

Data Mining: Concepts and Techniques (2nd edition)

Jiawei Han and Micheline Kamber
Morgan Kaufmann Publishers, 2006

Bibliographic Notes for Chapter 11 Applications and Trends in Data Mining

Many books discuss applications of data mining. For financial data analysis and financial modeling, see Benninga and Czaczkes [BC00] and Higgins [Hig03]. For retail data mining and customer relationship management, see books by Berry and Linoff [BL04] and Berson, Smith, and Thearling [BST99], and the article by Kohavi [Koh01]. For telecommunication-related data mining, see the book by Mattison [Mat97]. Chen, Hsu, and Dayal [CHD00] reported their work on scalable telecommunication tandem traffic analysis under a data warehouse/OLAP framework. For bioinformatics and biological data analysis, there are a large number of introductory references and textbooks. An introductory overview of bioinformatics for computer scientists was presented by Cohen [Coh04]. Recent textbooks on bioinformatics include Krane and Raymer [KR03], Jones and Pevzner [JP04], Durbin, Eddy, Krogh and Mitchison [DEKM98], Setubal and Meidanis [SM97], Orengo, Jones, and Thornton [OJT⁺03], and Pevzner [Pev03]. Summaries of biological data analysis methods and algorithms can also be found in many other books, such as Gusfield [Gus97], Waterman [Wat95], Baldi and Brunak [BB01], and Baxeavanis and Ouellette [BO04]. There are many books on scientific data analysis, such as Grossman, Kamath, Kegelmeyer, et al. (eds.) [GKK⁺01]. For geographic data mining, see the book edited by Miller and Han [MH01]. Valdes-Perez [VP99] discusses the principles of human-computer collaboration for knowledge discovery in science. For intrusion detection, see Barbará [Bar02] and Northcutt and Novak [NN02].

Many data mining books contain introductions to various kinds of data mining systems and products. KDnuggets maintains an up-to-date list of data mining products at www.kdnuggets.com/companies/products.html and the related software at www.kdnuggets.com/software/index.html, respectively. For a survey of data mining and knowledge discovery software tools, see Goebel and Gruenwald [GG99]. Detailed information regarding specific data mining systems and products can be found by consulting the Web pages of the companies offering these products, the user manuals for the products in question, or magazines and journals on data mining and data warehousing. For example, the Web page URLs for the data mining systems introduced in this chapter are www-4.ibm.com/software/data/iminer for IBM Intelligent Miner, www.microsoft.com/sql for Microsoft SQL Server, www.purpleinsight.com/products for MineSet of Purple Insight, www.oracle.com/ for Oracle Data Mining (ODM), www.spss.com/clementine for Clementine of SPSS, www.sas.com/technologies/analytics/datamining/miner for SAS Enterprise Miner, www.insightful.com/products/iminer for Insightful Miner of Insightful Inc, and www.R-project.org for the R environment for statistical computing and graphics. CART and See5/C5.0 are available from www.salford-systems.com and www.rulequest.com, respectively. Weka is available from the University of Waikato at www.cs.waikato.ac.nz/ml/weka. Since data mining systems and their functions evolve rapidly, it is not our intention to provide any kind of comprehensive survey on data mining systems in this book. We apologize if your data mining systems or tools were not included.

Issues on the theoretical foundations of data mining are addressed in many research papers. Mannila presented a summary of studies on the foundations of data mining in [Man00]. The data reduction view of data mining was summarized in *The New Jersey Data Reduction Report* by Barbará, DuMouchel, Faloutos, et al. [BDF⁺97]. The data compression view can be found in studies on the minimum description length (MDL) principle, such as Quinlan and Rivest [QR89] and Chakrabarti, Sarawagi, and Dom [CSD98]. The pattern discovery point of view of data mining is addressed in numerous machine learning and data mining studies, ranging from association mining, decision tree induction, and neural network classification to sequential pattern mining, clustering, and so on. The probability theory point of view can be seen in the statistics literature, such as in studies on Bayesian networks and hierarchical Bayesian models, as addressed in Chapter 6. Kleinberg, Papadimitriou, and Raghavan [KPR98] presented a microeconomic view, treating data mining as an optimization problem. The view of data mining as

the querying of inductive databases was proposed by Imielinski and Mannila [IM96].

