presents a prerequisite in order to build a conceptual digital dossier. Based on the conceptual graphs theory and their mathematization as concept graphs, a conceptual graph driven design for the underlying database is proposed. We represent both the structure of this database and the performed queries as conceptual graphs.

4. Conceptual Graphs and Contextual Logic

Contextual Logic has been introduced with the aim to support knowledge representation and knowledge processing. It is grounded on the traditional understanding of logic as the doctrine of the forms of human thinking ([18]). I. Kant explained this understanding of logic as "the theory of the three main essential functions of thinking: concepts, judgements and conclusions". The doctrine of concepts has been successfully mathematized by Formal Concept Analysis.

Conceptual graphs can be understood as formal judgements. A conceptual graph is a labeled graph that represents the literal meaning of a sentence. Conceptual graphs express 'meaning in a form that is logically precise, humanly readable, and computationally tractable' ([12]). They serve as an intermediate language for translating computer-oriented formalisms to and from natural languages. With their graphic representation, they serve as a readable, but formal design and specification language.

In particular, they are bipartite graphs that consist of concept nodes, which bear references as well as types of the references. The concept boxes are connected by edges, which are used to express different relationships between the referents of the attached concept boxes. Sowa provides rules for formal deduction procedures on conceptual graphs; hence the system of conceptual graphs offers a formalization of conclusions too.

As Formal Concept Analysis provides a formalization of concepts, and as conceptual graphs offer a formalization of judgements and conclusions, a convincing idea is to combine these approaches to gain a unified formal theory for concepts, judgements and conclusions, i.e., a formal theory of elementary logic. In [16], Wille marked the starting point for a such a theory. He provided a mathematization of conceptual graphs, the types being interpreted by formal concepts of a so-called power context family. The resulting graphs are called concept graphs. They form the mathematical basis for contextual logic.

The relations in a conceptual graph are understood as concepts of suitable chosen formal contexts. By this understanding, derivation of judgements (represented by conceptual graphs) is made possible. We use for this the so-called *power context family.*

Definition 1. A *power context family* is a sequence $\vec{\mathbb{K}} := (\mathbb{K}_0, \mathbb{K}_1, \mathbb{K}_2, \dots)$ of formal contexts $\mathbb{K}_j := (G_j, M_j, I_j)$ with $G_j \subseteq (G_0)^j$ for $j \in \mathbb{N} \setminus \{0\}$. The formal concepts of $\mathbb{K}_j$ with $j \in \mathbb{N} \setminus \{0\}$ are called *relation concepts*, because their extents represent k-ary relations on the object set G_0.

The concepts of $\mathbb{K}_0$ represent the concepts in the boxes and $\mathbb{K}_1, \mathbb{K}_2, \mathbb{K}_3, \dots$ yield the concepts of relations of arity $k = 1, 2, 3, \dots$.

Wille stated in [18] that this method can be effectively applied to develop information systems based on power context families representing the relevant knowledge. The central idea of these information systems is to present to the user, who has inputed his constraints, a conceptual graph representing all relevant informations. The conceptual graph should be understood as a logical structure which may have different graphical representations useful for quite different purposes.

A similar approach was presented in [15], focussed on the classical view of Conceptual Graphs as bipartite graphs. A software tool for interactive database access has been developed. The basic idea of using Conceptual Graphs as query interface to relational databases has been stated by J. Sowa. The CGDBInterface software, is a graphical tool to query an existing relational database. The actual version of application CGDBInterface can connect to MS SQL Server, Oracle and mySQL databases. Having a valid user name and password, the conceptual graph of the relational database is first displayed. The table names and their attributes become concepts. The attributes of a table are linked by conceptual relations to the concept corresponding to the table they belong to. The relationships between tables are designed by conceptual relations too.

Example 2. We have build a database for impressionist painters as underlying structure for a digital dossier. Figure 2 displays the conceptual graph of this digital dossier.

5. Conceptual Graph Driven Design

To exemplify the above stated ideas about the design of a digital dossier grounded on the Conceptual Knowledge Processing paradigm, especially Contextual Logic and Conceptual Graphs, we have created a digital dossier for impressionist art. We have chosen this topic, not only for its importance in the history of Art but also because it offers a large area of background knowledge.

We started from the types of impressionism (impressionism, post-impressionism, neo-impressionism) and we continued by adding artists for each of these types. Next, we wanted to see what do these artists have in common so, we included in the database their teachers and inspirations.

The impressionists work offered us the largest amount of information: for each of the artist we added a large number of works, accompanied, apart from media content, by data related to the techniques used (oil, pastel, gouache etc), motifs that appear in each work of art or collections and the museums in which they are displayed.

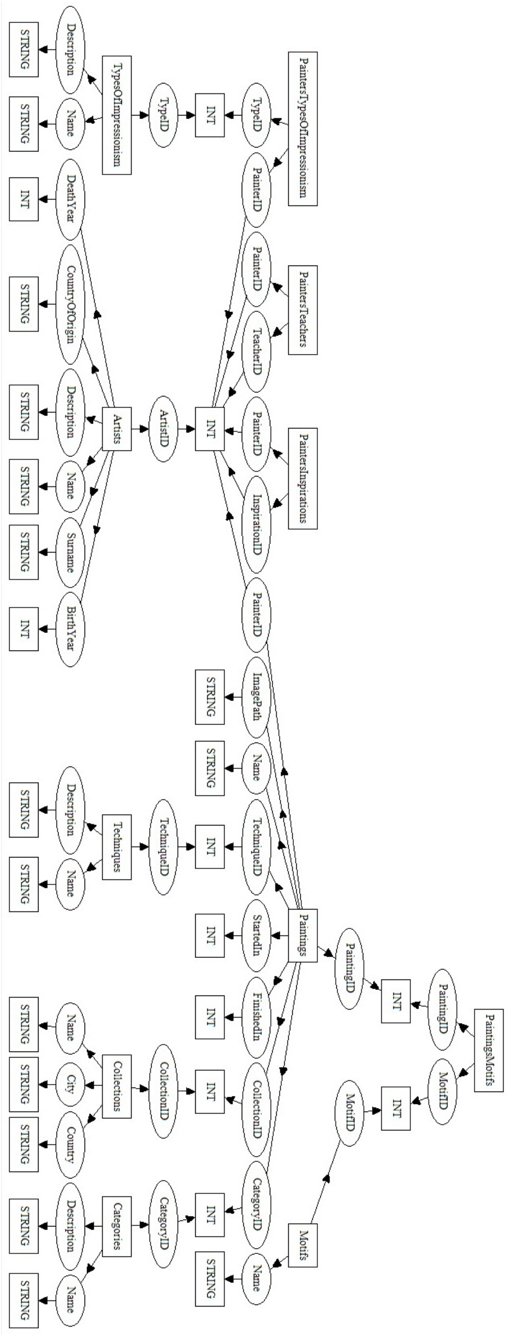


Figure 2. Conceptual graph for impressionism digital dossier

This approach was grounded on the ideas stated above of what a digital dossier should be, not just a digital archive, but an easy method of discovering the knowledge about impressionism art and artists. The paradigm of Conceptual Graphs as representations of judgements was used.

As a comparison with the digital archive for Marina Abramovic ([4]), that offers knowledge only on related concepts about one artist's work, we are trying to go up one level and see not only each painter as a concept, part of a power context family, but also the whole impressionist art.

Therefore, using the methods of data analysis offered by Formal Concept Analysis, we have built the formal context of impressionism, the painters representing the class of objects, and data about inspirations or teachers are classes of attributes. In the same time, each painter can be considered to give rise to a formal context itself, where the objects are its artworks and the attributes are the techniques, motifs, museums, collections and others.

In order to discover more knowledge about impressionists and their work, knowledge which is encoded in the digital dossier we have created, we make use again of the idea of power context family. For example, if we are interested in impressionist painters that appear in a certain collection, we consider a context having as object set pairs of impressionist painters and collections and as attribute set paintings, motifs, museums and others.

In the following we give some Query Conceptual Graphs examples referring the created digital dossier for impressionist art.

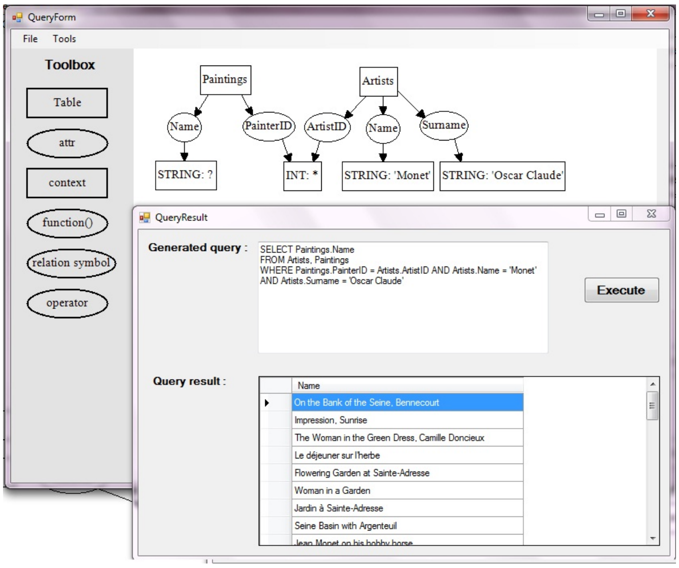
Example 3. Figure 3 presents the query as a conceptual graph about paintings title of Oscar Claude Monet.

We can see both the SQL query, the corresponding conceptual graph and the list of results.

Example 4. The query about impressionist painters using the *oil on canvas* technique is displayed in Figure 4.

Example 5. A query about impressionist painters who have painting on family motif can be seen in Figure 5. The query gives for these artist the number of their paintings.

The corresponding conceptual graph is nested, the query being more complex, the concept $T1$ at the upper left corner is considered the hypostatic abstraction of the concepts and relations which are included in that rectangle.



