Einoshin Suzuki
Setsuo Arikawa (Eds.)

LNAI 3245

# Discovery Science

7th International Conference, DS 2004
Padova, Italy, October 2004
Proceedings

DIALOGUES
2004

Springer

Einoshin Suzuki Setsuo Arikawa (Eds.)

# Discovery Science

7th International Conference, DS 2004
Padova, Italy, October 2-5, 2004
Proceedings

Springer

Volume Editors

Einoshin Suzuki
Yokohama National University
Electrical and Computer Engineering
79-5 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan
E-mail: suzuki@ynu.ac.jp

Setsuo Arikawa
Kyushu University, Department of Informatics
Hakozaki 6-10-1, Higashi-ku, Fukuoka 812-8581, Japan
E-mail: arikawa@i.kyushu-u.ac.jp

Library of Congress Control Number: 2004112902

CR Subject Classification (1998): I.2, H.2.8, H.3, J.1, J.2

ISSN 0302-9743
ISBN 3-540-23357-1 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 11324614 06/3142 5 4 3 2 1 0

# Lecture Notes in Artificial Intelligence 3245

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

# Preface

This volume contains the papers presented at the 7th International Conference on Discovery Science (DS 2004) held at the University of Padova, Padova, Italy, during 2–5 October 2004.

The main objective of the discovery science (DS) conference series is to provide an open forum for intensive discussions and the exchange of new information among researchers working in the area of discovery science. It has become a good custom over the years that the DS conference is held in parallel with the International Conference on Algorithmic Learning Theory (ALT). This co-location has been valuable for the DS conference in order to enjoy synergy between conferences devoted to the same objective of computational discovery but from different aspects. Continuing the good tradition, DS 2004 was co-located with the 15th ALT conference (ALT 2004) and was followed by the 11th Symposium on String Processing and Information Retrieval (SPIRE 2004). The agglomeration of the three international conferences together with the satellite meetings was called Dialogues 2004, in which we enjoyed fruitful interaction among researchers and practitioners working in various fields of computational discovery. The proceedings of ALT 2004 and SPIRE 2004 were published as volume 3244 of the LNAI series and volume 3246 of the LNCS series, respectively.

The DS conference series has been supervised by the international steering committee chaired by Hiroshi Motoda (Osaka University, Japan). The other members are Alberto Apostolico (University of Padova, Italy and Purdue University, USA), Setsuo Arikawa (Kyushu University, Japan), Achim Hoffmann (UNSW, Australia), Klaus P. Jantke (DFKI, Germany), Masahiko Sato (Kyoto University, Japan), Ayumi Shinohara (Kyushu University, Japan), Carl H. Smith (University of Maryland, College Park, USA), and Thomas Zeugmann (Hokkaido University, Japan).

In response to the call for papers 80 submissions were received. The program committee selected 20 submissions as long papers and 20 submissions as regular papers, of which 19 were submitted for publication. This selection was based on clarity, significance, and originality, as well as relevance to the field of discovery science. This volume consists of two parts. The first part contains the accepted long papers, and the second part contains the accepted regular papers.

We appreciate all individuals and institutions who contributed to the success of the conference: the authors of submitted papers, the invited speakers, the tutorial speakers, the steering committee members, the sponsors, and Springer. We are particularly grateful to the members of the program committee for spending their valuable time reviewing and evaluating the submissions and for participating in online discussions, ensuring that the presentations at the conference were of high technical quality. We are also grateful to the external additional referees for their considerable contribution to this process.

Last but not least, we express our deep gratitude to Alberto Apostolico, Massimo Melucci, Angela Visco, and the Department of Information Engineering of the University of Padova for their local arrangement of Dialogues 2004.