Statistical techniques for data analysis are described in several books, including *Intelligent Data Analysis* (2nd ed.), edited by Berthold and Hand [BH03]; *Probability and Statistics for Engineering and the Sciences* (6th ed.) by Devore [Dev03]; *Applied Linear Statistical Models with Student CD* by Kutner, Nachtsheim, Neter, and Li [KNNL04]; *An Introduction to Generalized Linear Models* (2nd ed.) by Dobson [Dob01]; *Classification and Regression Trees* by Breiman, Friedman, Olshen, and Stone [BFOS84]; *Mixed Effects Models in S and S-PLUS* by Pinheiro and Bates [PB00]; *Applied Multivariate Statistical Analysis* (5th ed.) by Johnson and Wichern [JW02]; *Applied Discriminant Analysis* by Huberty [Hub94]; *Time Series Analysis and Its Applications* by Shumway and Stoffer [SS05]; and *Survival Analysis* by Miller [Mil98].

For visual data mining, popular books on the visual display of data and information include those by Tufte [Tuf90, Tuf97, Tuf01]. A summary of techniques for visualizing data is presented in Cleveland [Cle93]. For information about StatSoft, a statistical analysis system that allows data visualization, see www.statsoft.inc. A VisDB system for database exploration using multidimensional visualization methods was developed by Keim and Kriegel [KK94]. Ankerst, Elsen, Ester, and Kriegel [AEEK99] present a perception-based classification approach, PBC, for interactive visual classification. The book, *Information Visualization in Data Mining and Knowledge Discovery*, edited by Fayyad, Grinstein, and Wierse [FGW01], contains a collection of articles on visual data mining methods.

There are many research papers on collaborative recommender systems. These include the GroupLens architecture for collaborative filtering by Resnick, Iacovou, Suchak, et al. [RIS⁺94]; empirical analysis of predictive algorithms for collaborative filtering by Breese, Heckerman, and Kadie [BHK98]; its applications in information tapestry by Goldberg, Nichols, Oki and Terry [GNOT92]; a method for learning collaborative information filters by Billsus and Pazzani [BP98]; an algorithmic framework for performing collaborative filtering proposed by Herlocker, Konstan, Borchers and Riedl [HKBR98]; item-based collaborative filtering recommendation algorithms by Sarwar, Karypis, Konstan, and Riedl [SKKR01] and Lin, Alvarez, and Ruiz [LAR02]; and content-boosted collaborative filtering for improved recommendations by Melville, Mooney and Nagarajan [MMN02].

Many examples of ubiquitous and invisible data mining can be found in an insightful and entertaining article by John [Joh99], and a survey of Web mining by Srivastava, Desikan, and Kumar [SDK04]. The use of data mining at Wal-Mart was depicted in Hays [Hay04]. Bob, the automated fast food management system of HyperActive Technologies, is described at www.hyperactivetechnologies.com. The book *Business @ the Speed of Thought: Succeeding in the Digital Economy* by Gates [Gat00] discusses e-commerce and customer relationship management, and provides an interesting perspective on data mining in the future. For an account on the use of Clementine by police to control crime, see Beal [Bea04]. Mena [Men03] has an informative book on the use of data mining to detect and prevent crime. It covers many forms of criminal activities, ranging from fraud detection, money laundering, insurance crimes, identity crimes, and intrusion detection.