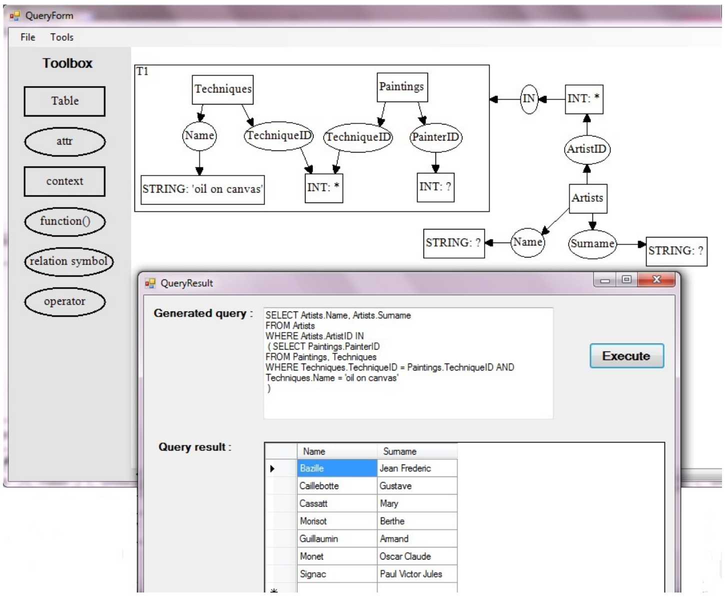
Generated query :

```
SELECT Paintings.Name
FROM Artists, Paintings
WHERE Paintings.PainterID = Artists.ArtistID AND Artists.Name = 'Monet'
AND Artists.Sumame = 'Oscar Claude'
```

Query result :

Name
On the Bank of the Seine, Bennecourt
Impression, Sunrise
The Woman in the Green Dress, Camille Doncieux
Le déjeuner sur l'herbe
Flowering Garden at Sainte-Adresse
Woman in a Garden
Jardin à Sainte-Adresse
Seine Basin with Argenteuil
Jean Monet on his hobby horse

FIGURE 3. Monet paintings query



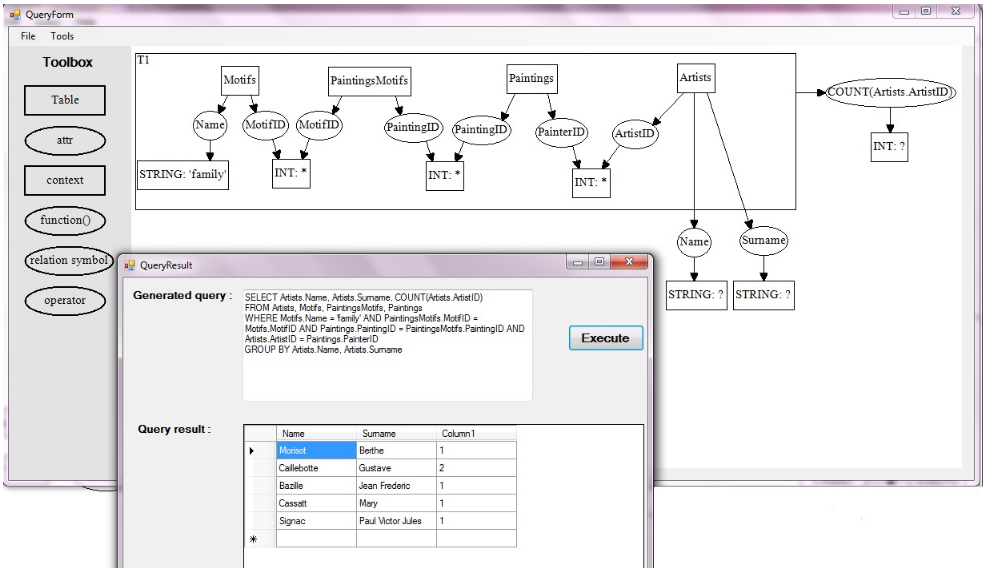
Generated query :

```
SELECT Artists.Name, Artists.Sumame
FROM Artists
WHERE Artists.ArtistID IN
 ( SELECT Paintings.PainterID
FROM Paintings, Techniques
WHERE Techniques.TechniqueID = Paintings.TechniqueID AND
Techniques.Name = 'oil on canvas'
 )
```

Query result :

Name	Sumame
Bazille	Jean Frederic
Caillebotte	Gustave
Cassatt	Mary
Morisot	Berthe
Guillaumin	Armand
Monet	Oscar Claude
Signac	Paul Victor Jules

FIGURE 4. Painting techniques query



Generated query :

```
SELECT Artists.Name, Artists.Surname, COUNT(Artists.ArtistID)
FROM Artists, Motifs, PaintingsMotifs, Paintings
WHERE Motifs.Name = 'family' AND PaintingsMotifs.MotifID =
Motifs.MotifID AND Paintings.PaintingID = PaintingsMotifs.PaintingID AND
Artists.ArtistID = Paintings.PainterID
GROUP BY Artists.Name, Artists.Surname
```

Query result :

	Name	Surname	Column1
▶	Morisot	Berthe	1
	Caillebotte	Gustave	2
	Bazille	Jean Frederic	1
	Cassatt	Mary	1
	Signac	Paul Victor Jules	1
*			

FIGURE 5. Impressionist painters query

6. CONCLUSION

Conceptual Digital Dossiers are proposed as a new paradigm for digital content management. We have presented here the prerequisites for a conceptual knowledge processing and representation management. Based on state of the art ideas and methods, conceptual graphs representation appears to be useful, both for design and for the end user, in order to extract valuable knowledge in a minimum amount of time.

There is still more work to be done. The knowledge base once structured, specific Formal Concept Analysis knowledge acquisition tools should be used. Conceptual hierarchies should be displayed or used to support navigation. There is a need for a navigation environment, in order to efficiently browse the digital content. Simultaneously, we need a visualization environment for a proper representation of the stored data, its connections and the underlying knowledge.

REFERENCES

[1] Carpineto, C., Romano, G.: *Concept Data Analysis, Theory and Applications*, Wiley and Sons, 2004.

[2] Dau, F.: *The Logic System of Concept Graphs with Negation And Its Relationship to Predicate Logic*, LNCS, Vol. 2892, Springer Berlin / Heidelberg (2003)

[3] Dau, F., Hereth, J. C.: *Nested Concept Graphs: Mathematical Foundations and Applications for Databases.* In: Ganter, B.; de Moor, A. (eds.): Using Conceptual Structures. Contributions to ICCS 2003. Shaker Verlag, Aachen, (2003), pp. 125-139.

[4] A. Eliens, C. van Riel, and Y. Wang, *Navigating media-rich information spaces using concept graphs - the abramovic dossier* in Proc. InSciT2006, 25-28 Oct. 2006, Merida, Spain.

[5] Hereth, J.: *Relational Scaling and Databases.* Proceedings of the 10th International Conference on Conceptual Structures: Integration and Interfaces LNCS 2393, Springer Verlag (2002) pp. 62-76

[6] Boksenbaum, C., Carbonneill, B., Haemmerle O., Libourel, T.: *Conceptual Graphs for Relational Databases* in Conceptual Graphs for Knowledge Representation., Guy, M. W., Moulin B., Sowa, J. F. eds. Lecture Notes in AI 699, Springer-Verlag, Berlin (1993).

[7] Peirce, C.S.: *The Simplest Mathematics*, in Collected Papers, CP4.235, CP4.227-323, 1902.

[8] C. Săcărea, *Conceptual Digital Dossier: Towards a new paradigm for digital content management*, Interdisciplinary New Media Studies Conference Proceedings, Cluj-Napoca 2009, pp. 107–111.

[9] C. Săcărea, V. Varga, *Conceptual Knowledge Processing for Databases. An Overview*, Studia Univ. Babeş- Bolyai, Informatica, Volume LIV, Number 2, 2009,pp. 59–70.

[10] Silberschatz, A., Korth, H. F.,Sudarshan, S.: *Database System Concepts*, McGraw-Hill, Fifth Edition, (2005)

[11] Sowa, J. F.: *Conceptual Graphs for a Database Interface.* In: IBM Journal of Research and Development, vol. 20, no. 4, (1976) pp. 336-357.

[12] Sowa, J. F.: *Conceptual Structures: Information Processing in Mind and Machine.* Addison Wesley Publishing Company Reading, (1984).

[13] Sowa, J. F.: *Conceptual graphs summary*, in Nagle, T. E.; Nagle, J. A.; Gerholz, L. and Eklund, P. W. (editors): Conceptual Structures: Current Research and Practice, Ellis Horwood, (1992), pp 3-51.

[14] Sowa, J. F.: *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA. (2000)

[15] Varga, V., Săcărea, C., Takács, A.: *A Software Tool for Interactive Database Access using Conceptual Graphs*, International Conference Knowledge Engineering Principles and Techniques, KEPT 2009, Cluj-Napoca, July 2-4.

[16] Wille, R.: *Conceptual Graphs and Formal Concept Analysis*, Lecture Notes In Computer Science; Vol. 1257, Proceedings of the Fifth International Conference on Conceptual Structures: Fulfilling Peirce's Dream, Springer Verlag (1997), pp 290 - 303.

[17] Wille, R.: *Conceptual Contents as Information - Basics for Contextual Judgement Logic*, Conceptual Structures for Knowledge Creation and Communication, ICCS 2003, LNAI 2746, Springer, pp. 1-15, 2003.

[18] Wille, R.: *Methods of Conceptual Knowledge Processing*, ICFCA 2006, LNAI 3874, Springer, pp. 1-29, 2006.

[19] Y. Wang, *The Presentation of Media-rich Collections of Culture Heritage in the Age of Digital Reproduction*, Master-Thesis, www.win.tue.nl/∼ywang/publications/yiwen-thesis.pdf

Babes-Bolyai University, Faculty of Mathematics and Computer Science, 400084 Cluj-Napoca, Romania

E-mail address: `florina.muntenescu@gmail.com`

E-mail address: `csacarea@math.ubbcluj.ro`

E-mail address: `ivarga@nessie.cs.ubbcluj.ro`

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LV**, Number 1, 2010

HANDWRITTEN DIGITS RECOGNITION USING NEURAL COMPUTING

CĂLIN ENĂCHESCU AND CRISTIAN-DUMITRU MIRON

Abstract. In this paper we present a method for the recognition of handwritten digits and a practical implementation of this method for real-time recognition. A theoretical framework for the neural networks used to classify the handwritten digits is also presented.

The classification task is performed using a Convolutional Neural Network (***CNN***). ***CNN*** is a special type of multi-layer neural network, being trained with an optimized version of the back-propagation learning algorithm. ***CNN*** is designed to recognize visual patterns directly from pixel images with minimal preprocessing, being capable to recognize patterns with extreme variability (such as handwritten characters), and with robustness to distortions and simple geometric transformations.

The main contributions of this paper are related to the original methods for increasing the efficiency of the learning algorithm by preprocessing the images before the learning process and a method for increasing the precision and performance for real-time applications, by removing the non useful information from the background.

By combining these strategies we have obtained an accuracy of ***96.76%***, using as training set the ***NIST*** (National Institute of Standards and Technology) database

Received by the editors: December 15, 2010.

2010 *Mathematics Subject Classification.* 68T05, 68T10, 62M45.

1998 *CR Categories and Descriptors.* I.7.5 [**Computing Methodologies**]: Document and text processing – *Optical character recognition (OCR)*; I.5.1 [**Computing Methodologies**]: Pattern Recognition – *Models – Neural Nets*.

Key words and phrases. Convolutional neural networks, supervised learning, handwritten digits recognition.

This paper has been presented at the International Conference Interdisciplinarity in Engineering (INTER-ENG 2009), Târgu-Mureş, Romania, November 12–13, 2009.

1. Introduction

The ***NIST*** database used to train the neural network [4] contains 60,000 images with distorted handwritten digits, 10,000 images being used for testing and considered as a reference benchmark for applications for handwritten digits recognition.

In order to demonstrate the efficiency of our proposed neural network we will compare the learning performances with a classical ***MLP*** (Multy Layer Perceptron) [5] neural network.

In order to measure the performances of our proposed neural network we will use two different types of neural networks architectures: the first one, a classical ***MLP*** neural network and the second an original ***CNN***, proposed in this paper.

