# Mining Structured Data

Nijssen, Siegfried Gerardus Remius

# Bibliography


[1] National Cancer Institute (NCI), DTP/2D and 3D structural information, `http://cactus.nci.nih.gov/ncidb2/download.html`.

[2] OpenBabel, `http://openbabel.sourceforge.net`.

[3] The predictive toxicology evaluation challenge, `http://web.comlab.ox.ac.uk/~research/areas/machlearn/pte/`.

[4] Valgrind, `http://www.valgrind.org`.

[5] P. Adriaans and D. Zantinge. *Data Mining*. Addison-Wesley, 1996.

[6] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 207–216. ACM Press, 1993.

[7] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI Press, 1996.

[8] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*, pages 487–499. Morgan Kaufmann Publishers, 1994.

[9] A. V. Aho, J. E. Hopcroft, and J. E. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, 1974.

[10] H. Albert-Lorincz and J.-F. Boulicaut. Mining frequent sequential patterns under regular expressions: a highly adaptive strategy for pushing constraints. In *Proceedings of the Third SIAM International Conference on Data Mining (SDM)*, pages 316–320. SIAM, 2003.

[11] K. Apt and E. Marchiori. Reasoning about Prolog programs: from modes through types to assertions. In *Formal Aspects of Computing*, volume 6, pages 743–765. Springer-Verlag, 1994.

[12] T. Asai, K. Abe, S. Kawasoe, H. Arimura, H. Satamoto, and S. Arikawa. Efficient substructure discovery from large semi-structured data. In *Proceedings of the Second SIAM International Conference on Data Mining*, pages 158–174. SIAM, 2002.

[13] T. Asai, H. Arimura, T. Uno, and S. Nakano. Discovering frequent substructures in large unordered trees. In *Proceedings of the 6th International Conference on Discovery Science (DS)*, volume 2843 of *Lecture Notes in Computer Science*, pages 47–61. Springer-Verlag, 2003.

[14] S. D. Bay and M. J. Pazzani. Detecting change in categorical data: Mining contrast sets. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 302–306. ACM Press, 1999.

[15] S. D. Bay and M. J. Pazzani. Detecting group differences: Mining contrast sets. In *Data Mining and Knowledge Discovery*, volume 5, pages 213–246. Kluwer Academic Publishers, 2001.

[16] R. J. Bayardo. Efficiently mining long patterns from databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 85–93. ACM Press, 1998.

[17] P. Berka. Workshop notes on discovery challenge PKDD-99. Technical report, University of Economics, Prague, Czech Republic, 1999.

[18] T. Beyer and S. M. Hedetniemi. Constant time generation of rooted trees. In *SIAM Journal of Computing*, volume 9, pages 706–711. SIAM, 1980.

[19] H. Blockeel and L. De Raedt. Top-down induction of first-order logical decision trees. In *Artificial Intelligence*, volume 101, pages 285–297. Elsevier Science, 1997.

[20] H. Blockeel, L. De Raedt, N. Jacobs, and B. Demoen. Scaling up inductive logic programming by learning from interpretations. In *Data Mining and Knowledge Discovery*, volume 3, pages 59–93. Kluwer Acadamic Publishers, 1999.

[21] H. Blockeel, L. Dehaspe, B. Demoen, G. Janssens, J. Ramon, and H. Vandecasteele. Improving the efficiency of inductive logic programming through the use of query packs. In *Journal of Artificial Intelligence Research*, volume 16, pages 135–166. AAAI Press, 2002.

[22] H. Blockeel, L. Dehaspe, J. Ramon, and J. Struyf. The ACE data mining system–User's manual. Technical report, Katholieke Universiteit Leuven, Belgium, 2004.

[23] F. Bonchi and B. Goethals. FP-Bonsai: the art of growing and pruning small FP-Trees. In *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, volume 3056 of *Lecture Notes in Computer Science*, pages 155–160. Springer-Verlag, 2004.

[24] C. Borgelt and M. R. Berthold. Mining molecular fragments: Finding relevant substructures of molecules. In *Proceedings of the Second IEEE International Conference on Data Mining (ICDM)*, pages 51–58. IEEE Press, 2002.

[25] C. Borgelt and R. Kruse. Induction of association rules: Apriori implementation. In *Proceedings of the 15th Conference on Computational Statistics (Compstat)*, pages 395–400. Physica Verlag, 2002.

[26] J.-F. Boulicaut, A. Bykowski, and C. Rigotti. Free-sets: a condensed representation of boolean data for the approximation of frequency queries. In *Data Mining and Knowledge Discovery, 7(1)*, pages 5–22. Kluwer Academic Publishers, 2003.

[27] J.-F. Boulicaut and B. Jeudy. Mining free itemsets under constraints. In *Proceedings of the International Database Engineering and Applications Symposium*, pages 322–329. ACM Press, 2001.

[28] J.-F. Boulicaut and B. Jeudy. Constraint-based data mining. In *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. Kluwer Academic Publishers, 2005.