October 2004

Setsuo Arikawa
Einoshin Suzuki

# Organization

## Conference Chair

| | |
|---|---|
| Setsuo Arikawa | Kyushu University, Japan |

## Program Committee

| | |
|---|---|
| Einoshin Suzuki (Chair) | Yokohama National University, Japan |
| Elisa Bertino | University of Milan, Italy |
| Wray Buntine | Helsinki Institute of Information Technology, Finland |
| Vincent Corruble | University of Pierre et Marie Curie, Paris, France |
| Manoranjan Dash | Nanyang Technological University, Singapore |
| Luc De Raedt | Albert Ludwigs University of Freiburg, Germany |
| Andreas Dress | Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany |
| Sašo Džeroski | Jožef Stefan Institute, Slovenia |
| Tapio Elomaa | Tampere University of Technology, Finland |
| Johannes Fürnkranz | Technical University of Darmstadt, Germany |
| Gunter Grieser | Technical University of Darmstadt, Germany |
| Fabrice Guillet | École Polytechnique of the University of Nantes, France |
| Mohand-Said Hacid | University of Claude Bernard, Lyon, France |
| Achim Hoffmann | University of New South Wales, Australia |
| Eamonn Keogh | University of California, Riverside, USA |
| Ramamohanarao Kotagiri | University of Melbourne, Australia |
| Aleksandar Lazarević | University of Minnesota, USA |
| Michael May | Fraunhofer Institute for Autonomous Intelligent Systems, Germany |
| Hiroshi Motoda | Osaka University, Japan |
| Jan Rauch | University of Economics, Prague, Czech Republic |
| Domenico Saccá | ICAR-CNR and University of Calabria, Italy |
| Tobias Scheffer | Humboldt University of Berlin, Germany |
| Rudy Setiono | National University of Singapore, Singapore |
| Masayuki Takeda | Kyushu University, Japan |
| Kai Ming Ting | Monash University, Australia |
| Ljupčo Todorovski | Jožef Stefan Institute, Slovenia |
| Hannu Toivonen | University of Helsinki, Finland |
| Akihiro Yamamoto | Kyoto University, Japan |
| Djamel A. Zighed | Lumière University, Lyon, France |

## Local Arrangements

| | |
|---|---|
| Melucci Massimo | University of Padova |

## Subreferees

Helena Ahonen-Myka
Fabrizio Angiulli
Hideo Bannai
Maurice Bernadet
Sourav S. Bhowmick
Steffen Bickel
Ulf Brefeld
Christoph Bscher
Lourdes Peña Castillo
Narendra S. Chaudhari
Damjan Demšar
Isabel Drost
Alfredo Garro
Vivekanand Gopalkrishnan
Mounira Harzallah
Daisuke Ikeda
Akira Ishino
Matti Kääriäinen
Branko Kavšek
Kristian Kersting
Heidi Koivistoinen
Jussi Kujala
Kari Laasonen
Sau Dan Lee
Remi Lehn
Jussi T. Lindgren
Elio Masciari
Taneli Mielikäinen
Phu Chien Nguyen
Riccardo Ortale
Martijn van Otterlo
Luigi Palopoli
Clara Pizzuti
Juho Rousu
Alexandr Savinov
Francesco Scarcello
Alexandre Termier
Charles-David Wajnberg
Tetsuya Yoshida
Bernard Ženko

## Sponsors

Department of Information Engineering of the University of Padova
Yokohama National University
Research Institute on High Performance Computing and Networking,
Italian National Research Council (ICAR-CNR)

# Remembering Carl Smith, 1950-2004

Sadly, Carl Smith passed away 10:30PM, July 21, 2004. He had had a 1.5 year battle with an aggressive brain tumor. He fought this battle with calm optimism, dignity, and grace. He is survived by his wife, Patricia, his son, Austin, and his sister, Karen Martin.

Carl was very active in the algorithmic or computational learning communities, especially in the inductive inference subarea which applies recursive function theory techniques.

I first met Carl when I interviewed for my faculty position at SUNY/Buffalo in the Spring of 1973. He was then a graduate student there and told me he was interested in recursive function theory. After I joined there, he naturally became my Ph.D. student, and that's when we both began working on inductive inference. We spent a lot of time together, pleasantly blurring the distinction between the relationships of friendship and advisor-student.

After Buffalo, Carl had faculty positions at Purdue and, then, the University of Maryland.