2. The Convolutional Neural Network - CNN

CNN is a feed-forward neural network capable to extract topological properties from an image. It extracts features from the input raw image using the first hidden layer and classifies the patterns with the help of the final hidden layer.

The first two layers of the neural network can be considered as a trainable feature extractor [2]. We can add a trainable classifier to the feature extractor, considering 2 fully connected layers (a universal classifier) [2].

The number of hidden neurons contained in the hidden layers is variable and by varying this number we can control the learning capacity and the generalization capacity of the overall classifier [5].

The architecture of the ***CNN*** used to learn the ***NIST*** database is depicted in Figure 1.

The general processing strategy for a ***CNN*** is to extract simple features at a higher resolution, and then to convert them into complex features at a coarser resolution. Each convolutional hidden layer can reduce the resolution of the initial image by a factor of $(n-3)/2$. For this purpose, each neuron contained in a convolutional hidden layer will process a single region from the map of the previous layer. In this way, a neuron is functioning like an independent micro kernel, named ***convolutional kernel*** [2]. The convolutional kernel contains as inputs the corresponding region from the previous map to be processed.

The width of the ***convolutional kernel*** is chosen to be centered on a unit-pixel (odd size), to have sufficient overlap for not losing information (3 units

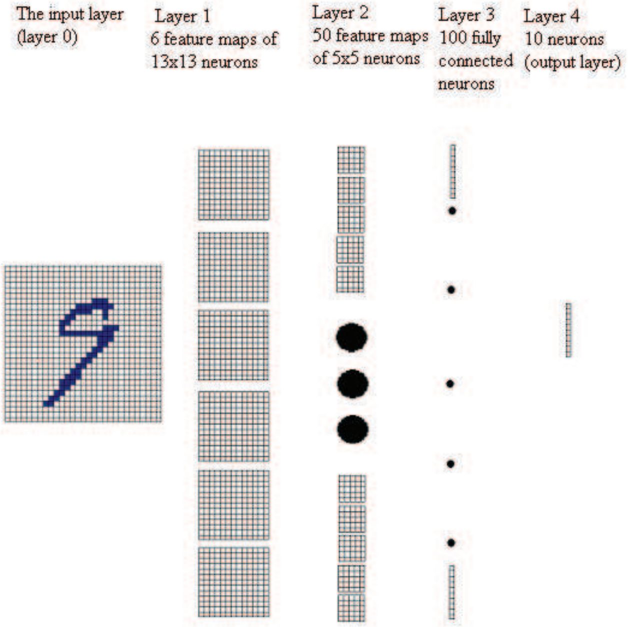


FIGURE 1. The architecture of the ***CNN***

are too small with only one unit overlap, 7 units represents a 70% overlap). A ***convolution kernel*** of size 5 is shown in Figure 2.

With no padding, a sub-sampling [2] of 2, and a kernel size of 5, each ***convolution layer*** reduces the feature size from n to $(n-3)/2$. Since the initial ***NIST*** database contains images of size 28×28, the nearest value which generates an integer size after 2 layers of convolution is 29×29.

The ***CNN*** architecture is composed of 4 layers. The first layer is a convolutional layer composed of 6 feature maps of 13×13 units. Each neuron processes only a region of 5×5 pixels from the input image, ignoring the rest of the information from the picture. The second layer is also a convolution layer that is processing a region of 5×5 outputs from each feature map of the previous layer, the other outputs being ignored. The third layer is composed of 100 neurons fully connected with the neurons from the second layer. The output layer is composed of 10 neurons, one neuron for each pattern that has to be classified.

A trainable weight is assigned to each connection, but all units of one feature map share the same weights. This characteristic is justified by the fact

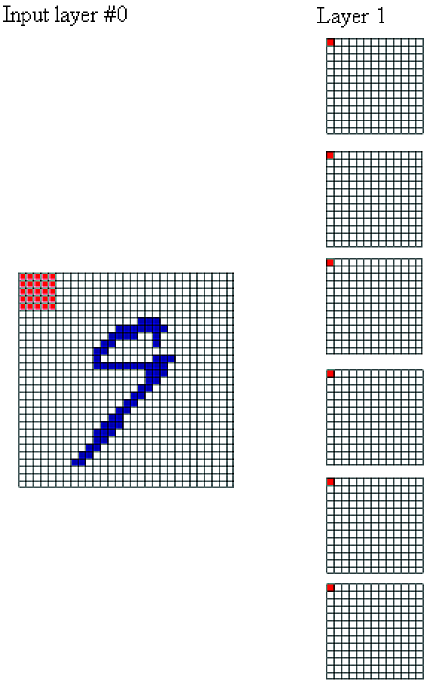


FIGURE 2. The inputs of the first neuron from each feature map in Layer1

that an elementary feature detector useful on one part of the image is unlikely to be useful for the entire image. Moreover, the weight sharing technique allows the reduction of the numbers of trainable parameters. For instance, ***CNN*** has only 132,750 trainable parameters out of 303,450 connections.

2.1. **Cutting out useless information.** Although ***CNN*** are known to be able to extract information from images with minimum preprocessing, the removal of the un-useful pixels from the background improves performances, not only during the learning process, but more important, during the use of the ***CNN*** in practice. One of the reasons is that we have the ***NIST*** training database that contains 60,000 images of handwritten digits but, all of them are centered and have the same width. In practice we will have to extract out digits from bigger images and resize them to 29 × 29 pixels, so we can present them as inputs to the neural network. Trying to cut out a digit from an image, resizing and centering the image, we can obtain the same results as with the digits contained in the ***NIST*** database. Another advantage of this strategy is obtained when an image of the digit is zoomed in and the relevant patterns are accentuated. In Figure 3 we have on the left a sample image from the ***NIST*** database and on the right the same image with the background removed.

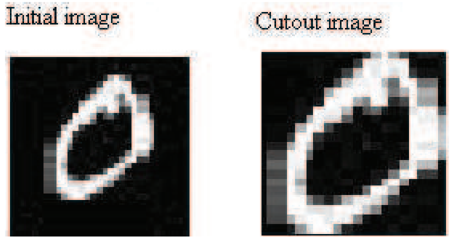


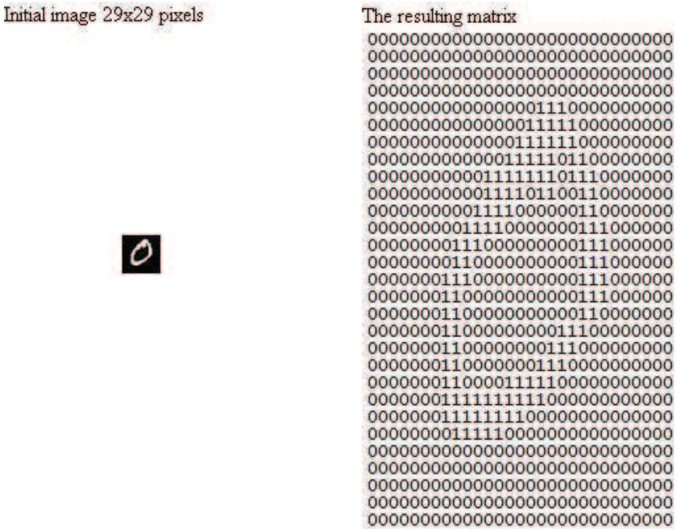
Figure 3. Background removal

2.2. **Pre-encoding of the training images.** The encoding of the images used in the learning process is a crucial improvement of the learning process because it dramatically decreases the time needed to pass one learning epoch [5]. The images are codified using the following rule: the useful information, the white pixels, are set to 1 and the background pixels are set to 0.

After this encoding process we will store the encoding images in a special training file. Before starting the learning process we will load all this samples from the training file into the internal memory of the computer. It is very important to run the learning process with the deactivated virtual memory option of the operation system - this will eliminate the need to use the external memory which is slow compared with the internal memory. An example of encoding is presented in Figure 4.

3. The learning process

The learning process was based on the ***NIST*** training database that contains 60,000 images of handwritten digits. Because the training set is huge, the duration of a single learning epoch took almost 73 minutes on a computer with Intel Dual Core processor - 1.86 GHz and with 3 GB of RAM. This long time span of a learning epoch was given by the fact that all of the images had to be loaded from the hard disk, a slow external memory device, compared with the internal memory. To reduce the duration of a learning epoch we



Initial image 29x29 pixels

The resulting matrix

```
00000000000000000000000000000
00000000000000000000000000000
00000000000000000000000000000
00000000000000000000000000000
00000000000000001110000000000
00000000000000011111000000000
00000000000000111111000000000
00000000000001111101100000000
00000000000111111101110000000
00000000000111101100110000000
00000000001111000000110000000
00000000011110000000111000000
00000000111000000000111000000
00000000110000000000111000000
00000001110000000000111000000
00000001100000000000111000000
00000001100000000000110000000
00000001100000000011100000000
00000001100000000111000000000
00000001100000001110000000000
00000001100001111100000000000
00000001111111111000000000000
00000001111111100000000000000
00000000111110000000000000000
00000000000000000000000000000
00000000000000000000000000000
00000000000000000000000000000
00000000000000000000000000000
00000000000000000000000000000
```

Figure 4. Encoding example for digit 0

decided to upload all the images from the ***NIST*** learning database into the internal memory (RAM).

By pre-encoding the learning images and by loading them directly into the RAM at the start of the learning process, we have reduced the time span of a learning epoch to approximately 52 minutes, which means an improvement of **40%** in performance. An important observation is to deactivate the virtual memory option before the start of the learning process, in order to be sure that the training dataset is not stored on the HDD, in the swap file of the operating system.

The initial values of the ***synaptic weights*** have an important effect on the learning capabilities of the ***CNN***. Through experiments, we have found that the best results are obtained with initial random synaptic weights taking values in the interval [-0.005, 0.005].

The ***learning rate*** has also an important effect on the learning process. We have started the learning process with a learning rate of 0.01, afterwards the learning rate being decreased with 10% at every 3 epochs.

We have used as ***activation function*** the ***hyperbolic tangent*** [5], instead of the ***sigmoidal function*** [5], because it has the advantage of being symmetrical related to the origin, and has a greater range of values in interval [-1,1] instead of interval [0,1]. A good practice is to limit the range of values for the hyperbolic tangent activation function, we have trained our ***CNN*** and calculated the error of the output neurons using the interval [-0.8, 0.8] instead of interval (-1. 1).