Data mining issues regarding privacy and data security are substantially addressed in literature. One of the first papers on data mining and privacy was by Clifton and Marks [CM96]. The Fair Information Practices discussed in Section 11.4.2 were presented by the Organization for Economic Co-operation and Development (OECD) [OEC98]. Laudon [Lau96] proposes a regulated national information market that would allow personal information to be bought and sold. Cavoukian [Cav98] considered opt-out choices and data security-enhancing techniques. Data security-enhancing techniques and other issues relating to privacy were discussed in Walstrom and Roddick [WR01]. Data mining for counterterrorism and its implications for privacy were discussed in Thuraisingham [Thu04]. A survey on privacy-preserving data mining can be found in Verykios, Bertino, Fovino, and Provenza [VBFP04]. Many algorithms have been proposed, including work by Agrawal and Srikant [AS00], Evfimievski, Srikant, Agrawal and Gehrke [ESAG02], and Vaidya and Clifton [VC03]. Agrawal and Aggarwal [AA01] proposed a metric for assessing privacy preservation, based on differential entropy. Clifton, Kantarcioğlu, and Vaidya [CKV04] discussed the need to produce a rigorous definition of privacy and a formalism to prove privacy-preservation in data mining.

Data mining standards and languages have been discussed in several forums. The new book *Data Mining with SQL Server 2005* by Tang and MacLennan [TM05] describes Microsoft's OLE DB for Data Mining. Other efforts towards standardized data mining languages include Predictive Model Markup Language (or PMML), described at www.dmg.org, and Cross-Industry Standard Process for Data Mining (or CRISP-DM), described at www.crisp-

dm.org.

There have been lots of discussions on trend and research directions in data mining in various forums and occasions. A recent book that collects a set of articles on trends and challenges of data mining was edited by Kargupta, Joshi, Sivakumar, and Yesha [KJSY04]. For a tutorial on distributed data mining, see Kargupta and Sivakumar [KS04]. For multirelational data mining, see the introduction by Dzeroski [Dze03], as well as work by Yin, Han, Yang, and Yu [YHYY04]. For mobile data mining, see Kargupta, Bhargava, Liu, et al. [KBL⁺04]. Washio and Motoda [WM03] presented a survey on graph-based mining, that also covers several typical pieces of work, including Su, Cook, and Holder [SCH99], Kuramochi and Karypis [KK01], and Yan and Han [YH02]. ACM SIGKDD Explorations had special issues on several of the topics we have addressed, including DNA microarray data mining (volume 5, number 2, December 2003); constraints in data mining (volume 4, number 1, June 2002); multirelational data mining (volume 5, number 1, July 2003); and privacy and security (volume 4, number 2, December 2002).

Bibliography

- [AA01] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proc. 2001 ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems (PODS'01)*, pages 247–255, Santa Barbara, CA, May 2001.
- [AEEK99] M. Ankerst, C. Elsen, M. Ester, and H.-P. Kriegel. Visual classification: An interactive approach to decision tree construction. In *Proc. 1999 Int. Conf. Knowledge Discovery and Data Mining (KDD'99)*, pages 392–396, San Diego, CA, Aug. 1999.
- [AS00] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00)*, pages 439–450, Dallas, TX, May 2000.
- [Bar02] D. Barbará. *Applications of Data Mining in Computer Security (Advances in Information Security, 6)*. Kluwer Academic, 2002.
- [BB01] P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach* (2nd ed.). MIT Press, 2001.
- [BC00] S. Benninga and B. Czaczkes. *Financial Modeling* (2nd ed.). MIT Press, 2000.
- [BDF⁺97] D. Barbará, W. DuMouchel, C. Faloutsos, P. J. Haas, J. H. Hellerstein, Y. Ioannidis, H. V. Jagadish, T. Johnson, R. Ng, V. Poosala, K. A. Ross, and K. C. Servcik. The New Jersey data reduction report. *Bull. Technical Committee on Data Engineering*, 20:3–45, Dec. 1997.
- [Bea04] B. Beal. Case study: Analytics take a bite out of crime. In *searchCRM.com (www.searchCRM.techtarget.com)*, Jan. 2004.
- [BFOS84] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth International Group, 1984.
- [BH03] M. Berthold and D. J. Hand. *Intelligent Data Analysis: An Introduction* (2nd ed.). Springer-Verlag, 2003.
- [BHK98] J. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proc. 1998 Conf. Uncertainty in Artificial Intelligence*, pages 43–52, Madison, WI, July 1998.
- [BL04] M. J. A. Berry and G. S. Linoff. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. John Wiley & Sons, 2004.
- [BO04] A. Baxevanis and B. F. F. Ouellette. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* (3rd ed.). John Wiley & Sons, 2004.
- [BP98] D. Billsus and M. J. Pazzani. Learning collaborative information filters. In *Proc. 1998 Int. Conf. Machine Learning (ICML'98)*, pages 46–54, Madison, WI, Aug. 1998.
- [BST99] A. Berson, S. J. Smith, and K. Thearling. *Building Data Mining Applications for CRM*. McGraw-Hill, 1999.