Carl had a very productive career. He was a master collaborator working with many teams around the world. Of course he also produced a number of papers about inductive inference by teams — as well as papers about anomalies, queries, memory limitation, procrastination, and measuring mind changes by counting down from notations for ordinals. I had the reaction to some of his papers of wishing I'd thought of the idea. This especially struck me with his 1989 *TCS* paper (with Angluin and Gasarch) in which it is elegantly shown that the learning of some classes of tasks can be done only sequentially after or in parallel with other classes.

Carl played a significant leadership role in theoretical computer science. In 1981, with the help of Paul Young, Carl organized the Workshop on Recursion Theoretic Aspects of Computer Science. This became the well known, continuing series of Computational Complexity conferences. Carl provided an improvement in general theoretical computer science funding level during his year as Theory Program Director at NSF. He was involved, in many cases from the beginning, in the COLT, AII, ALT, EuroCOLT, and DS conferences, as a presenter of papers, as a member of many of their program committees and, in some cases, steering committees. He spearheaded the development of COLT's Mark Fulk Award for best student papers and managed the finances.

Carl was very likable. He had a knack for finding funding to make good things happen. He was a good friend and colleague. He is missed.

August 2004 John Case

# Lecture Notes in Artificial Intelligence (LNAI)

Vol. 3249: B. Buchberger, J.A. Campbell (Eds.), Artificial Intelligence and Symbolic Computation. X, 285 pages. 2004.

Vol. 3245: E. Suzuki, S. Arikawa (Eds.), Discovery Science. XIV, 430 pages. 2004.

Vol. 3244: S. Ben-David, J. Case, A. Maruoka (Eds.), Algorithmic Learning Theory. XIV, 505 pages. 2004.

Vol. 3238: S. Biundo, T. Frühwirth, G. Palm (Eds.), KI 2004: Advances in Artificial Intelligence. XI, 467 pages. 2004.

Vol. 3229: J.J. Alferes, J. Leite (Eds.), Logics in Artificial Intelligence. XIV, 744 pages. 2004.

Vol. 3215: M.G. Negoita, R.J. Howlett, L. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. XXX, 900 pages. 2004.

Vol. 3214: M.G. Negoita, R.J. Howlett, L. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. XXX, 1200 pages. 2004.

Vol. 3213: M.G. Negoita, R.J. Howlett, L. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. XXX, 1200 pages. 2004.

Vol. 3206: P. Sojka, I. Kopecek, K. Pala (Eds.), Text, Speech and Dialogue. XIII, 667 pages. 2004.

Vol. 3202: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Knowledge Discovery in Databases: PKDD 2004. XIX, 560 pages. 2004.

Vol. 3201: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Machine Learning: ECML 2004. XVIII, 580 pages. 2004.

Vol. 3194: R. Camacho, R. King, A. Srinivasan (Eds.), Inductive Logic Programming. XI, 361 pages. 2004.

Vol. 3192: C. Bussler, D. Fensel (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. XIII, 522 pages. 2004.

Vol. 3191: M. Klusch, S. Ossowski, V. Kashyap, R. Unland (Eds.), Cooperative Information Agents VIII. XI, 303 pages. 2004.

Vol. 3187: G. Lindemann, J. Denzinger, I.J. Timm, R. Unland (Eds.), Multiagent System Technologies. XIII, 341 pages. 2004.

Vol. 3176: O. Bousquet, U. von Luxburg, G. Rätsch (Eds.), Advanced Lectures on Machine Learning. IX, 241 pages. 2004.

Vol. 3171: A.L.C. Bazzan, S. Labidi (Eds.), Advances in Artificial Intelligence – SBIA 2004. XVII, 548 pages. 2004.

Vol. 3159: U. Visser, Intelligent Information Integration for the Semantic Web. XIV, 150 pages. 2004.

Vol. 3157: C. Zhang, H. W. Guesgen, W.K. Yeap (Eds.), PRICAI 2004: Trends in Artificial Intelligence. XX, 1023 pages. 2004.