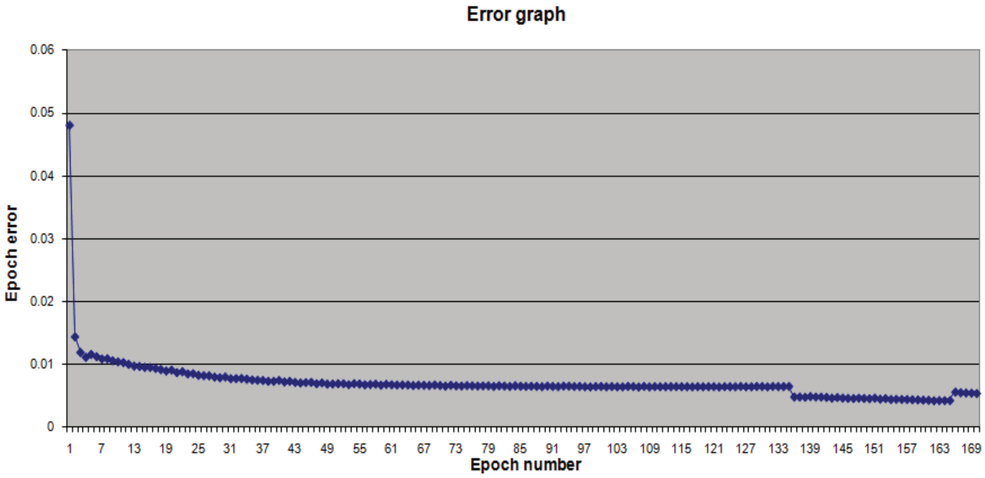


Figure 5. ***CNN*** learning error based on the ***NIST*** training images

In practical applications, when we try to extract digits from an image it is very difficult to find the exact algorithm for centering and resizing the digits to 29 × 29 resolution, similar to the ***NIST*** training set. This is why we trained the ***CNN*** using the initial ***NIST*** training images and a new training set that we have obtained after the removal of the non-useful information from the training images, using the technique presented before.

4. The testing process

After stopping the training process, we have tested the generalization capabilities of the neural network using **10,000** images (the initial ***NIST*** testing database and the testing images obtained from cutting out the useless information from the ***NIST*** testing images). To obtain better performance and to save time we have used in the testing process the pre-encoding method and the loading procedure of the testing images in the internal memory. The ***NIST*** testing database is considered to be a benchmark and a very difficult testing set for handwritten digits recognition applications.

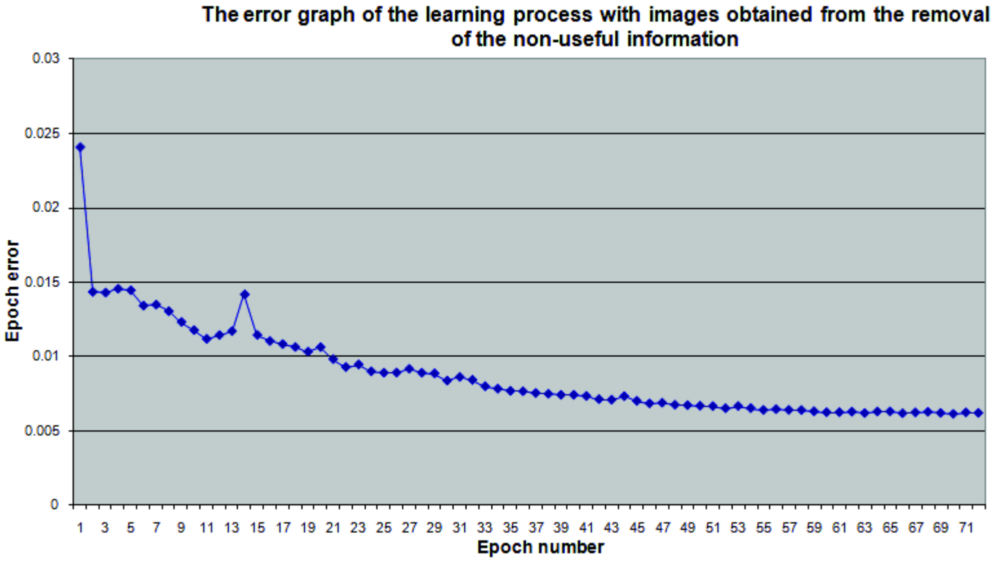


Figure 6. ***CNN*** learning error based on the ***NIST*** training images with the useless information removed

Our ***CNN*** obtained the following performances: with the original ***NIST*** testing images **96.74%** accuracy and **96.56%** accuracy with the images without background.

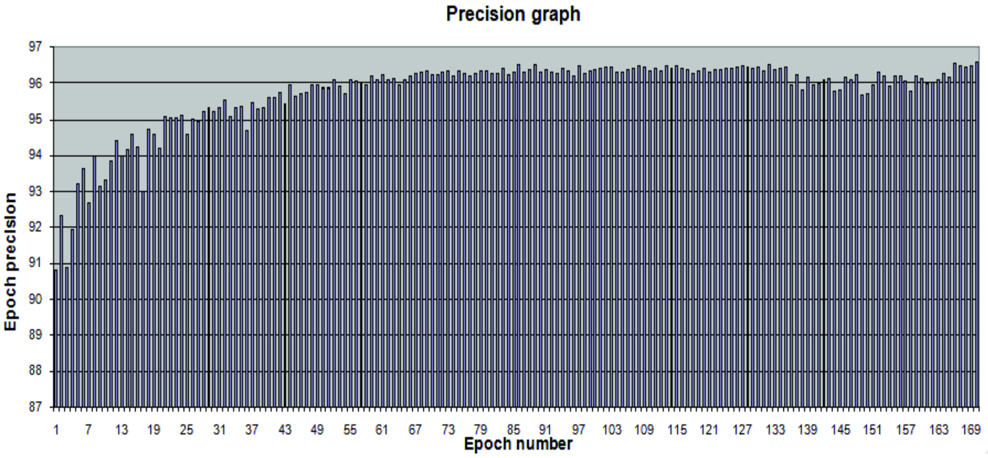
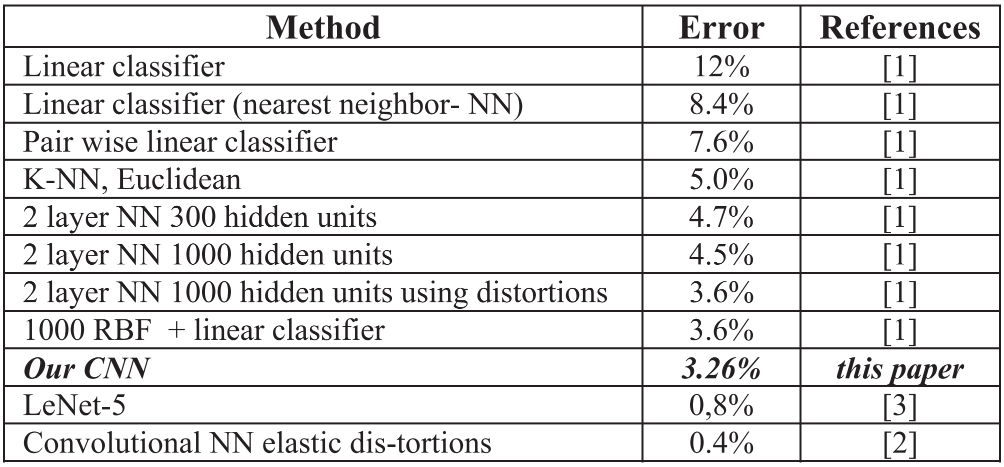


Figure 7. ***CNN*** accuracy based on the ***NIST*** testing images

Table 1 contains the best results obtained so far in the world, they range from an error of 12% up to an error of 0.4%, using the ***NIST*** testing database with 10,000 images.

Table 1. Results of the learning process for different neural networks



Method	Error	References
Linear classifier	12%	[1]
Linear classifier (nearest neighbor- NN)	8.4%	[1]
Pair wise linear classifier	7.6%	[1]
K-NN, Euclidean	5.0%	[1]
2 layer NN 300 hidden units	4.7%	[1]
2 layer NN 1000 hidden units	4.5%	[1]
2 layer NN 1000 hidden units using distortions	3.6%	[1]
1000 RBF + linear classifier	3.6%	[1]
Our CNN	***3.26%***	***this paper***
LeNet-5	0,8%	[3]
Convolutional NN elastic dis-tortions	0.4%	[2]

5. Conclusions

This article presents a method for analysis of visual documents and can be considered a starting point for the problem of handwritten letters recognition and handwriting recognition. In the future we will try to extend this architecture to support the recognition of handwritten letters based on the ***NIST*** database of handwritten letters containing 800,000 images. New methods for training the neural network must be found in order to speed up the learning process and to manage this high dimensional learning set.

References

[1] E. Kussul, T. Baidyk, "Improved method of handwritten digits recognition. UNAM", Centro de Instrumentos, Cd. Universitaria A.P. 70-186 , CP 04510, Mexico D.F, 2002.

[2] P.Y. Simard, D. Steinkraus, J.C. Platt, "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis", Microsoft Research, One Microsoft Way, Redmond WA 98052, 2003.

[3] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-based learning applied to document recognition", Proceedings of the IEEE, v. 86, pp 2278-2324, 1998.

[4] http://www.cs.toronto.edu/ roweis/data.html

[5] C. Enăchescu, ”Calculul neuronal”, Casa Cărţii de Ştiinţă, ISBN: 978-973-133-460-8, Cluj Napoca, 2009.

“Petru Maior” University of Târgu-Mureş, N. Iorga 1, 540088, Târgu-Mures, Romania

E-mail address: `ecalin@upm.ro`

E-mail address: `miron.cristian@stud.upm.ro`

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LV**, Number 1, 2010

SOME COMBINATORIAL ASPECTS OF THE KSAm-LIKE ALGORITHMS SUITABLE FOR RC4 STREAM CIPHER

BOGDAN CRAINICU, FLORIAN MIRCEA BOIAN

ABSTRACT. RC4 remains one of the most widely used stream cipher. In order to face the main critical weaknesses, a number of proposals for modifying RC4 algorithm have been advanced. In this paper we analyze some combinatorial aspects regarding the randomness of a variant of the Key-Scheduling Algorithm (KSA), called KSAm, proposed by Crainicu and Boian in [2] as a better protection against Initialization Vectors (IVs) weakness of Wired Equivalent Privacy (WEP) cryptosystems. Based on a model presented by Mironov in [19], we calculate the sign of the entries' permutation of the internal state table S after KSAm, which provides a negligible advantage of guessing a particular bit. Then, we analyze the probability of the event where a particular initial value follows a linear forward movement through the vector S, with possible undesirable consequences in predicting the value during that movement.

1. INTRODUCTION

RC4 is a stream cipher which was designed by Ron Rivest in 1987 for RSA Security. RC4 was kept as a trade secret until an alleged copy of it was anonymously posted to the Cypherpunks mailing list in 1994.

Because of its simplicity and speed, RC4 is one of the most widely used stream cipher; for example, it is used in the SSL/TLS (Secure Socket Layer/ Transport Layer Security) standards, WEP (Wired Equivalent Privacy), WPA (Wi-Fi Protected Access), and it can be also found in email encryption products.

There were discovered many significant weaknesses of RC4 and RC4-based WEP implementations: weak IVs/keys [1, 4, 8, 9, 10, 13, 14, 21, 22, 26, 28], invariance weakness [4], bias in the second output [16], related key attack [4,

Received by the editors: January 2, 2010.

2010 *Mathematics Subject Classification.* 68P25, 94A60, 11T71.