- [Cav98] A. Cavoukian. Data mining: Staking a claim on your privacy. In *Office of the Information and Privacy Commissioner, Ontario (www.ipc.on.ca/docs/datamine.pdv, viewed Mar. 2005)*, Jan. 1998.
- [CHD00] Q. Chen, M. Hsu, and U. Dayal. A data-warehouse/OLAP framework for scalable telecommunication tandem traffic analysis. In *Proc. 2000 Int. Conf. Data Engineering (ICDE'00)*, pages 201–210, San Diego, CA, Feb. 2000.
- [CKV04] C. Clifton, M. Kantarcioğlu, and J. Vaidya. Defining privacy for data mining. In H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha, editors, *Data Mining: Next Generation Challenges and Future Directions*, pages 255–270. AAAI/MIT Press, 2004.
- [Cle93] W. Cleveland. *Visualizing Data*. Hobart Press, 1993.
- [CM96] C. Clifton and D. Marks. Security and privacy implications of data mining. In *Proc. 1996 SIGMOD'96 Workshop Research Issues on Data Mining and Knowledge Discovery (DMKD'96)*, pages 15–20, Montreal, Canada, June 1996.
- [Coh04] J. Cohen. Bioinformatics—an introduction for computer scientists. *ACM Computing Surveys*, 36:122–158, 2004.
- [CSD98] S. Chakrabarti, S. Sarawagi, and B. Dom. Mining surprising patterns using temporal description length. In *Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98)*, pages 606–617, New York, NY, Aug. 1998.
- [DEKM98] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probability Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.
- [Dev03] J. L. Devore. *Probability and Statistics for Engineering and the Sciences* (6th ed.). Duxbury Press, 2003.
- [Dob01] A. J. Dobson. *An Introduction to Generalized Linear Models* (2nd ed.). Chapman and Hall, 2001.
- [Dze03] S. Dzeroski. Multirelational data mining: An introduction. *ACM SIGKDD Explorations*, 5:1–16, July 2003.
- [ESAG02] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *Proc. 2002 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'02)*, pages 217–228, Edmonton, Canada, July 2002.
- [FGW01] U. Fayyad, G. Grinstein, and A. Wierse. *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann, 2001.
- [Gat00] B. Gates. *Business @ the Speed of Thought: Succeeding in the Digital Economy*. Warner Books, 2000.
- [GG99] M. Goebel and L. Gruenwald. A survey of data mining and knowledge discovery software tools. *SIGKDD Explorations*, 1:20–33, 1999.
- [GKK⁺01] R. L. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, and R. R. Namburu. *Data Mining for Scientific and Engineering Applications*. Kluwer Academic, 2001.
- [GNOT92] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Comm. ACM*, 35:61–70, 1992.
- [Gus97] D. Gusfield. *Algorithms on Strings, Trees and Sequences, Computer Science and Computation Biology*. Cambridge University Press, 1997.
- [Hay04] C. L. Hays. What Wal-Mart knows about customers' habits. In *New York Times*, Sec. 3, page 1, col. 1, Nov. 14, 2004.
- [Hig03] R. C. Higgins. *Analysis for Financial Management* (7th ed.). Irwin/McGraw-Hill, 2003.