Vol. 3155: P. Funk, P.A. González Calero (Eds.), Advances in Case-Based Reasoning. XIII, 822 pages. 2004.

Vol. 3139: F. Iida, R. Pfeifer, L. Steels, Y. Kuniyoshi (Eds.), Embodied Artificial Intelligence. IX, 331 pages. 2004.

Vol. 3131: V. Torra, Y. Narukawa (Eds.), Modeling Decisions for Artificial Intelligence. XI, 327 pages. 2004.

Vol. 3127: K.E. Wolff, H.D. Pfeiffer, H.S. Delugach (Eds.), Conceptual Structures at Work. XI, 403 pages. 2004.

Vol. 3123: A. Belz, R. Evans, P. Piwek (Eds.), Natural Language Generation. X, 219 pages. 2004.

Vol. 3120: J. Shawe-Taylor, Y. Singer (Eds.), Learning Theory. X, 648 pages. 2004.

Vol. 3097: D. Basin, M. Rusinowitch (Eds.), Automated Reasoning. XII, 493 pages. 2004.

Vol. 3071: A. Omicini, P. Petta, J. Pitt (Eds.), Engineering Societies in the Agents World. XIII, 409 pages. 2004.

Vol. 3070: L. Rutkowski, J. Siekmann, R. Tadeusiewicz, L.A. Zadeh (Eds.), Artificial Intelligence and Soft Computing - ICAISC 2004. XXV, 1208 pages. 2004.

Vol. 3068: E. André, L. Dybkjær, W. Minker, P. Heisterkamp (Eds.), Affective Dialogue Systems. XII, 324 pages. 2004.

Vol. 3067: M. Dastani, J. Dix, A. El Fallah-Seghrouchni (Eds.), Programming Multi-Agent Systems. X, 221 pages. 2004.

Vol. 3066: S. Tsumoto, R. Słowiński, J. Komorowski, J.W. Grzymała-Busse (Eds.), Rough Sets and Current Trends in Computing. XX, 853 pages. 2004.

Vol. 3065: A. Lomuscio, D. Nute (Eds.), Deontic Logic in Computer Science. X, 275 pages. 2004.

Vol. 3060: A.Y. Tawfik, S.D. Goodwin (Eds.), Advances in Artificial Intelligence. XIII, 582 pages. 2004.

Vol. 3056: H. Dai, R. Srikant, C. Zhang (Eds.), Advances in Knowledge Discovery and Data Mining. XIX, 713 pages. 2004.

Vol. 3055: H. Christiansen, M.-S. Hacid, T. Andreasen, H.L. Larsen (Eds.), Flexible Query Answering Systems. X, 500 pages. 2004.

Vol. 3040: R. Conejo, M. Urretavizcaya, J.-L. Pérez-de-la-Cruz (Eds.), Current Topics in Artificial Intelligence. XIV, 689 pages. 2004.

Vol. 3035: M.A. Wimmer (Ed.), Knowledge Management in Electronic Government. XII, 326 pages. 2004.

Vol. 3034: J. Favela, E. Menasalvas, E. Chávez (Eds.), Advances in Web Intelligence. XIII, 227 pages. 2004.

Vol. 3030: P. Giorgini, B. Henderson-Sellers, M. Winikoff (Eds.), Agent-Oriented Information Systems. XIV, 207 pages. 2004.

Vol. 3029: B. Orchard, C. Yang, M. Ali (Eds.), Innovations in Applied Artificial Intelligence. XXI, 1272 pages. 2004.

Vol. 3025: G.A. Vouros, T. Panayiotopoulos (Eds.), Methods and Applications of Artificial Intelligence. XV, 546 pages. 2004.

Vol. 3020: D. Polani, B. Browning, A. Bonarini, K. Yoshida (Eds.), RoboCup 2003: Robot Soccer World Cup VII. XVI, 767 pages. 2004.

Vol. 3012: K. Kurumatani, S.-H. Chen, A. Ohuchi (Eds.), Multi-Agnets for Mass User Support. X, 217 pages. 2004.