1998 *CR Categories and Descriptors.* G.2.1 [**Mathematics of Computing**]: Combinatorics – *Permutation and combinations*; G.3 [**Mathematics of Computing**]: Probability and Statistics – *Random number generation*; E.3 [**Data**]: Data Encrypting – *Code breaking.*

Key words and phrases. KSA, KSAm, RC4, sign of permutation, randomness of state table.

7], state recovery attack [11, 20, 23, 27], distinguishing attack [3, 5, 16-19, 24, 25], biased distribution of RC4 initial permutation [24, 28].

In order to address its critical weaknesses, a number of proposals for modifying RC4 algorithm have been advanced: Paul and Preneel present in [25] a new pseudorandom bit generator called RC4A, Zoltak proposes in [29] the VMPS stream cipher, and Gong, Gupta, Hell and Nawaz also propose in [6] a new 32/64-bit RC4-like keystream generator.

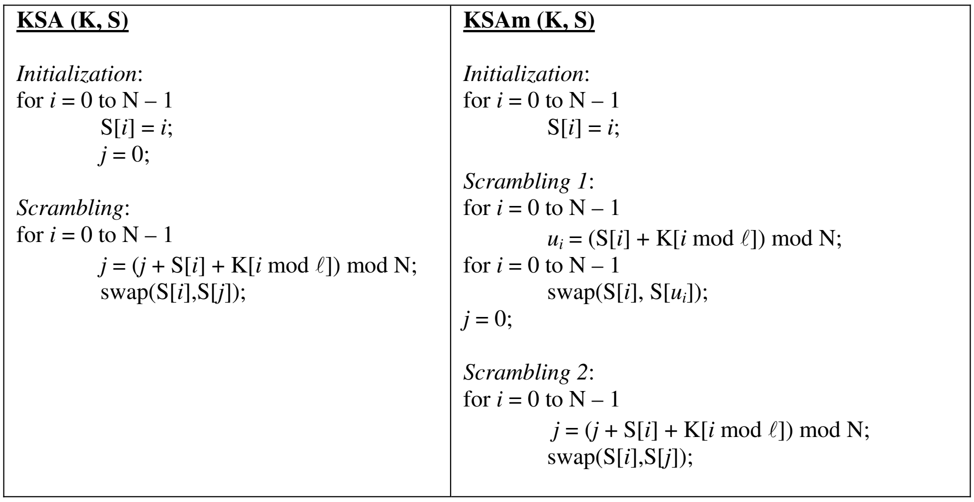
Crainicu and Boian proposed in [2] a modified version of KSA, called KSAm, whose primary goal was to face the weakness exhibited by Fluhrer, Mantin and Shamir (FMS) in [4], where certain IVs leaks information about the secret key in WEP mode of operation. The authors demonstrate that the attacker has no possibilities to manipulate KSAm permutation in order to reach the FMS resolved condition. Based on the Roos' experimental observation [26], there is a weaker probabilistic correlation between the first three words of the secret key and the first three entries of the state table after KSAm, which causes a negligible bias of the first word of the RC4_{KSAm} (RC4 with KSAm as Key-Scheduling Algorithm) output stream towards the sum of the first three words of the secret key. The effect of this negligible bias can be easily avoided by discarding only the first word from the RC4_{KSAm} output stream.

In this paper, we examine two combinatorial properties of the KSAm in its normal mode of operation, and not from the perspective of a particular implementation. Firstly, based on a model of a shuffling technique presented by Mironov in [19], where an idealized RC4 stream cipher is involved, we comute the sign of the permutation of S after KSAm, whose values help to predict a bit b with a probability of 0.91% over a random guess. This advantage is too small to be feasible in an attack. We also analyze the state table entries during the KSAm steps, with special focus on calculating the probability of a linear advance movement of an initial value from a particular state table entry during KSAm. The results prove that it is very unlikely to find a location in S during such movement where that value may be predicted with a probability significantly greater than 1/N.

2. KSAm

Crainicu and Boian suggest in [2] a modified version of the original KSA, called KSAm (Fig. 1), for addressing the FMS weakness of WEP-like cryptosystems, mainly when IV precedes the secret key.

The KSAm takes the secret key and initializes a vector of indices $u_0, u_1, \ldots,$ u_{N-1}; the values of indices u_i are not necessarily unique within the vector of indices, and they are kept secret. Then, it swaps the two values of S pointed to



```
KSA (K, S)

Initialization:
for $i$ = 0 to N – 1
        S[$i$] = $i$;
        $j$ = 0;

Scrambling:
for $i$ = 0 to N – 1
        $j$ = ($j$ + S[$i$] + K[$i$ mod $\ell$]) mod N;
        swap(S[$i$],S[$j$]);
```

```
KSAm (K, S)

Initialization:
for $i$ = 0 to N – 1
        S[$i$] = $i$;

Scrambling 1:
for $i$ = 0 to N – 1
        $u_i$ = (S[$i$] + K[$i$ mod $\ell$]) mod N;
for $i$ = 0 to N – 1
        swap(S[$i$], S[$u_i$]);
$j$ = 0;

Scrambling 2:
for $i$ = 0 to N – 1
        $j$ = ($j$ + S[$i$] + K[$i$ mod $\ell$]) mod N;
        swap(S[$i$],S[$j$]);
```

Figure 1. KSA vs KSAm [2]

by i and u_i, so that the *Scrambling 1* stage of KSAm ends with a secret state, which is different from the identity permutation with a very high probability. The rest of operations (*Scrambling 2*) remain the same as in the original KSA: it applies the scrambling rounds $N = 2^n$ times, stepping i across S, updating j by adding the previous value of $j, S[i]$ and the next word of the key.

In fact, KSAm comprises a family of key scheduling algorithms, where *Scrambling 1* sequence tries to follow the Knuth's observation [12]: instead of swapping $S[i]$ with a random entry, it must be swapped with an entry randomly chosen from $S[i]$ to $S[N-1]$ (the implementation of this concept remains still problematic due to the randomness of the secret key K).

At a glance, the first observation is that there are now two different scrambling processes, both of them based on the same secret key. Even if the computation/running time of KSAm is almost twice as long as that of KSA, the additional time is insignificant (the software implementation remains very fast).

The security of KSAm comes also from its huge internal state. The internal state of RC4$_{KSA}$ is approximately 1700 bits for 8-bits words. Instead, KSAm provides a much larger size and, as result, it is much harder to reconstruct its internal state. Crainicu and Boian present in [2] the formula for calculating L$_{RC4-KSAm}$, which represents the size of the RC4$_{KSAm}$' internal state (the values of indices u_i are not necessarily unique; therefore, the number of all possibilities of distributing 2^n elements into 2^n cells where repetitions are allowed is $(2^n)^{2^n}$) [2]:

$$L_{RC4-KSAm} = \log_2(2^n! \times (2^n)^{2^n} \times (2^n)^2)] = [\log_2(2^n!) + (n \times 2^n) + 2n]$$
$$L_{RC4-KSAm,n=8} \approx 3748 \text{ bits}$$

Beside KSA, KSAm needs only additional 256 bytes of memory for the indices u_i (for $N = 8$), which represents a negligible amount of supplementary memory.

3. On Randomness of KSAm

Two independent scrambling processes are involved with KSAm; therefore, after running consecutively both of them, each element of the state table will be swapped at least twice (possibly with itself).

Based on a series of significant studies on the original KSA [5, 16, 17, 19, 20, 23, 24], two of the most important approaches for analyzing KSAm are to deduce the probability of a linear advance movement of a value b which each step along the locations of vector S, and also to calculate the probability of a value b to end up in any location a (including the probability of identity permutation).

3.1. **The sign of the permutation S after KSAm.**

Mironov calculates in [19] the limiting distribution for the two possible values (+1 and −1) of the sign of S after KSA:

$$P(sign(S) = (-1)^N) = \frac{1}{2}\left(1 + e^{-2}\right)$$
$$P(sign(S) = (-1)^{N-1}) = \frac{1}{2}\left(1 - e^{-2}\right)$$

Therefore, Mironov demonstrates that it is possible to predict the sign of the permutation S after KSA with probability $1 - \frac{1}{2} \cdot e^{-2}$, and he shows that this value of about 6.7% over a random guess becomes also the advantage of guessing the bit b correctly.

Next, we compute the sign of the permutation S after KSAm. Two scrambling sequences are involved in KSAm, and therefore we have $2N$ rounds. At each round and regardless of what scrambling sequence is about, the swap process changes the sign of the permutation only if $i \neq j$. This happens with probability $\left(1 - \frac{1}{N}\right)$. If $i = j$, with probability $\frac{1}{N}$, the values of S pointed by i and j remain unchanged. The probability that $i \neq j$ during all $2N$ rounds of KSAm is $\left(1 - \frac{1}{N}\right)^{2N}$, and the probability that $i = j$ during all $2N$ rounds of KSAm is $\left(\frac{1}{N}\right)^{2N}$. The sign of the identity permutation is +1.

Based on these observations, the probability that the sign of S is changed an even number of times and at the end of KSAm is negative, is:

$$P(sign(S)=(-1)^{2N})=\left(1-\frac{1}{N}\right)^{2N}+C_{2N}^{2}\cdot\left(1-\frac{1}{N}\right)^{2N-2}\cdot\frac{1}{N^2}+$$
$$C_{2N}^{4}\cdot\left(1-\frac{1}{N}\right)^{2N-4}\cdot\frac{1}{N^4}+\ldots+C_{2N}^{N}\cdot\left(1-\frac{1}{N}\right)^{N}\cdot\frac{1}{N^N}+$$
$$C_{2N}^{N+2}\cdot\left(1-\frac{1}{N}\right)^{N-2}\cdot\frac{1}{N^{N+2}}+\ldots+C_{2N}^{2N}\cdot\left(1-\frac{1}{N}\right)^{0}\cdot\frac{1}{N^{2N}}=$$
$$\left(1-\frac{1}{N}\right)^{2N}\cdot\left[1+\frac{2N\cdot(2N-1)}{2!\cdot N^2}\cdot\left(1-\frac{1}{N}\right)^{-2}+\right.$$
$$\left.\frac{2N\cdot(2N-1)\cdot(2N-2)\cdot(2N-3)}{4!\cdot N^4}\cdot\left(1-\frac{1}{N}\right)^{-4}+\ldots+\frac{1}{N^{2N}}\right]\rightarrow$$
$$e^{-2}\cdot\left(1+\frac{4}{2!}+\frac{16}{4!}+\frac{64}{6!}+\ldots\right)=$$
$$\frac{e^{-2}}{2}\cdot\left(2+\frac{8}{2!}+\frac{32}{4!}+\frac{128}{6!}+\ldots\right)\xrightarrow[N\to\infty]{}\frac{e^{-2}}{2}(e^2+e^{-2})=\frac{1}{2}\cdot\left(1+e^{-4}\right)$$

The probability that the sign of S is changed an odd number of times and at the end of KSAm is negative, is:

$$P(sign(S)=(-1)^{2N-1})=1-P(sign(S)=(-1)^{2N})=\frac{1}{2}\cdot\left(1-e^{-4}\right).$$

After KSAm, the sign of S can be predicted with an very small advantage of $\frac{e^{-4}}{2}\approx 0,0091$ over the random guess, which means approximately 0.91%. Mironov obtains in [19] approximately the same result by running consecutively two times the original KSA.