[29] U. Brandes, M. Eiglsperger, I. Herman, M. Himsolt, and M. Marshall. GraphML progress report: Structural layer proposal. In *Proceedings of the 9th International Symposium on Graph Drawing (GD)*, volume 2265 of *Lecture Notes in Computer Science*, pages 501–512. Springer-Verlag, 2001.

[30] B. Bringmann. Matching in frequent tree discovery. In *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM)*, pages 335–338. IEEE Press, 2004.

[31] B. Bringmann, L. De Raedt, and T. Horvath. Mining frequent hypergraphs, 2005. Personal communication.

[32] C. Bucilă, J. Gehrke, D. Kifer, and W. White. Dualminer: A dual-pruning algorithm for itemsets with constraints. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 42–51. ACM Press, 2002.

[33] T. Calders and B. Goethals. Mining all non-derivable frequent itemsets. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*, volume 2431 of *Lecture Notes in Computer Science*, pages 74–84. Springer-Verlag, 2002.

[34] G. Casas-Garriga. Towards a formal framework for mining general patterns from structured data. In *Proceedings of the 2nd Workshop on Multi-Relational Data Mining*, 2003.

[35] R. Chalmers and K. Almeroth. On the topology of multicast trees. In *IEEE/ACM Transactions on Networking*, volume 11, pages 153–165. IEEE Press and ACM Press, 2003.

[36] W. Chen. More efficient algorithm for ordered tree inclusion. In *Journal of Algorithms*, volume 26, pages 370–385. Elsevier Science, 1998.

[37] Y. Chi, S. Nijssen, R. R. Muntz, and J. N. Kok. Frequent subtree mining—An overview. In *Fundamenta Informaticae*, volume 66, pages 161–198. IOS Press, 2005.

[38] Y. Chi, Y. Yang, and R. R. Muntz. Indexing and mining free trees. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM)*, pages 509–512. IEEE Press, 2003.

[39] Y. Chi, Y. Yang, and R. R. Muntz. HybridTreeMiner: An efficient algorithm for mining frequent rooted trees and free trees using canonical forms. In *Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM)*, pages 11–20. IEEE Press, 2004.

[40] Y. Chi, Y. Yang, Y. Xia, and R. R. Muntz. CMTreeMiner: Mining both closed and maximal frequent subtrees. In *Proceedings of the Eighth Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, volume 3056 of *Lecture Notes in Computer Science*, pages 63–73. Springer-Verlag, 2004.

[41] M. J. Chung. $O(n^{2.5})$ time algorithms for the subgraph homeomorphism problem on trees. In *Journal of Algorithms*, volume 8, pages 106–112. Elsevier Science, 1987.

[42] A. Clare and R. King. Data mining the yeast genome in a lazy functional language. In *Proceedings of the 5th International Symposium on Practical Aspects of Declarative Languages (PADL)*, volume 2562 of *Lecture Notes In Computer Science*, pages 19–36. Springer-Verlag, 2003.

[43] A. Clare, H. Williams, and N. Lester. Scalable multi-relational association mining. In *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM)*, pages 355–358. IEEE Press, 2004.

[44] P. Clark and T. Niblett. The CN2 induction algorithm. In *Machine Learning*, volume 3, pages 261–283. Kluwer Academic Publishers, 1989.

[45] M. Cohen and E. Gudes. Diagonally subgraphs pattern mining. In *Proceedings of the ACM SIGMOD Workshop on Research issues in data mining and knowledge discovery*, pages 51–58. ACM Press, 2004.

[46] D. J. Cook and L. B. Holder. Substructure discovery using minimum description length and background knowledge. In *Journal of GG Research*, volume 1, pages 231–255. AAAI Press, 1994.

[47] V. S. Costa, A. Srinivasan, R. Camacho, H. Blockeel, B. Demoen, G. Janssens, J. Struyf, H. Vandecasteele, and W. V. Laer. Query transformations for improving the efficiency of ILP systems. In *Journal of Machine Learning Research*, volume 4, pages 465–491. MIT Press, 2003.

[48] T. Dasu, T. Johnson, S. Muthukrishnan, and V. Shkapenyuk. Mining database structure; or, how to build a data quality browser. In *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, pages 240–251. ACM Press, 2002.

[49] L. De Raedt. Data mining as constraint logic programming. In *From Logic Programming into the Future (In honour of Bob Kowalski)*, pages 526–547. Springer-Verlag, 2002.

[50] L. De Raedt. A perspective on inductive databases. In *SIGKDD Explorations*, volume 4, pages 69–77. ACM Press, 2003.

[51] L. De Raedt, H. Blockeel, L. Dehaspe, and W. Van Laer. Three companions for data mining in first order logic. In *Relational Data Mining*, pages 105–137. Springer-Verlag, 2001.

[52] L. De Raedt and L. Dehaspe. Clausal discovery. In *Machine Learning*, volume 26, pages 99–146. Kluwer Academic Publishers, 1997.