- [HKBR98] J. L. Herlocker, J. A. Konstan, J. R. A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Proc. 1999 Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'99)*, pages 230–237, Berkeley, CA, Aug. 1998.
- [Hub94] C. H. Huberty. *Applied Discriminant Analysis*. New York, 1994.
- [IM96] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. *Comm. ACM*, 39:58–64, 1996.
- [Joh99] G. H. John. Behind-the-scenes data mining: A report on the KDD-98 panel. *SIGKDD Explorations*, 1:6–8, 1999.
- [JP04] N. C. Jones and P. A. Pevzner. *An Introduction to Bioinformatics Algorithms*. MIT Press, 2004.
- [JW02] R. A. Johnson and D. A. Wichern. *Applied Multivariate Statistical Analysis* (5th ed.). Prentice Hall, 2002.
- [KBL⁺04] H. Kargupta, B. Bhargava, K. Liu, M. Powers, P. Blair, S. Bushra, J. Dull, K. Sarkar, M. Klein, M. Vasa, and D. Handy. VEDAS: A mobile and distributed data stream mining system for real-time vehicle monitoring. In *Proc. 2004 SIAM Int. Conf. Data Mining (SDM'04)*, Lake Buena Vista, FL, April 2004.
- [KJSY04] H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha. *Data Mining: Next Generation Challenges and Future Directions*. AAAI/MIT Press, 2004.
- [KK94] D. A. Keim and H.-P. Kriegel. VisDB: Database exploration using multidimensional visualization. In *Computer Graphics and Applications*, pages 40–49, 1994.
- [KK01] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *Proc. 2001 Int. Conf. Data Mining (ICDM'01)*, pages 313–320, San Jose, CA, Nov. 2001.
- [KNNL04] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models with Student CD*. Irwin, 2004.
- [Koh01] R. Kohavi. Mining e-commerce data: The good, the bad, and the ugly. In *Proc. 2001 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'01)*, pages 8–13, San Fransisco, CA, Aug. 2001.
- [KPR98] J. M. Kleinberg, C. Papadimitriou, and P. Raghavan. A microeconomic view of data mining. *Data Mining and Knowledge Discovery*, 2:311–324, 1998.
- [KR03] D. Krane and R. Raymer. *Fundamental Concepts of Bioinformatics*. Benjamin Cummings, 2003.
- [KS04] H. Kargupta and K. Sivakumar. Existential pleasures of distributed data mining. In H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha, editors, *Data Mining: Next Generation Challenges and Future Directions*, pages 3–25. AAAI/MIT Press, 2004.
- [LAR02] W. Lin, S. Alvarez, and C. Ruiz. Efficient adaptive-support association rule mining for recommender systems. *Data Mining and Knowledge Discovery*, 6:83–105, 2002.
- [Lau96] K. C. Laudon. Markets and privacy. *Comm. ACM*, 39:92–104, Sept. 96.
- [Man00] H. Mannila. Theoretical frameworks of data mining. *SIGKDD Explorations*, 1:30–32, 2000.
- [Mat97] R. Mattison. *Data Warehousing and Data Mining for Telecommunications*. Artech House, 1997.
- [Men03] J. Mena. *Investigative Data Mining with Security and Criminal Detection*. Butterworth-Heinemann, 2003.
- [MH01] H. Miller and J. Han. *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, 2001.
- [Mil98] R. G. Miller. *Survival Analysis*. Wiley-Interscience, 1998.