Vol. 3010: K.R. Apt, F. Fages, F. Rossi, P. Szeredi, J. Váncza (Eds.), Recent Advances in Constraints. VIII, 285 pages. 2004.

Vol. 2990: J. Leite, A. Omicini, L. Sterling, P. Torroni (Eds.), Declarative Agent Languages and Technologies. XII, 281 pages. 2004.

Vol. 2980: A. Blackwell, K. Marriott, A. Shimojima (Eds.), Diagrammatic Representation and Inference. XV, 448 pages. 2004.

Vol. 2977: G. Di Marzo Serugendo, A. Karageorgos, O.F. Rana, F. Zambonelli (Eds.), Engineering Self-Organising Systems. X, 299 pages. 2004.

Vol. 2972: R. Monroy, G. Arroyo-Figueroa, L.E. Sucar, H. Sossa (Eds.), MICAI 2004: Advances in Artificial Intelligence. XVII, 923 pages. 2004.

Vol. 2969: M. Nickles, M. Rovatsos, G. Weiss (Eds.), Agents and Computational Autonomy. X, 275 pages. 2004.

Vol. 2961: P. Eklund (Ed.), Concept Lattices. IX, 411 pages. 2004.

Vol. 2953: K. Konrad, Model Generation for Natural Language Interpretation and Analysis. XIII, 166 pages. 2004.

Vol. 2934: G. Lindemann, D. Moldt, M. Paolucci (Eds.), Regulated Agent-Based Social Systems. X, 301 pages. 2004.

Vol. 2930: F. Winkler (Ed.), Automated Deduction in Geometry. VII, 231 pages. 2004.

Vol. 2926: L. van Elst, V. Dignum, A. Abecker (Eds.), Agent-Mediated Knowledge Management. XI, 428 pages. 2004.

Vol. 2923: V. Lifschitz, I. Niemelä (Eds.), Logic Programming and Nonmonotonic Reasoning. IX, 365 pages. 2004.

Vol. 2915: A. Camurri, G. Volpe (Eds.), Gesture-Based Communication in Human-Computer Interaction. XIII, 558 pages. 2004.

Vol. 2913: T.M. Pinkston, V.K. Prasanna (Eds.), High Performance Computing - HiPC 2003. XX, 512 pages. 2003.

Vol. 2903: T.D. Gedeon, L.C.C. Fung (Eds.), AI 2003: Advances in Artificial Intelligence. XVI, 1075 pages. 2003.

Vol. 2902: F.M. Pires, S.P. Abreu (Eds.), Progress in Artificial Intelligence. XV, 504 pages. 2003.

Vol. 2892: F. Dau, The Logic System of Concept Graphs with Negation. XI, 213 pages. 2003.

Vol. 2891: J. Lee, M. Barley (Eds.), Intelligent Agents and Multi-Agent Systems. X, 215 pages. 2003.

Vol. 2882: D. Veit, Matchmaking in Electronic Markets. XV, 180 pages. 2003.

Vol. 2871: N. Zhong, Z.W. Raś, S. Tsumoto, E. Suzuki (Eds.), Foundations of Intelligent Systems. XV, 697 pages. 2003.

Vol. 2854: J. Hoffmann, Utilizing Problem Structure in Planing. XIII, 251 pages. 2003.

Vol. 2843: G. Grieser, Y. Tanaka, A. Yamamoto (Eds.), Discovery Science. XII, 504 pages. 2003.

Vol. 2842: R. Gavaldá, K.P. Jantke, E. Takimoto (Eds.), Algorithmic Learning Theory. XI, 313 pages. 2003.

Vol. 2838: N. Lavrač, D. Gamberger, L. Todorovski, H. Blockeel (Eds.), Knowledge Discovery in Databases: PKDD 2003. XVI, 508 pages. 2003.

Vol. 2837: N. Lavrač, D. Gamberger, L. Todorovski, H. Blockeel (Eds.), Machine Learning: ECML 2003. XVI, 504 pages. 2003.