[53] L. De Raedt and S. Džeroski. First order *jk*-clausal theories are PAC-learnable. In *Artificial Intelligence*, volume 70, pages 375–392. Elsevier Science, 1994.

[54] L. De Raedt, M. Jaeger, S. D. Lee, and H. Mannila. A theory of inductive query answering (extended abstract). In *Proceedings of the Second IEEE International Conference on Data Mining (ICDM)*, pages 123–130. IEEE Press, 2002.

[55] L. De Raedt and S. Kramer. The levelwise version space algorithm and its application to molecular fragment finding. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 853–862. Morgan Kaufmann Publishers, 2001.

[56] L. De Raedt and J. Ramon. Condensed representations for inductive logic programming. In *Proceedings of Ninth International Conference on the Principles of Knowledge Representation and Reasoning*, pages 438–446. AAAI Press, 2004.

[57] L. De Raedt and W. Van Laer. Inductive constraint logic. In *Proceedings of the Sixth International Workshop on Algorithmic Learning Theory*, volume 997 of *Lecture Notes in Artificial Intelligence*, pages 80–94. Springer-Verlag, 1995.

[58] L. Dehaspe and L. De Raedt. Dlab: A declarative language bias formalism. In *Proceedings of the Ninth International Symposium on Methodologies for Intelligent Systems (ISMIS)*, volume 1079 of *Lecture Notes in Artificial Intelligence*, pages 613–622. Springer-Verlag, 1996.

[59] L. Dehaspe and L. De Raedt. Mining association rules with multiple relations. In *Proceedings of the 7th International Workshop on Inductive Logic Programming*, volume 1297 of *Lecture Notes in Artificial Intelligence*, pages 125–132. Springer-Verlag, 1997.

[60] L. Dehaspe and H. Toivonen. Frequent query discovery: a unifying ILP approach to association rule mining. Technical Report CW-258, Katholieke Universiteit Leuven, Belgium, 1998.

[61] L. Dehaspe and H. Toivonen. Discovery of frequent Datalog patterns. In *Data Mining and Knowledge Discovery*, volume 3, pages 7–36. Kluwer Acadamic Publishers, 1999.

[62] L. Dehaspe, H. Toivonen, and R. D. King. Finding frequent substructures in chemical compounds. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 30–36. AAAI Press, 1998.

[63] M. Deshpande, M. Kuramochi, and G. Karypis. Frequent sub-structure based approaches for classifying chemical compounds. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM)*, pages 35–42. IEEE Press, 2003.

[64] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 43–52. ACM Press, 1999.

[65] G. Dong and J. Li. Mining border descriptions of emerging patterns from dataset pairs. In *Knowledge and Information Systems*, volume 8, pages 178–202. Springer-Verlag, 2005.

[66] G. Dong, X. Zhang, L. Wong, and J. Li. CAEP: Classification by aggregating emerging patterns. In *Discovery Science*, volume 1721 of *Lecture Notes in Computer Science*, pages 30–42. Springer-Verlag, 1999.

[67] M. Dunham. *Data Mining: Introductory and Advanced Topics*. Prentice-Hall, 2003.

[68] D. Eppstein. Subgraph isomorphism in planar graphs and related problems. In *Journal of Graph Algorithms and Applications*, volume 3, pages 1–27. World Scientific Publishing, 1999.

[69] F. Esposito, N. Fanizzi, S. Ferilli, and G. Semeraro. A generalization model based on OI-implication for ideal theory refinement. In *Fundamenta Informaticae*, volume 47, pages 15–33. IOS Press, 2001.

[70] J. Fischer and L. De Raedt. Towards optimizing conjunctive inductive queries. In *Proceedinges of the Eigth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, volume 3056 of *Lecture Notes in Computer Science*, pages 625–637. Springer-Verlag, 2004.

[71] P. Foggia, C. Sansone, and M. Vento. An improved algorithm for matching large graphs. In *Proceedings of the Third International Workshop on Graph-based Representation in Pattern Recognition*, pages 176–187, 2001.

[72] J. Fürnkranz and P. Flach. ROC 'n' rule learning — towards a better understanding of covering algorithms. In *Machine Learning*, volume 58, pages 39–77. Kluwer Academic Publishers, 2005.

[73] M. R. Garey and D. S. Johnson. *Computers and Intractability*. Freeman, 1979.

[74] M. N. Garofalakis, R. Rastogi, and K. Shim. SPIRIT: sequential pattern mining with regular expression constraints. In *Proceedings of the 25th International Conference on Very Large Databases (VLDB)*, pages 223–234. Morgan Kaufmann Publishers, 1998.

[75] B. Goethals and J. Van den Bussche. Relational association rules: getting WARMER. In *Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery in Data Mining*, volume 2447 of *Lecture Notes in Artificial Intelligence*, pages 125–139. Springer-Verlag, 2002.

[76] B. Goethals and M. J. Zaki. Advances in frequent itemset mining implementations: report on FIMI'03. In *SIGKDD Explorations Newsletter*, volume 6, pages 109–117. ACM Press, 2004.

[77] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 1–12. ACM Press, 2000.