Even if we found a small bias towards the sign of the permutation after KSAm, the value $\frac{e^{-4}}{2}$ is totally useless for attacking RC4 based on KSAm. As further precaution, discarding only the first three words of the output of RC4 ensures the security of the algorithm.

3.2. **Probability of a linear advance movement of an initial value from a particular state table entry during KSAm.** We define the minimum probability of a given initial entry $S_0[a]=a$ as the probability that this entry remains unchanged during each step of one of the two scrambling processes of KSAm. Thus, the minimum probability of the identity permutation is defined as the multiplication of all minimum probabilities corresponding to each initial entry $S_0[a]=a$.

We refine the Theorem 1 from [2]:

Theorem 1. *The minimum probability of a given initial entry $S_0[a] = a$ after N steps is:*

$$P(S_N[a] = a) = \frac{1}{N} \cdot \left(1 - \frac{1}{N}\right)^{N-1} \tag{1}$$

Proof. At a some point, the index i touches the value a. In this round, with probability $1/N$, $i = j = a$, and therefore $S[a]$ will be swapped with itself. For the rest of the $(N-1)$ rounds we have $i \neq a$, and $j \neq a$ with probability $\left(1 - \frac{1}{N}\right)$.

The minimum probability of the identity permutation, which means that all N entries of table S remain unchanged after completion of one of the two scrambling process is:

$$P(S_N = \text{identity\_permutation}) = \left[\frac{1}{N} \cdot \left(1 - \frac{1}{N}\right)^{N-1}\right]^N \tag{2}$$

□

Result 1. [19]: The probability of the value b to end up in location a in S at the end of the KSA round (the distribution of KSA outputs) is:

$$P[S_N[a] = b] = \begin{cases} \frac{1}{N}\left[\left(1 - \frac{1}{N}\right)^{N-a-1} + \left(1 - \frac{1}{N}\right)^b\right] & \text{if } a < b \\ \frac{1}{N}\left[\left(1 - \frac{1}{N}\right)^b + \left(1 - \left(1 - \frac{1}{N}\right)^b\right)\left(1 - \frac{1}{N}\right)^{N-a-1}\right] & \text{if } a \geq b \end{cases}$$

For example, the minimum probability for event $S_N[0] = 0$ is $\frac{1}{N} \cdot \left(1 - \frac{1}{N}\right)^{N-1}$, and according to Result 1, the probability for event $S_N[0] = 0$ is $\frac{1}{N}$.

Theorem 2. *Assuming that both indices u_i and j take values independently and uniformly at random at each round of the two scrambling processes of KSAm, the probability of the value b to end up in location $a+1$ in S in the round a of Scrambling 1 or Scrambling 2, where $S_0[0] = b$, is:*

$$P(S_a[(a+1) \bmod N] = b) = \frac{1}{N}, a \in [1, N] \text{ and } b \in [0, N-1] \tag{3}$$

Proof. At each round, taking into account that the value of i is known, we can analyze the probability of the event $S_a[(a+1) \bmod N] = b$, which depends on the value taken by u_i or j. The value b, during the a rounds, must not remain behind a location pointed by $i (i \in [0, (a+1) \bmod N])$ nor to reach a position after the $[(a+1) \bmod N]^{th}$ location in S, because, in these situations, there are no possibilities to manipulate the value b so that it ends up in location $[(a+1)\bmod N]$. Following the steps of either *Scrambling 1* or *Scrambling 2*, we have:

$S_1[2]=b$ if $j_1=2$ (the initial condition is $S_0[0]=b$) $\Rightarrow$

$P(S_1[2]=b)=P(j_1=2)=\dfrac{1}{N}$;

$S_2[3]=b$ if $<j_1=1$ and $j_2=3>$ or $<j_1=3$ and $j_2\neq 3>\Rightarrow$

$P(S_2[3]=b)=P(j_1=1)\cdot P(j_2=3)+P(j_1=3)\cdot P(j_2\neq 3)=$

$\dfrac{1}{N^2}+\dfrac{1}{N}\cdot\left(1-\dfrac{1}{N}\right)=\dfrac{1}{N}$;

$S_3[4]=b$ if $<j_1=1$ and $j_2=2$ and $j_3=4>$ or $<j_1=1$ and $j_2=4$ and $j_3\neq 4>$ or

$<j_1=2$ and $j_2\neq 2$ and $j_3=4>$ or $<j_1=4$ and $j_2\neq 4$ and $j_3\neq 4>\Rightarrow$

$P(S_3[4])=P(j_1=1)\cdot P(j_2)=2)\cdot P(j_3=4)+P(j_1=1)\cdot P(j_2=4)\cdot P(j_3\neq 4)+$

$P(j_1=2)\cdot P(j_2\neq 2)\cdot P(j_3=4)+P(j_1=4)\cdot P(j_2\neq 4)\cdot P(j_3)\neq 4)=$

$\dfrac{1}{N^3}+\dfrac{2}{N^2}\cdot\left(1-\dfrac{1}{N}\right)+\dfrac{1}{N}\cdot\left(1-\dfrac{1}{N}\right)^2=\dfrac{1}{N}$;

$$\begin{aligned}
&\ldots\\
&P(S_4[5]=b)=\frac{1}{N^4}+\frac{3}{N^3}\cdot\left(1-\frac{1}{N}\right)+\frac{3}{N^2}\cdot\left(1-\frac{1}{N}\right)^2+\\
&+\frac{1}{N}\cdot\left(1-\frac{1}{N}\right)^3=\frac{1}{N}\\
&P(S_5[6]=b)=\frac{1}{N^5}+\frac{4}{N^4}\cdot\left(1-\frac{1}{N}\right)+\frac{6}{N^3}\cdot\left(1-\frac{1}{N}\right)^2+\\
&+\frac{4}{N^2}\cdot\left(1-\frac{1}{N}\right)^3+\frac{1}{N}\cdot\left(1-\frac{1}{N}\right)^4=\frac{1}{N};\\
&\ldots
\end{aligned}$$

$$P(S_a[a+1]=b) = \frac{1}{N}\cdot\left(1-\frac{1}{N}\right)^{a-1} + \frac{(a-1)}{1!\cdot N^2}\cdot\left(1-\frac{1}{N}\right)^{a-2} + \frac{(a-1)\cdot(a-2)}{2!\cdot N^3}\cdot\left(1-\frac{1}{N}\right)^{a-3} + \frac{(a-1)\cdot(a-2)\cdot(a-3)}{3!\cdot N^4}\cdot\left(1-\frac{1}{N}\right)^{a-4} + \frac{(a-1)\cdot(a-2)\cdot(a-3)\cdot(a-4)}{4!\cdot N^5}\cdot\left(1-\frac{1}{N}\right)^{a-5} + \ldots + \frac{1}{N^a} = \frac{1}{N}$$

□

For *Scrambling 1*, $b = 0$, and for *Scrambling 2*, $b \in [0, 255]$. Applying the Result 1, where the initial state is the identity permutation, for $b = 0$ and $a = 1$, we obtain $P[S_N[1] = 0] = \frac{1}{N}$, and applying the Theorem 2 for $a = N$, we have the same result: $P(S_N[(N+1)\text{mod}N] = 0) = \frac{1}{N}$.

Theorem 2 can be adapted for all entries of the initial permutation S_0. The significance of Theorem 2 lies in the fact that the event that a particular value b follows a linear path through the vector S, and consequently ends up in an expected location after *Scrambling1*, has a probability around $\frac{1}{N}$. This result has also to be considered in the context of starting *Scrambling 2* with a state table S which is different from the identity permutation with high probability.

4. Conclusions

We investigated KSAm [2] from the point of view of shuffling algorithm presented by Mironov in [19]. We calculated as well the probability of the sign of the permutation S, but after KSAm, finding a value which may help in predicting that sign with an advantage of 0.91% over a random guess. Mironov obtains in [19] about the same result by running consecutively two times the original KSA, but KSAm benefits from a much larger size of the internal state and the running of two different scrambling processes. The mentioned advantage of 0.91% over a random guess, biased towards the sign of the permutation after KSAm, is though too small, so that an attack against RC4_{KSAm} could not rely on it. As precaution, an additional measure for thwarting against this weakness consists in discarding only the first three words of the output of RC4_{KSAm}.

The second part of our analysis is focused on calculating the probability of a particular event, namely, a linear advance movement of the state table entry $S_0[0] = b$ during KSAm rounds. The value obtained is about $\frac{1}{N}$, which demonstrates that such event happens randomly. The result can be extended to the others entries of the initial state table S.

References

[1] A. Bittau, *Additional weak IV classes for the FMS attack*, Department of Computer Science, University College London, 2003. Available: http://www.cs.ucl.ac.uk/staff/a.bittau/sorwep.txt.

[2] B. Crainicu, F. M. Boian, *KSAm – An Improved RC4 Key-Scheduling Algorithm for Securing WEP*, in Novel Algorithms and Techniques in Telecommunications and Networking, Springer, Netherlands 2010, ISBN 978-90-481-3661-2, pp. 391-396.

[3] S. Fluhrer, D. McGrew, *Statistical analysis of the alleged RC4 keystream Generator*, in. Proc. 7th International Workshop, FSE 2000, New York, Lecture Notes in Computer Science, Vol. 1978, Springer-Verlag, 2001, pp. 66-71.

[4] S. Fluhrer, I. Mantin, A. Shamir, *Weaknesses in the key scheduling algorithm of RC4*, in Proc. 8th Annual International Workshop, SAC 2001, Toronto, Lecture Notes in Computer Science, Vol. 2259, Springer-Verlag, 2001 pp. 1-24.

[5] J. Dj. Goli, *Linear statistical weakness of alleged RC4 keystream generator*, in. Proc. International Conference on the Theory and Application of Cryptographic Techniques, EUROCRYPT '97, Konstanz, Lecture Notes in Computer Science, Vol. 1233, Springer-Verlag, 1997, pp. 226-238.

[6] G. Gong, K. C. Gupta, M. Hell, Y. Nawaz, *Towards a General RC4-like Keystream Generator*, in Proc. First SKLOIS Conference*, CISC 2005*, Beijing, Lecture Notes in Computer Science, Vol. 3822, Springer-Verlag, 2005, pp. 162-174.

[7] A. L. Grosul, D. S. Wallach, *A related key cryptanalysis of RC4*, Technical Report TR-00-358, Department of Computer Science, Rice University, 2000. Available: www.weizmann.ac.il/mathusers/itsik/RC4/Papers/GrosulWallach.ps

[8] D. Hulton, *Practical exploitation of RC4 weaknesses in WEP environments*, 2001. Available: http://www.datastronghold.com/security-articles/hacking-articles/practical-exploitation-of-rc4-weaknesses-in-wep-environments.html

[9] KoreK, *Need security pointers*, 2004. Available: http://www.netstumbler.org/showthread.php?postid=89036#post89036.