Vol. 2835: T. Horváth, A. Yamamoto (Eds.), Inductive Logic Programming. X, 401 pages. 2003.

Vol. 2821: A. Günter, R. Kruse, B. Neumann (Eds.), KI 2003: Advances in Artificial Intelligence. XII, 662 pages. 2003.

Vol. 2807: V. Matoušek, P. Mautner (Eds.), Text, Speech and Dialogue. XIII, 426 pages. 2003.

Vol. 2801: W. Banzhaf, J. Ziegler, T. Christaller, P. Dittrich, J.T. Kim (Eds.), Advances in Artificial Life. XVI, 905 pages. 2003.

Vol. 2797: O.R. Zaïane, S.J. Simoff, C. Djeraba (Eds.), Mining Multimedia and Complex Data. XII, 281 pages. 2003.

Vol. 2792: T. Rist, R.S. Aylett, D. Ballin, J. Rickel (Eds.), Intelligent Virtual Agents. XV, 364 pages. 2003.

Vol. 2782: M. Klusch, A. Omicini, S. Ossowski, H. Laamanen (Eds.), Cooperative Information Agents VII. XI, 345 pages. 2003.

Vol. 2780: M. Dojat, E. Keravnou, P. Barahona (Eds.), Artificial Intelligence in Medicine. XIII, 388 pages. 2003.

Vol. 2777: B. Schölkopf, M.K. Warmuth (Eds.), Learning Theory and Kernel Machines. XIV, 746 pages. 2003.

Vol. 2752: G.A. Kaminka, P.U. Lima, R. Rojas (Eds.), RoboCup 2002: Robot Soccer World Cup VI. XVI, 498 pages. 2003.

Vol. 2741: F. Baader (Ed.), Automated Deduction – CADE-19. XII, 503 pages. 2003.

Vol. 2705: S. Renals, G. Grefenstette (Eds.), Text- and Speech-Triggered Information Access. VII, 197 pages. 2003.

Vol. 2703: O.R. Zaïane, J. Srivastava, M. Spiliopoulou, B. Masand (Eds.), WEBKDD 2002 - MiningWeb Data for Discovering Usage Patterns and Profiles. IX, 181 pages. 2003.

Vol. 2700: M.T. Pazienza (Ed.), Extraction in the Web Era. XIII, 163 pages. 2003.

Vol. 2699: M.G. Hinchey, J.L. Rash, W.F. Truszkowski, C.A. Rouff, D.F. Gordon-Spears (Eds.), Formal Approaches to Agent-Based Systems. IX, 297 pages. 2002.

Vol. 2691: V. Mařík, J.P. Müller, M. Pechoucek (Eds.), Multi-Agent Systems and Applications III. XIV, 660 pages. 2003.

# Table of Contents

## Long Papers

### Pattern Mining

### Classification

### Outlier Detection

### Clustering

## Feature Construction and Generation

## Knowledge Acquisition

## Discovery Science in Reality

## Regular Papers

### Pattern Mining

### Machine Learning Algorithms

### Web Mining

## Applications of Predictive Methods

## Interdisciplinary Approaches

# Predictive Graph Mining

Andreas Karwath and Luc De Raedt

Albert-Ludwigs-Universität Freiburg, Institut für Informatik,
Georges-Köhler-Allee 079, D-79110 Freiburg, Germany
{karwath,deraedt}@informatik.uni-freiburg.de

**Abstract.** Graph mining approaches are extremely popular and effective in molecular databases. The vast majority of these approaches first derive interesting, i.e. frequent, patterns and then use these as features to build predictive models. Rather than building these models in a two step indirect way, the SMIREP system introduced in this paper, derives predictive rule models from molecular data directly. SMIREP combines the SMILES and SMARTS representation languages that are popular in computational chemistry with the IREP rule-learning algorithm by Fürnkranz. Even though SMIREP is focused on SMILES, its principles are also applicable to graph mining problems in other domains. SMIREP is experimentally evaluated on two benchmark databases.