[78] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, 2001.

[79] N. Helft. Induction as nonmonotonic inference. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 149–156. Morgan Kaufmann Publishers, 1989.

[80] H. Hirsh. Theoretical underpinnings of version spaces. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 665–670. Morgan Kaufmann Publishers, 1991.

[81] H. Hofer, C. Borgelt, and M. Berthold. Large scale mining of molecular fragments with wildcards. In *Proceedings of the 5th International Symposium on Intelligent Data Analysis (IDA)*, pages 380–389. Springer-Verlag, 2003.

[82] J. Hopcroft and R. Karp. A $n^{\frac{5}{2}}$ algorithm for maximum matching in bipartite graphs. In *SIAM Journal of Computing*, volume 2, pages 225–231. SIAM, 1973.

[83] J. Hopcroft and R. Tarjan. Isomorphism of planar graphs. In *Complexity of Computer Computations*, pages 131–152. Plenum Press, 1972.

[84] J. Hopcroft and R. Tarjan. Efficient planarity testing. In *Journal of the Association for Computing Machinery*, volume 21, pages 549–568. ACM Press, 1974.

[85] J. Huan, W. Wang, D. Bandyopadhyay, J. Snoeyink, J. Prins, and A. Tropsha. Mining family specific residue packing patterns from protein structure graphs. In *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 308–315. ACM Press, 2004.

[86] J. Huan, W. Wang, and J. Prins. Efficient mining of frequent subgraphs in the presence of isomorphism. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM)*, pages 549–552. IEEE Press, 2003.

[87] J. Huan, W. Wang, J. Prins, and J. Yang. SPIN: Mining maximal frequent subgraphs from graph databases. In *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 581–586. ACM Press, 2004.

[88] J. Huan, W. Wang, J. Prins, and J. Yang. SPIN: Mining maximal frequent subgraphs from graph databases. Technical report, University of North Carolina at Chapel Hill, USA, 2005.

[89] Y. Huhtala, J. Kärkkäinen, P. Porkka, and H. Toivonen. Efficient discovery of functional and approximate dependencies using partitions. In *Proceedings of the 14th International Conference on Data Engineering (ICDE)*, pages 392–401. IEEE Press, 1998.

[90] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. In *Communications of the ACM*, volume 39, pages 58–64. ACM Press, 1996.

[91] A. Inokuchi. Mining generalized substructures from a set of labeled graphs. In *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM)*, pages 415–418. IEEE Press, 2004.

[92] A. Inokuchi, T. Washio, and H. Motoda. An APRIORI-based algorithm for mining frequent substructures from graph data. In *Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, volume 1910 of *Lecture Notes in Artificial Intelligence*, pages 13–23. Springer-Verlag, 2000.

[93] A. Inokuchi, T. Washio, K. Nishimura, and H. Motoda. A fast algorithm for mining frequent connected subgraphs. Technical Report RT0448, IBM Research, Tokyo Research Laboratory, 2002.

[94] A. Inokuchi, T. Washio, T. Okada, and H. Motoda. Applying the APRIORI-based graph mining method to mutagenesis data analysis. In *Journal of Computer Aided Chemistry*, volume 15, pages 87–92. Kluwer Academic Publishers, 2001.

[95] B. Kavšek, N. Lavrač, and V. Jovanoski. APRIORI-SD: Adapting association rule learning to subgroup discovery. In *Proceedings of the Fifth International Symposium on Intelligent Data Analysis*, volume 2810 of *Lecture Notes in Computer Science*, pages 230–241. Springer-Verlag, 2003.

[96] P. Kilpeläinen. Tree matching problems with applications to structured text database (Ph.D. dissertation). Technical Report A-1992-6, University of Helsinki, Finland, 1992.

[97] R. D. King, A. Srinivasan, and L. Dehaspe. WARMR: A data mining tool for chemical data. In *Journal of Computer Aided Molecular Design*, volume 15, pages 173–181. Kluwer Academic Publishers, 2001.

[98] G. Kirchhoff. Uber die Ausflösung der Gleichungen auf welche man bei der Untersuchungen der linearen Verteilung galvanisher Ströme geführt wird. In *Poggendorf's Annalen der Physik und Chemie*, volume 72, pages 497–508, 1847.

[99] W. Klösgen. EXPLORA: A multipattern and multistrategy discovery assistant. In *Advances in Knowledge Discovery and Data Mining*, pages 249–271. AAAI Press, 1996.

[100] A. Knobbe. *Multi-Relational Data Mining (Ph.D. dissertation)*. Utrecht University, The Netherlands, 2004.

[101] A. Knobbe, A. Siebes, H. Blockeel, and D. van der Wallen. Multi-relational data mining, using UML for ILP. In *Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, volume 1910 of *Lecture Notes in Artificial Intelligence*, pages 1–12. Springer-Verlag, 2000.

[102] D. Knuth. *The Art of Computer Programming (fascicles)*, volume 4. Addison-Wesley, 2005.