[10] KoreK, *Next generation of WEP attacks?*, 2004. Available http://www.netstumbler.org/showpost.php?p=93942&postcount=35

[11] L. R. Knudsen, W. Meier, B. Preneel, V. Rijmen, S. Verdoolaege, *Analysis Methods for (Alleged) RC4*, in Proc. International Conference on the Theory and Application of Cryptology and Information Security, ASIACRYPT'98, Beijing, Lecture Notes in Computer Science, Springer-Verlag, Vol.1514, 1998, pp.327–341.

[12] E. Knuth, *The Art of Computer Programming*, Third edition, Volume 2, Addison-Wesley, 1997.

[13] K. Kobara, H. Imai, *Key-Dependent Weak IVs and Weak Keys in WEP – How to Trace Conditions Back to Their Patterns –*, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol. E89-A, No. 8, 2006, pp. 2198-2206.

[14] K. Kobara, H. Imai, *IVs to Skip for Immunizing WEP against FMS Attack*, IEICE Transactions on Communications, Vol.E91–B, No.1, 2008, pp. 218-227.

[15] I. Mantin, *The Security of the Stream Cipher RC4*, Master Thesis, The Weizmann Institute of Science, 2001.

[16] I. Mantin, A. Shamir, *A practical attack on broadcast RC4*, in Proc. 8th International Workshop, FSE 2001, Yokohama, Lecture Notes in Computer Science, Springer-Verlag, Vol. 2355, 2002, pp. 87-104.

[17] I. Mantin, *Predicting and Distinguishing Attacks on RC4 Keystream Generator*, in. Proc. 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, EUROCRYPT 2005, Aarhus, Lectures Notes in Computer Science, Vol. 3494, Springer-Verlag, 2005, pp. 491-506.

[18] I. Mantin, *A Practical Attack on the Fixed RC4 in the WEP Mode*, in Proc. 11th International Conference on the Theory and Application of Cryptology and Information Security, ASIACRYPT 2005, Chennai, Lecture Notes in Computer Science, Springer-Verlag, Vol. 3788, 2005, pp. 395-411.

[19] I. Mironov, *(Not So) Random Shuffles of RC4*, in Proc. 22nd Annual International Cryptology Conference, Advances in Cryptology, CRYPTO 2002, Santa Barbara, Lecture Notes in Computer Science, Springer-Verlag, Vol. 2442, 2002, pp. 304–319.

[20] S. Mister, S. E. Tavares, *Cryptanalysis of RC4-like Ciphers*, in Proc. 5th Annual International Workshop, SAC 1998, Kingston, Lecture Notes in Computer Science, Springer-Verlag, Vol.1556, 1999, pp. 131–143.

[21] T. Ohigashi, Y. Shiraishi, M. Morii, *Most IVs of FMS Attack-Resistant WEP Implementation Leak Secret Key Information*, in Proc. 2005 Symposium on Cryptography and Information Security, Maiko, Vol. 4, 2005, pp. 1957–1962.

[22] T. Ohigashi, Y. Shiraishi, M. Morii, *FMS Attack-Resistant WEP Implementation Is Still Broken – Most IVs Leak a Part of Key Information –* , in Proc. International Conference, CIS 2005, Xi'an, Lecture Notes in Computer Science, Springer-Verlag, Vol. 3802, 2005, pp. 17-26.

[23] T. Ohigashi, Y. Shiraishi, M. Morii, *New Weakness in the Key-Scheduling Algorithm of RC4*, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol. E91-A, No. 1, 2008, pp. 3-11.

[24] S. Paul, B. Preneel, *Analysis of Non-fortuitous Predictive States of the RC4 Keystream Generator*, in Proc. 4th International Conference on Cryptology in India, INDOCRYPT 2003, New Delhi, Lecture Notes in Computer Science, Springer-Verlag, Vol. 2904, 2002, pp. 52-67.

[25] S. Paul, B. Preneel, *A New Weakness in the RC4 Keystream Generator and an Approach to Improve the Security of the Cipher*, in Proc. 11th International Workshop, FSE 2004, Delhi, Lecture Notes in Computer Science, Springer-Verlag, Vol. 3017, 2004, pp. 245–259.

[26] A. Roos, *Class of weak keys in the RC4 stream cipher*, Two posts in sci.crypt, message-id 43u1eh$1j3@hermes.is.co.za and 44ebge$llf@hermes.is.co.za, 1995.

[27] Y. Shiraishi, T. Ohigashi, M. Morii, *An improved Internal-State Reconstruction Method of a Stream Cipher RC4*, in Proc. IASTED International Conference on Communication, Network, and Information Security, CNIS 2003, New York, 2003, pp. 132-135.

[28] D. Wagner, *My RC4 weak keys*, Post in sci.crypt, message-id 447o1l$cbj@cnn.princeton.edu, 1995. Available: http://www.cs.berkeley.edu/˜daw/my-posts/my-rc4-weak-keys

[29] B. Zoltak, *VMPC One-Way Function and Stream Cipher*, in Proc. 11th International Workshop, FSE 2004, Delhi, Lectures Notes in Computer Science, Vol. 3017, Springer-Verlag, 2004, pp. 210–225.

"Petru Maior" University of Tîrgu-Mureş
E-mail address: cbogdan@upm.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LV**, Number 1, 2010

ENCRYPTION SYSTEM OVER THE SYMMETRICAL GROUP OF ORDER N

STELIAN FLONTA, LIVIU-CRISTIAN MICLEA, ENYEDI SZILÁRD

Abstract. The encryption systems with public keys are related to algebraic structures, which have to ensure, by means of their computational properties or their dimensions, a high level of security for the generated keys, which are secret. The ElGamal algorithm is defined over the group of remainder classes modulo n, which is a suitable structure for an encryption algorithm. This paper chooses another structure for which the ElGamal algorithm is to be applied to. This structure is the symmetrical group of order n. Properties which ensure the opportunity of choosing this structure are: the cardinal of the group is high enough, the operation of composition of the permutations is simple from a computational point of view and every permutation can be decomposed uniquely into a product of disjunctive cycles.

1. Introduction

The main goal of this paper is to develop an encryption system. Among the secondary objectives, which are necessary in order to achieve the main goal, the following stand out: elaborating a method for encoding / decoding a message by means of a permutation and presenting a method for choosing the permutation for the key generation. The paper starts with a paragraph that outlines different results of the research community. It continues with the presentation of the algorithm for encoding / decoding a message by the help of a permutation and the description of the cipher of type ElGamal. In the end, a study regarding the security perspective of the system is conducted and the author's contributions are pointed out.

Received by the editors: October 10, 2009.

2010 *Mathematics Subject Classification.* 94A60, 11T71.

1998 *CR Categories and Descriptors.* E.3 [**DATA ENCRYPTION**]: Subtopic – *Public key cryptosystems*;

Key words and phrases. ElGamal over the permutations group, discrete logarithm problem.

2.Related Work

The encryption systems are conceived starting from the properties of certain mathematical structures. Algorithms with public keys highlight those properties that ensure the generation of some keys sufficiently "secure" in a relatively simple way from computational point of view. Starting from the problem of the discrete logarithm over a group, an idea is presented in [9, p. 107] [8, p. 294] [1, ch. 12, p. 2] [7, p. 2]. A structure can be chosen to ensure a high level of difficulty for the discrete logarithm's determination from computational point of view. The problem of the discrete logarithm can be stated using different group structures. There are versions where the group is formed over the remainder classes or over the set of points which belong to the elliptical curves. Starting from these structures, the ElGamal and ElGamal over the elliptical curves encryption systems have been defined. The paper [13, p. 218] proposes a group structure formed over the set of points which belong to the conical curves. In this case, a conic is considered and the operation is defined so that a group structure is obtained over which the ElGamal system is implemented.

Another approach, which is described in [11, p. 473], concerns the mathematical structure over which the encryption system is defined. In the research community, some papers present hybrid systems obtained by combining some encryption systems with the ElGamal system [12, p. 436]. Other existing papers analyze the properties of the encryption function, proving that it is a homomorphism [3, p. 645]. Different approaches of the systems, which are based on the discrete logarithm problem [10, p. 210] [6, p. 2] [2, p. 9], outline the properties related to the method of calculus. There are limited possibilities for choosing a structure for which the problem of the discrete logarithm would be difficult to solve. Another solution [5, p. 445, tome III] presents an ElGamal type cryptographic primitive, with the key divided among the modulo n remainder classes. This differs from the present paper in that we propose an ElGamal type algorithm without a key divided over the n order symmetrical group. The paper [4, p. 226-234] proposes an encryption system based on discrete logarithm in symmetric group of n order problem. The differences between [4, p. 226-234] and the model developed in this paper consist in the coding mode of the message and the encryption algorithm. These differences will be presented in detail furthermore in the paper. The properties of the symmetrical group of order n recommend choosing it in the case of implementing an algorithm of ElGamal type. In this situation, the structure is a finite group which is not commutative.

3. ElGamal over the permutation group

This system will be defined in the following paragraph and it will be noted EGPGP. In order to construct a sufficiently secure system, it is necessary to choose a group with a great number of elements which allows the generation of some keys comparable with the ones frequently used, that is the ones of length 512 bits, 1024 bits or 2048 bits. Such a group is S_n, which has $n!$ elements. Applying the Stirling formula, which approximates $n!$, an estimation can be made for the cardinal of the group S_{128} and S_{1024}, so:

$128! \approx \left(\frac{128}{e}\right)^{128} \cdot \sqrt{2\pi 128} \geq 47^{128} \cdot 2^4 \geq 32^{128} \cdot 2^4 = 2^{644}$

$1024! \approx \left(\frac{1024}{e}\right)^{1024} \cdot \sqrt{2\pi 1024} \geq (2^8)^{1024} \cdot 2^6 = 2^{8198}$.

Another important aspect to be able to elaborate ElGamal over S_n is the codification of the information so that the product between the coded message m and the permutation h to be possible. From this results the necessity for the message m to be coded in a permutation, meaning that to every message m one and only one permutation will be associated. The association must be made using an easily computable bijective function. It is also important that the inverse of this function should be easily computed, this property being necessary in the decryption stage of the algorithm.

We consider A a set of symbols having the cardinal $n - k$. A message, of fixed length k, is $m = \alpha_1\alpha_2...\alpha_k$ where $\alpha_j \in A, \forall j = \overline{1,k}$.