## 1 Introduction

In recent years, the problem of graph mining in general, and its application to chemical and biological problems, has become an active research area in the field of data-mining. The vast majority of graph mining approaches first derives interesting, i.e. frequent, patterns and then uses these as features to build predictive models. Several approaches have been suggested [1–8] for the task of identifying fragments which can be used to build such models. The earliest approaches to compute such fragments are based on techniques from inductive logic programming [1]. Whereas inductive logic programming techniques are theoretically appealing because of the use of expressive representation languages, they exhibit significant efficiency problems, which in turn implies that their application has been restricted to finding relatively small fragments in relatively small databases. Recently proposed approaches to mining frequent fragments in graphs such as gSpan [5], CloseGraph [9], FSG [2], and AGM [7] are able to mine complex subgraphs more efficiently. However, the key difficulty with the application of these techniques is – as for other frequent pattern mining approaches – the number of patterns that are generated. Indeed, [6] report on $10^6$ of patterns being discovered. Furthermore, frequent fragments are not necessarily of interest to a molecular scientist. Therefore, [3] and [6] propose approaches that take into account the classes of the molecules. Kramer *et al.* compute all simple patterns that are frequent in the positives (or actives) and infrequent in the negatives (or inactives) [3]. Inokuchi *et al.* compute correlated patterns [6] . While Inokuchi *et al.* claim that the discovered patterns can be used for predictive purposes, they do not report on any predictive accuracies.

E. Suzuki and S. Arikawa (Eds.): DS 2004, LNAI 3245, pp. 1–15, 2004.

The approach taken here is different: SMIREP produces predictive models (in the form of rule-sets) directly. SMIREP combines SMILES, a chemical representation language that allows the representation of complex graph structures as strings, with IREP [10,11], a well-known rule-learner from the field of machine learning. SMIREP has been applied to two benchmark data sets (the mutagenicity dataset [12] and the AIDS Antiviral Screening Database [3]) and the experiments show that SMIREP produces *small* rule sets containing possibly complex fragments, that SMIREP is competitive in terms of predictive accuracy and that SMIREP is quite efficient as compared to other data mining systems.

Although, the SMIREP system is tailored towards the chemical domain, the approach can be employed as a general approach to build predictive models for graph mining domains.

The paper is organized as follows: in section 2 we give an overview of SMILES and SMARTS as language for chemical compounds and fragments as well as their applicability to other structured data, like non-directed graphs; section 3 gives an overview of the SMIREP system; in section 4, we report on experiments and findings conducted with SMIREP;finally, in section 5, we touch upon related work and conclude.

## 2 SMILES and SMARTS

SMILES (Simplified Molecular Input Line Entry System) [13] is a linear string representation language for chemical molecules. The SMILES language is commonly used in computational chemistry and is supported by most major software tools in the field, like the commercial Daylight toolkit or the Open-Source OpenBabel library.

The SMILES notations of chemical compounds are comprised of atoms, bonds, parathesis, and numbers:

- *Atoms:* Atoms are represented using their atomic symbols. E.g. C for carbon, N for nitrogen, or S for sulfur. For aromatic atoms, lower case letters are used, and upper case letters otherwise. Atoms with two letter symbols, like chlorine (Cl) or bromine (Br), are always written with the first letter in upper case and the second letter can be written upper or lower case. With a rare few exceptions, hydrogen atoms are not included in the string representation of a molecule.
- *Bonds:* Four basic bond types are used in the SMILES language: single, double, triple, and aromatic bonds, represented by the symbols: '-', '=', '#', and ':' respectively. Single and aromatic bonds are usually omitted from SMILES strings. Not belonging to the four basic bonds are ionic bonds, or *disconnections*, represented by a '.'.
- *Branches:* Branches are specified by enclosing brackets, "(" and ")", and indicate side-structures. A branch can, and often does, contain other branches.
- *Cycles:* Cyclic structures are represented by breaking one bond in each ring. The atoms adjacent to the bond obtain the same number. E.g. 'cccccc' denotes a (linear) sequence of six aromatic carbons and 'c1ccccc1' denotes