[103] D. Knuth, J. Morris, and V. Pratt. Fast pattern matching in strings. In *SIAM Journal on Computing*, volume 6, pages 323–350. SIAM, 1977.

[104] W. Kosters, W. Pijls, and V. Popova. Complexity analysis of depth first and FP-GROWTH implementations of APRIORI. In *Machine Learning and Data Mining in Pattern Recognition (MLDM)*, volume 2734 of *Lecture Notes in Artificial Intelligence*, pages 284–292. Springer-Verlag, 2003.

[105] M. Kryszkiewicz and M. Gajek. Concise representation of frequent patterns based on generalized disjunction-free generators. In *Proceedinges of the Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, volume 2336 of *Lecture Notes in Computer Science*, pages 159–171. Springer-Verlag, 2002.

[106] T. Kudo. An implementation of FREQT, 2003. `http://chasen.org/~taku/software/freqt/`.

[107] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *Proceedings of the First IEEE International Conference on Data Mining (ICDM)*, pages 313–320. IEEE Press, 2001.

[108] M. Kuramochi and G. Karypis. Discovering frequent geometric subgraphs. In *Proceedings of the Second IEEE International Conference on Data Mining (ICDM)*, pages 258–265. IEEE Press, 2002.

[109] M. Kuramochi and G. Karypis. An efficient algorithm for discovering frequent subgraphs. Technical Report 02-026, University of Minnesota, USA, 2002.

[110] M. Kuramochi and G. Karypis. Finding frequent patterns in a large sparse graph. In *Proceedings of the Fourth SIAM International Conference on Data Mining*. SIAM, 2004.

[111] M. Kuramochi and G. Karypis. GREW—a scalable frequent subgraph discovery algorithm. In *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM)*, pages 439–442. IEEE Press, 2004.

[112] N. Lavrač, P. Flach, and B. Zupan. Rule evaluation measures: A unifying view. In *Proceedings of the 9th International Workshop on Inductive Logic Programming*, volume 1634 of *Lecture Notes in Computer Science*, pages 174–185. Springer-Verlag, 1999.

[113] N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski. Subgroup discovery with CN2-SD. In *Journal of Machine Learning Research*, volume 5, pages 153–188. MIT Press, 2004.

[114] S. D. Lee and L. De Raedt. An algebra for inductive query evaluation. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM)*, pages 147–154. IEEE Press, 2003.

[115] S. D. Lee and L. De Raedt. Constraint based mining of first-order sequences in SeqLog. In *Database Support for Data Mining Applications*, volume 2682 of *Lecture Notes in Computer Science*, pages 155–176. Springer-Verlag, 2004.

[116] G. Li and F. Ruskey. The advantages of forward thinking in generating rooted and free trees. In *Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 939–940. SIAM, 1999.

[117] W. Li, J. Han, and J. Pei. CMAR: Accurate and efficient classification based on multiple class-association rules. In *Proceedings of the First IEEE International Conference on Data Mining (ICDM)*, pages 369–376. IEEE Press, 2001.

[118] F. Lisi, S. Ferilli, and N. Fanizzi. Object identity as search bias for pattern spaces. In *Proceedings of the Fifteenth European Conference on Artificial Intelligence (ECAI)*, pages 375–379. IOS Press, 2002.

[119] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 80–86. ACM Press, 1998.

[120] B. Liu, Y. Ma, and C.-K.Wong. Improving an exhaustive search based rule learner. In *Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, volume 1910 of *Lecture Notes in Artificial Intelligence*, pages 504–509. Springer-Verlag, 2000.

[121] B. Lui, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 125–134. ACM Press, 1999.

[122] D. Malerba and F. Lisi. Discovering associations between spatial objects: An ILP application. In *Proceedings of the 11th International Conference on Inductive Logic Programming (ILP)*, volume 2157 of *Lecture Notes in Artificial Intelligence*, pages 156–163. Springer-Verlag, 2001.

[123] D. Malerba and F. Lisi. Inducing multi-level association rules from multiple relations. In *Machine Learning*, volume 55, pages 175–210. Kluwer Academic Publishers, 2004.

[124] H. Mannila and H. Toivonen. Multiple uses of frequent sets and condensed representations. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 189–194. AAAI Press, 1996.

[125] H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. In *Data Mining and Knowledge Discovery*, volume 1, pages 241–258. Kluwer Academic Publishers, 1997.

[126] H. Mannila, H. Toivonen, and A. Verkamo. Efficient algorithms for discovering association rules. In *Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases (KDD)*, pages 181–192. AAAI Press, 1994.

[127] M. P. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 114–119. Morgan Kauffman Publishers, 1994.

[128] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. In *Computational Linguistics*, volume 19, pages 313–330. MIT Press, 1993.

[129] T. Matsuda, T. Horiuchi, H. Motoda, and T. Washio. Extension of graph-based induction for general graph structured data. In *Proceedinges of the Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, volume 1805 of *Lecture Notes in Computer Science*, pages 420–431. Springer-Verlag, 2000.