Each symbol is encoded by a bijective function $f : A \to \{k+1, ..., n\}$. For each message m, a function $g_m : \{\alpha_1, \alpha_2, ..., \alpha_k\} \to \{1, ..., k\}$, $g_m(\alpha_j) = j, \forall j = \overline{1,k}$ is defined. This function associates to each symbol from the message m a number which means the symbol's position in the message. It is obvious the fact that the function g_m is bijective. We define the function

$$h : \{1, ..., n\} \to A, h(x) = \left\{ \begin{array}{c} g_m^{-1}(x), x \in \{1, ..., k\} \\ f^{-1}(x), x \in \{k+1, ..., n\} \end{array} \right.$$

For each $\alpha_j \in m, j = \overline{1,k}$, we chose $i_1^j, i_2^j, ..., i_{(l-1)_j}^j, i_{l_j}^j \in \{1, ..., k\}$ such that $h(i_1^j) = h(i_2^j) = ... = h(i_{(l-1)_j}^j) = h(i_{l_j}^j) = \alpha_j$ and is attached the cycle

$$c_j = (i_1^j, i_2^j, ..., i_{(l-1)_j}^j, i_{l_j}^j) \in S_n, \ \ i_1^j \leq i_2^j \leq ... \leq i_{(l-1)_j}^j \leq i_{l_j}^j.$$

Through this process we attach to each symbol $\alpha_j, \forall j = \overline{1,k}$, which appears in the message, a disjoint cycle of the other cycles. The length of this cycle is $s + 1$, where s is the number of times the symbol appears within the message. The permutation $p(m) = \prod c_j$ with the property that the cycles are all disjoint two by two. They are also longer than or equal to two. The reverse process through which the message m of $p(m)$ is determined is described in the following lines.

All disjoint cycles of $p(m)$, longer than or equal to two, are paired, and for each cycle the symbols $(i_1, i_2, , .., i_{l-1}, i_l), \quad i_1 \leq i_2 \leq ... \leq i_{l-1} \leq i_l$ are determined such that $\alpha_{i_1} = \alpha_{i_2} = ... = \alpha_{i_{l-1}} = h(\alpha_{i_l})$. The message m recovery is realized by merging the symbols in increasing order of indices.

Of course, the cycles are permutations from the group where EGPGP is applied to, that is S_n. In practice, this can be S_{128} or S_{256} and the standard length of a message can be 32 characters or 64 characters. This choice is determined by the structure of the ASCII code. Starting from a standard length of the message, this idea can be generalized. If the standard length of the message is k, then the chosen group is S_{k+96}. We will come back to these possibilities with detailed explanations later in this paper. In [4, p. 230] the message m is an integer number. The permutation's determination, which is associated, is made using the representation in factorial base system, which is a positional system. In this paper the message m is formed by concatenation of alphanumeric symbols. The encoding through a permutation is made using a permutation's decomposition property in a unique mode in a product of disjoint cycles, making abstraction of order, and a numeric code for the used symbols, for example the ASCII code. Therefore the coding mode is totally different from [4, p. 230]. Further, a version of EGPGP over the group S_n will be presented. This variant is different from [1, ch. 12, p. 2] and [4, p. 226-234] by the mode of the keys generation, of encryption and decryption.

The steps of this algorithm are:

Key generation

A permutation $g \in S_n$ is considered such that the problem of the discrete logarithm is difficult to solve and $ordg = r$ is determined. Also the numbers $x_1, x_2 \in Z_{|\mathrm{H}|}$ are chosen such that $(x_1, r) = 1$, that is the two numbers are prime between them, where $H = \langle g \rangle$ and $h_1 = g^{x_1}, h_2 = g^{x_2}$ are computed. Also a number $s < n$ is chosen such that s divides the number x_2. From the extended Euclid algorithm the numbers α, β are determined such that $x_1\alpha + r\beta = 1$. The public key is $\{h_1, h_2, s\}$ and the secret key is $\{x_2, \alpha\}$. The elements g, r, x_1, β are secret, but they are not keys, consequently they will not be transmitted to the user of the secret keys.

Message encryption

If we desire the encryption of message m, with a maximum length k, than we determine $p(m)$, which is the associated permutation. Then $y \in Z_{|\mathrm{H}|}, t \in S_n$ are chosen, where t is a cycle of order s and $c_1 = t \cdot {h_1}^y$and $c_2 = p(m) \cdot {h_2}^y$ are computed. The encrypted message is (c_1, c_2).

Message decryption

The first step for decrypting the message (c_1, c_2) is to compute

$$\frac{c_2}{(c_1)^{x_2\alpha}} = \frac{p(m)\cdot(g^{x_2})^y}{t^{x_2\alpha}\cdot((g^{x_1})^y)^{x_2\alpha}} =$$

$$= \frac{p(m)\cdot g^{x_2 y}}{e\cdot g^{yx_2(-\beta r+1)}}\frac{p(m)\cdot g^{x_2 y}}{(g^{(-\beta r+1)})^{yx_2}} = \frac{p(m)\cdot g^{x_2 y}}{(g)^{yx_2}} = p(m) \ .$$

In the next step, message m is determined from $p(m)$ using the decoding algorithm [1, 8].

4. Choosing the Permutation

A few observations regarding the actual implementation of the EGPGP cipher are necessary. In order to generate efficient keys, a permutation g is needed for which the problem of the discrete logarithm is difficult to solve. This is true if the order of g is sufficiently high. A way of generating such a permutation is based on the theorem of decomposing a permutation into disjunctive cycles and on the way of computing the order of the permutation. From these results, one deduces that a permutation has a great order if it is the product of some disjunctive cycles whose the least common multiple of lengths is maximum with the restriction that their sum is constant. The choice of some disjunctive cycles from S_n is reduced to choosing a disjunctive subset of the set $\{1, ..., n\}$. Every subset $\{i_{1,}i_{2,...,}i_s\}$ generated the cycle $(i_{1,}i_{2,...,}i_s)$. Consequently, an optimization problem can be stated: Determine the numbers $k_{1,}k_2, ..., k_j \in Z^*$, prime among them two by two, such that

$$\begin{cases} k_1k_2...k_j \to \max \\ k_1 + k_2 + ... + k_j = n \end{cases} .$$

A solution to this problem corresponds to a maximum order permutation which can also be written as a product of disjunctive cycles that have the lengths $k_{1,}k_2, ..., k_j \in Z^*$.

5. Conclusions

In the ElGamal version $({Z_q}^*, \cdot)$, which is the classical version, there are computed multiplications, additions, exponentiations, modulo q inversions. If q is a great number there are adequate resources. The ElGamal over $(S_n, \cdot)$ version assumes smaller exponents and the product and the inversion of the permutation are simple operations. This is another reason for which ElGamal over $(S_n, \cdot)$is preferred. In the process of key generation for the ElGamal over $(S_n, \cdot)$ system it is very important to be able to choose permutations of high order to ensure the resistance against breaking by means of "brute force". A solution to this problem corresponds to a permutation of maximum order which is written as a product of cycles that have the lengths $k_{1,}k_2, ..., k_j \in Z^*$. Practically, n is given, $k_{1,}k_2, ..., k_j \in Z^*$ are chosen, prime among them, such that $k_1 + k_2 + ... + k_j = n$ and then the set $\{1, ..., n\}$ is decomposed in sets with respectively $k_{1,}k_2, ..., k_j$ elements and after that the cycles obtained from these sets are multiplied. The product obtained this way is the permutation g that can be used for key generation. The next computations, using the sum respectively the product of the prime numbers smaller or equal to 97 and the approximation of the factorial number, using the Stirling formula, shows that in $(S_{1096}, \cdot)$ exist permutations which have the order greater or equal to 2^{128}. Also, the order of the group is very large 2^{8773}.

- $\begin{array}{l} 2+3+5+7+11+13+17+19+23+29+31+ \\ +37+41+43+47+53+59+61+67+71+ \\ +73+79+83+89+97=1060 \end{array}$
- $\begin{array}{l} 2\cdot3\cdot5\cdot7\cdot11\cdot13\cdot17\cdot19\cdot23\cdot29\cdot31\cdot37\cdot41\cdot \\ \cdot43\cdot47\cdot53\cdot59\cdot61\cdot67\cdot71\cdot73\cdot79\cdot83\cdot89\cdot97\approx \\ \approx 2,3\cdot10^{36}\approx 2^{128} \end{array}$
- $$1096! \geq 2^{8773}$$

Consequently the group $(S_{1096}, \cdot)$ has at least the order 2^{8773}. Also, from this group, a permutation of order 2^{128} can be easily chosen.

We developed an encryption system of ElGamal type over the symmetrical group of order n, and we compared this system with the cipher defined over the modulo n remainder classes. Also, we have presented an algorithm for encoding / decoding a message by using permutations and a method for choosing the permutation for the key generation.

References

[1] A. Atanasiu, *Cryptography*, course notes, Bucharest, 2009, http://www.galaxyng.com/adrian_atanasiu/cript.htm

[2] B. Chevallier-Mames, P. Paillier and D. Pointcheval, *Encoding-free ElGamal encryption without random oracles* , Public Key Cryptography - PKC 2006, pp. 91-104.

[3] L. Chen, Y. Xu, W. Fang and C. Gao, *A New ElGamal-Based Algebraic Homomorphism and Its Application*, 2008 ISECS International Colloquium on Computing, Communication, Control and Management, Guangzhou, 2008, pp. 643-648.

[4] J. N. Doliskani, E. Malekian and A. Zakerolhosseini, *A Cryptosystem Based on the Symmetric Group Sn*, IJCSNS International Journal of Computer Science and Network Security, vol. 8 No. 2, 2008, pp. 226-234

[5] S. Flonta and L. Miclea, *An extension of the El Gamal encryption algorithm, Proceedings of 2008 IEEE International Conference on Automation, Quality and Testing, Robotics, Cluj-Napoca, AQTR 2008, pp. 444-446.*

[6] M. P. Jhanwar and R. Barua, *A Public Key Encryption In Standard Model Using Cramer-Shoup Paradigm*, Cryptology ePrint Archive, 2008.

[7] A. Mahalanobis, *The discrete logarithm problem in the group of non-singular circulant matrices*, Cryptology ePrint Archive, 2009.

[8] A. Menezes, P. Oorschot and S. Vanstome, *Handbook of Applied Cryptography*, CRC Press, 1996.

[9] V. V. Patriciu, M. Ene-Pietroșanu, I. Bica and J. Priescu *Electronic Signatures and Security in Informatics*, Editura All, Bucureşti 2006.

[10] R. Schmitz, *Public Key Cryptography - A Dynamical Systems Perspective*, 2008 Second International Conference on Emerging Security Information, Systems and Technologies, Cap Esterel, 2008, pp. 209-212.

[11] S. H. Paeng, K. C. Ha, J. H. Kim, S. Chee and C. Park, *New public key cryptosystem using finite non-abelian groups*, Crypto 2001 (J. Kilian, ed.), LNCS, vol. 2139, Springer-Verlag, 2001, pp. 470-485.

[12] Q. Bing-cheng, Y. Yang-zin, Z. Xi-min and C. Yin-dong, *Iterative Composite Encryption Algorithm Based on Tea and Elgamal*, 2009 WRI World Congress on Computer Science and Information Engineering, Los Angeles, 2009, pp. 435-438.

[13]Dalu Zhang , Min Liu , Zhe Yang,*Zero-Knowledge Proofs of Identity Based on ELGAMAL on Conic*, IEEE International Conference on E-Commerce Technology for Dynamic E-Business, CEC-East'04, Beijing, 2004, pp. 216-223.

Technical University of Cluj-Napoca, Cluj-Napoca, Romania

E-mail address: `Stelian.FLONTA@aut.utcluj.ro`, `Liviu.Miclea@aut.utcluj.ro`, `Szilard.Enyedi@aut.utcluj.ro`