- [MMN02] P. Melville, R. J. Mooney, and R. Nagarajan. Content-boosted collaborative filtering for improved recommendations. In *Proc. 2002 Nat. Conf. Artificial Intelligence (AAAI'02)*, pages 187–192, Edmonton, Canada, July 2002.
- [NN02] S. Northcutt and J. Novak. *Network Intrusion Detection*. Sams, 2002.
- [OEC98] OECD. *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*. Organization for Economic Co-operation and Development, 1998.
- [OJT⁺03] C. A. Orengo, D. T. Jones, J. M. Thornton, Proteins Bioinformatics: Genes, and Computers. BIOS Scientific Pub., 2003.
- [PB00] J. C. Pinheiro and D. M. Bates. *Mixed Effects Models in S and S-PLUS*. Springer-Verlag, 2000.
- [Pev03] J. Pevzner. *Bioinformatics and Functional Genomics*. Wiley-Liss, 2003.
- [QR89] J. R. Quinlan and R. L. Rivest. Inferring decision trees using the minimum description length principle. *Information and Computation*, 80:227–248, Mar. 1989.
- [RIS⁺94] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proc. 1994 Conf. Computer Supported Cooperative Work (CSCW'94)*, pages 175–186, Chapel Hill, NC, Oct. 1994.
- [SCH99] S. Su, D. J. Cook, and L. B. Holder. Knowledge discovery in molecular biology: Identifying structural regularities in proteins. *Intelligent Data Analysis*, 3:413–436, 1999.
- [SDK04] J. Srivastava, P. Desikan, and V. Kumar. Web mining—concepts, applications, and research directions. In H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha, editors, *Data Mining: Next Generation Challenges and Future Directions*, pages 405–423. AAAI/MIT Press, 2004.
- [SKKR01] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proc. 2001 Int. World Wide Web Conf. (WWW'01)*, pages 158–167, Hong Kong, China, May 2001.
- [SM97] J. C. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology*. PWS Pub Co., 1997.
- [SS05] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications*. Springer, 2005.
- [Thu04] B. Thuraisingham. Data mining for counterterrorism. In H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha, editors, *Data Mining: Next Generation Challenges and Future Directions*, pages 157–183. AAAI/MIT Press, 2004.
- [TM05] Z. Tang and J. MacLennan. *Data Mining with SQL Server 2005*. John Wiley & Sons, 2005.
- [Tuf90] E. R. Tufte. *Envisioning Information*. Graphics Press, 1990.
- [Tuf97] E. R. Tufte. *Visual Explanations : Images and Quantities, Evidence and Narrative*. Graphics Press, 1997.
- [Tuf01] E. R. Tufte. *The Visual Display of Quantitative Information* (2nd ed.). Graphics Press, 2001.
- [VBFP04] V. S. Verykios, E. Bertino, I. N. Fovino, and L. P. Provenza. State-of-the-art in privacy preserving data mining. *SIGMOD Record*, 33:50–57, March 2004.
- [VC03] J. Vaidya and C. Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In *Proc. 2003 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'03)*, Washington, DC, Aug 2003.
- [VP99] P. Valdes-Perez. Principles of human-computer collaboration for knowledge-discovery in science. *Artificial Intelligence*, 107:335–346, 1999.

- [Wat95] M. S. Waterman. *Introduction to Computational Biology: Maps, Sequences, and Genomes (Interdisciplinary Statistics)*. CRC Press, 1995.
- [WM03] T. Washio and H. Motoda. State of the art of graph-based data mining. *SIGKDD Explorations*, 5:59–68, 2003.
- [WR01] K. Wahlstrom and J. F. Roddick. On the impact of knowledge discovery and data mining. In *Selected Papers from the 2nd Australian Institute of Computer Ethics Conference (AICE2000)*, pages 22–27, Canberra, Australia, 2001.
- [YH02] X. Yan and J. Han. gSpan: Graph-based substructure pattern mining. In *Proc. 2002 Int. Conf. Data Mining (ICDM'02)*, pages 721–724, Maebashi, Japan, Dec. 2002.
- [YHYY04] X. Yin, J. Han, J. Yang, and P. S. Yu. CrossMine: Efficient classification across multiple database relations. In *Proc. 2004 Int. Conf. Data Engineering (ICDE'04)*, pages 399–410, Boston, MA, Mar. 2004.