[130] D. W. Matula. An algorithm for subtree identification. In *SIAM Review*, volume 10, pages 273–274. SIAM, 1968.

[131] B. McKay. Practical graph isomorphism. In *Congressus Numerantium*, volume 30, pages 45–87. Utilitas Mathematica Publishing, 1981.

[132] C. Mellish. The automatic generation of mode declarations for Prolog programs. Technical Report 163, Department of Artificial Intelligence, University of Edinburgh, UK, 1981.

[133] R. Meo, G. Psaila, and S. Ceri. An extension to SQL for mining association rules in SQL. In *Data Mining and Knowledge Discovery*, volume 2, pages 195–224. Kluwer Academic Publishers, 1998.

[134] T. Mitchell. Generalization as search. In *Artificial Intelligence*, volume 18, pages 203–226. Elsevier Science, 1982.

[135] V. Morell. TreeBASE: The roots of phylogeny. In *Science*, volume 273, page 569, 1996.

[136] S. Morishita and J. Sese. Traversing itemset lattices with statistical metric pruning. In *Proceedings of the Nineteenth ACM SIGACT-SIGMOD-SIGART Symposium on Database Systems (PODS)*, pages 226–236. ACM Press, 2000.

[137] S. Muggleton. Inverse entailment and Progol. In *New Generation Computing*, volume 13, pages 245–286. Springer-Verlag, 1995.

[138] S. Muggleton and L. De Raedt. Inductive logic programming: Theory and methods. In *Journal of Logic Programming*, volume 19/20, pages 629–679. Elsevier Science, 1994.

[139] S. Nakano and T. Uno. Efficient generation of rooted trees. Technical Report 2003-005E, National Institute of Informatics, Japan, 2003.

[140] R. Ng, L. V. S. Lakshmanan, J. Han, and T. Mah. Exploratory mining via constrained frequent set queries. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 556–558. ACM Press, 1999.

[141] R. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained associations rules. In *Proceedings of the ACM SIGMOD international conference on Management of data*, pages 13–24. ACM Press, 1998.

[142] S.-H. Nienhuys-Cheng and R. de Wolf. *Foundations of Inductive Logic Programming*, volume 1228 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, 1997.

[143] S. Nijssen. Homepage for mining structured data, `http://hms.liacs.nl/`. 2003.

[144] S. Nijssen and J. N. Kok. Faster association rules for multiple relations. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 891–896. Morgan Kaufmann Publishers, 2001.

[145] S. Nijssen and J. N. Kok. Efficient discovery of frequent unordered trees. In *Proceedings of the First International Workshop on Mining Graphs, Trees and Sequences (MGTS)*, pages 55–64, 2003.

[146] S. Nijssen and J. N. Kok. Efficient frequent query discovery in FARMER. In *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, volume 2838 of *Lecture Notes in Computer Science*, pages 350–362. Springer-Verlag, 2003.

[147] S. Nijssen and J. N. Kok. Proper refinement of Datalog clauses using primary keys. In *Proceedings of the 15th Belgium-Netherlands Conference on Artificial Intelligence*, pages 227–234. Belgium-Netherlands Association for Artificial Intelligence, 2003.

[148] S. Nijssen and J. N. Kok. Frequent graph mining and its application to molecular databases. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE Press, 2004.

[149] S. Nijssen and J. N. Kok. The GASTON tool for frequent subgraph mining. In *Proceedings of the International Workshop on Graph-Based Tools, (Grabats)*. Elsevier Science, 2004.

[150] S. Nijssen and J. N. Kok. Ideal refinement of Datalog clauses using primary keys. In *Proceedings of the 16th Eureopean Conference on Artificial Intelligence (ECAI)*, pages 520–524. IOS Press, 2004.

[151] S. Nijssen and J. N. Kok. A quickstart in frequent structure mining can make a difference. In *Proceedings of the 2004 International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 647–652. ACM Press, 2004.

[152] S. Nijssen and J. N. Kok. On multi-class correlated pattern mining. In *Proceedings of the Fourth International Workshop on Knowledge Discovery in Inductive Databases (KDID)*, 2005.

[153] C. A. Orengo, D. T. Jones, and J. M. Thornton. *Bioinformatics: Genes, Proteins and computers*. BIOS Scientific Publishers, 2003.

[154] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In *Proceedings of the 7th International Conference on Database Theory (ICDT)*, volume 1540 of *Lecture Notes in Computer Science*, pages 398–416. Springer-Verlag, 1999.

[155] J. Pei and J. Han. Can we push more constraints into frequent pattern mining? In *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 350–354. ACM Press, 2000.

[156] J. Pei, J. Han, and L. V. S. Lakshmanan. Mining frequent item sets with convertible constraints. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, pages 433–442. IEEE Press, 2001.

[157] J. Pei, J. Han, and R. Mao. CLOSET: An efficient algorithm for mining frequent closed itemsets. In *Proceedings of the 5th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 21–30, 2000.

[158] J. Pei, J. Han, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*. IEEE Press.

[159] W. Pijls and J. C. Bioch. Mining frequent itemsets in memory-resident databases. In *Proceedings of the 4th World Multiconference on Systemics, Cybernetics and Informatics (SCI)*, pages 93–98. IIIS Publishers, 2000.

[160] G. Plotkin. *Automatic Methods of Inductive Inference (Ph.D. Dissertation)*. Edinburgh University, 1971.

[161] F. Provost and T. Fawcett. Robust classification for imprecise environments. In *Machine Learning*, volume 42, pages 203–231. Kluwer Academic Publishers, 2001.

[162] J. Punin and M. Krishnamoorthy. WWWPal system–a system for analysis and synthesis of web pages. In *WebNet 98 Conference*, November 1998.

[163] J. Punin, M. Krishnamoorthy, and M. J. Zaki. LOGML —log markup language for web usage mining. In *WEBKDD 2001 — Mining Web Log Data Across All Customers Touch Points, Third International Workshop*, volume 2356 of *Lecture Notes in Artificial Intelligence*, pages 88–112. Springer-Verlag, 2002.

[164] S. Reyner. An analysis of a good algorithm for the subtree problem. In *SIAM Journal of Computing*, volume 6, pages 906–908. SIAM, 1977.

[165] N. Robertson and P. D. Seymour. Graph minors. II. Algorithmic aspects of tree-width. In *Journal of Algorithms*, volume 7, pages 309–322. Elsevier Science, 1986.

[166] U. Rückert and S. Kramer. Frequent free tree discovery in graph data. In *Proceedings of the 2004 ACM symposium on Applied computing (SAC)*, pages 564–570. ACM Press, 2004.

[167] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice-Hall, 2003.

[168] D. C. Schmid and L. E. Druffel. A fast backtracking algorithm to test directed graphs for isomorphism using distance matrices. In *Journal of the Association for Computing Machinery*, volume 23, pages 433–445. ACM Press, 1976.

[169] H. Scoins. Placing trees in lexicographic order. In *Machine Intelligence*, volume 3, pages 43–60. Edinburgh University Press, 1968.

[170] D. Y. Seid and S. Mehrotra. Efficient relationship pattern mining using multi-relational iceberg-cubes. In *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM)*, pages 515–518. IEEE Press, 2004.

[171] S. Sekine. *Corpus-Based Parsing and Sublanguages Studies (Ph.D. Dissertation)*. New York University, USA, 1998.

[172] J. Setubal. Sequential and parallel experimental results with bipartite matching algorithms. Technical Report IC-96-09, Institute of Computing, State University of Campinas, Brazil, 1996.

[173] R. Shamir and D. Tsur. Faster subtree isomorphism. In *Journal of Algorithms*, volume 33, pages 267–280. Elsevier Science, 1999.

[174] P. Shenoy, J. R. Haritsa, S. Sundarshan, G. Bhalotia, M. Bawa, and D. Shah. Turbocharging vertical mining of large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 22–33. ACM Press, 2000.

[175] A. Shioura, A. Tamura, and T. Uno. An optimal algorithm for scanning all spanning trees of undirected graphs. In *SIAM Journal on Comuting*, volume 26, pages 678–692. SIAM, 1997.

[176] R. Srikant and R. Agrawal. Mining generalized association rules. In *Proceedings of the 21st International Conference on Very Large Databases (VLDB)*, pages 407–419. Morgan Kaufmann Publishers, 1995.

[177] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *Proceedings of the 5th International Conference on Extending Database Technology (EDBT)*, volume 1057 of *Lecture Notes in Computer Science*, pages 3–17. Springer-Verlag, 1996.

[178] R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 67–73. AAAI Press, 1997.

[179] A. Termier, M.-C. Rousset, and M. Sebag. TreeFinder: A first step towards XML data mining. In *Proceedings of the Second IEEE International Conference on Data Mining (ICDM)*, pages 450–457. IEEE Press, 2002.

[180] A. Termier, M.-C. Rousset, and M. Sebag. Dryade: A new approach for discovering closed frequent trees in heterogeneous tree databases. In *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM)*, pages 543–546. IEEE Press, 2004.

[181] L. Todorovski, P. A. Flach, and N. Lavrač. Predictive performance of weighted relative accuracy. In *Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, volume 1910 of *Lecture Notes in Computer Science*, pages 255–264. Springer-Verlag, 2000.

[182] E. Ukkonen. On-line construction of suffix trees. In *Algorithmica*, volume 14, pages 249–260. Springer-Verlag, 1995.

[183] J. R. Ullman. An algorithm for subgraph isomorphism. In *Journal of the Association for Computing Machinery*, volume 23, pages 31–42. ACM Press, 1976.

[184] R. Verma and S. Reyner. An analysis of a good algorithm for the subtree problem, corrected. In *SIAM Journal of Computing*, volume 18, pages 906–908. SIAM, 1989.

[185] J. Vilo. Discovering frequent patterns from strings. Technical Report C-1998-9, Department of Computer Science, University of Helsinki, 1998.

[186] C. Wang, M. Hong, J. Pei, H. Zhou, W. Wang, and B. Shi. Efficient pattern-growth methods for frequent tree pattern mining. In *Proceedinges of the Eigth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, volume 3056 of *Lecture Notes in Computer Science*, pages 441–451. Springer-Verlag, 2004.

[187] C. Wang, W. Wang, J. Pei, Y. Zhu, and B. Shi. Scalable mining of large disk-based graph databases. In *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 316–325. ACM Press, 2004.

[188] J. Wang and S. Zhang. Unordered tree mining with applications to phylogeny. In *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM)*, pages 708–719. IEEE Press, 2004.

[189] K. Wang and H. Liu. Discovering typical structures of documents: A road map approach. In *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 146–154. ACM Press, 1998.

[190] G. I. Webb. Efficient search for association rules. In *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 99–107. ACM Press, 2000.

[191] G. I. Webb and S. Zhang. *k*-optimal-rule-discovery. In *Data Mining and Knowledge Discovery*, volume 10, pages 39–79. Kluwer Acadamic Publishers, 2005.

[192] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. In *Journal of Chemical Information and Computer Sciences*, volume 28, pages 31–36. ACS Publications, 1988.

[193] D. Weininger, A. Weininger, and J. L. Weininger. SMILES 2. Algorithm for generation of unique SMILES notation. In *Journal of Chemical Information and Computer Sciences*, volume 29, pages 97–101. ACS Publications, 1989.

[194] A. Winter. Exchanging graphs with GXL. In *Proceedings of the 9th International Symposium on Graph Drawing (GD)*, volume 2265 of *Lecture Notes in Computer Science*, pages 485–500. Springer-Verlag, 2001.

[195] I. H. Witten and E. Frank. *Data Mining*. Morgan Kauffman Publishers, 2000.

[196] D. Wood. *Data Structures, Algorithms and Performance*. Addison-Wesley, 1993.

[197] R. Wright, B. Richmond, A. Odzlyzko, and B. McKay. Constant time generation of free trees. volume 15, pages 540–548. SIAM, 1986.

[198] Y. Xiao, J.-F. Yao, Z. Li, and M. Dunham. Efficient data mining for maximal frequent subtrees. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM)*, pages 379–386. IEEE Press, 2003.

[199] X. Yan and J. Han. gSpan: Graph-based substructure pattern mining. In *Proceedings of the Second IEEE International Conference on Data Mining (ICDM)*, pages 721–724. IEEE Press, 2002.

[200] X. Yan and J. Han. gSpan: Graph-based substructure pattern mining — Expanded version. Technical Report UIUCDCS-R-2002-2296, University of Illinois at Urbana-Champaign, 2002.

[201] X. Yan and J. Han. CloseGraph: Mining closed frequent graph patterns. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 286–295. ACM Press, 2003.

[202] X. Yan, P. S. Yu, and J. Han. Graph indexing: A frequent structure-based approach. In *Proceedings of the ACM SIGMOD International conference on Management of Data*, pages 335–346. ACM Press, 2004.

[203] M. J. Zaki. Efficiently mining frequent trees in a forest. In *Proceedings of the 8th International Conference Knowledge Discovery and Data Mining (KDD)*, pages 71–80. ACM Press, 2002.

[204] M. J. Zaki. Efficiently mining frequent embedded unordered trees. In *Fundamenta Informaticae*, volume 66, pages 33–52. IOS Press, 2005.

[205] M. J. Zaki and K. Gouda. Fast vertical mining using diffsets. In *Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 326–335. ACM Press, 2003.

[206] M. J. Zaki and C.-J. Hsiao. CHARM: An efficient algorithm for closed itemset mining. In *Proceedings of the Second SIAM International Conference on Data Mining (SDM)*, pages 457–473. SIAM, 2002.

[207] M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 283–286. AAAI Press, 1997.

[208] S. Zhang and J. Wang. Frequent agreement subtree mining, 2005. `http://aria.njit.edu/mediadb/fast/`.

[209] A. Zimmermann and L. De Raedt. Cluster-grouping: From subgroup discovery to clustering. In *Proceedings of the 15th European Conference on Machine Learning (ECML)*, volume 3201 of *Lecture Notes in Computer Science*, pages 575–577. Springer-Verlag, 2004.

[210] A. Zimmermann and L. De Raedt. CorClass: Correlated association rule mining for classification. In *Proceedings of the 7th International Conference on Discovery Science (DS)*, volume 3245 of *Lecture Notes in Artificial Intelligence*, pages 60–72. Springer-Verlag, 2004.

# Index

# Acknowledgements

I would like to thank Eric-Wubbo Lameijer for building the molecular model of Cuneane of which a photo is included in this thesis. I enjoyed the discussions that I had with Eric-Wubbo and Jeroen Kazius about mining molecular databases. These discussions have motivated me very much, and I would like thank them for that. Of course I would also like to thank all colleagues that I used to have lunch and 'coffee' breaks with for making the period in Leiden an enjoyable one.