Chapter 13

Trends and Research Frontiers in Data Mining

13.1 Bibliographic Notes

For mining complex types of data, there are many research papers and books covering various themes. We will leave the detailed discussions of the research history and literature in Volume II and list here some recent books and well-cited survey or research articles for references.

Time-series analysis has been studied in statistics and computer science communities for decades, with many textbooks, such as Box, Jenkins and Reinsel [BJR08], Brockwell and Davis [BD02], Chatfield [Cha03b], Hamilton [Ham94], and Shumway and Stoffer [SS05]. A fast subsequence matching method in time-series databases was presented by Faloutsos, Ranganathan, and Manolopoulos [FRM94]. Agrawal, Lin, Sawhney, and Shim [ALSS95] developed a method for fast **similarity search** in the presence of noise, scaling, and translation in time-series databases. Shasha and Zhu present an overview of the methods for high performance discovery in time series [SZ04].

Sequential pattern mining methods have been studied by many researchers, such as Agrawal and Srikant [SA96], Zaki [Zak01], Pei, Han, Mortazavi-Asl, et al. [PHMA+04], Yan, Han and Afshar [YHA03]. The study on **sequence classification** include Ji, Bailey and Dong [JBD05], and Ye and Keogh [YK09], with a survey by Xing, Pei and Keogh [XPK10]. Dong and Pei [DP07] provides an overview on **sequence data mining** methods.

Methods for **analysis of biological sequences** including **Markov chains** and **hidden Markov models** are introduced in many books or tutorials, such as Waterman [Wat95], Setubal and Meidanis [SM97], Durbin, Eddy, Krogh and Mitchison [DEKM98], Baldi and Brunak [BB01], Krane and Raymer [KR03], Rabiner [Rab89], Jones and Pevzner [JP04], and Baxevanis and Ouellette [BO04]. Information about BLAST (see also, Korf, Yandell, and Bedell [KYB03]) can be found at NCBI Web site http://www.ncbi.nlm.nih.gov/BLAST/.

Graph pattern mining has been studied extensively, including Holder, Cook and Djoko [HCD94], Inokuchi, Washio, and Motoda [IWM98], Kuramochi and Karypis [KK01], Yan and Han [YH02, YH03], Borgelt and Berthold [BB02], Huan, Wang, Bandyopadhyay, et al. [HWB+04], and Gaston by Nijssen and Kok [NK04].

There has been a great deal of research on **social and information network analysis**, including Newman [New10], Easley and Kleinberg [EK10], Yu, Han and Faloutsos [YHF10], Wasserman and Faust [WF94], Watts [Wat03], Newman, Barabasi, and Watts [NBW06]. **Statistical modeling of networks** is studied popularly, such as Albert and Barbasi [AB99], Watts [Wat03], Faloutsos, Faloutsos, and Faloutsos [FFF99], Kumar, Raghavan, Rajagopalan, et al. [KRR+00], and Leskovec, Kleinberg, and Faloutsos [LKF05]. **Data cleaning, integration and validation by information network analysis** was studied by many, such as Bhattacharya and Getoor [BG04], and Yin, Han and Yu [YHY07, YHY08]. **Clustering, ranking and classification in networks** was studied extensively, such as Brin and Page [BP98], Chakrabarti, Dom, and Indyk [CDI98], Kleinberg [Kle99a], Getoor, Friedman, Koller, and Taskar [GFKT01], Newman and M. Girvan [NG04], Yin, Han, Yang, and Yu [YHYY04], Yin, Han, and Yu [YHY05], Xu, Yuruk, Feng and Schweiger [XYFS07], Kulis, Basu, Dhillon and Mooney [KBDM09], Sun, Han, Zhao, et al. [SHZ+09], Neville, Gallaher, and Eliassi-Rad [NGER09], and Ji, Sun, Danilevsky et al. [JSD+10]. **Role discovery and link prediction in information networks** have been studied extensively as well, such as Krebs [Kre02], Kubica, Moore, and Schneider [KMS03], Liben-Nowell and Kleinberg [LNK03], and Wang, Han, Jia, et al. [WHJ+10]. **Similarity search and OLAP in information networks** has been studied by many, such as Tian, Hankins and Patel [THP08], and Chen, Yan, Zhu, et al. [CYZ+08]. **Evolution of social and information networks** has been studied by many researchers, such as Chakrabarti, Kumar, and Tomkins [CKT06], Chi, Song, Zhou, et al. [CSZ+07], Tang, Liu, Zhang, and Nazeri [TLZN08], Xu, Zhang, Yu, and Long [XZYL08], Kim and Han [KH09], and Sun, Tang and Han [STH+10].

Spatial and spatiotemporal data mining has been studied extensively, with a collection of papers by Miller and Han [MH09], and introduced in some textbooks, such as Shekhar and Chawla [SC03], and Hsu, Lee and Wang [HLW07]. Spatial clustering algorithms have been studied extensive in Chapters 10 and 11. Research has been conducted on spatial warehouse and OLAP, such as Stefanovic, Han, and Koperski [SHK00], and spatial and spatiotemporal data mining, such as Koperski and Han [KH95], Mamoulis, Cao, Kollios, Hadjieleftheriou, et al. [MCK+04], Tsoukatos and Gunopulos [TG01], and Hadjieleftheriou, Kollios, Gunopulos, and Tsotras [HKGT03]. **Mining moving object data** has been studied by many, such as Vlachos, Gunopulos, and Kollios [VGK02], Tao, Faloutsos, Papadias, and Liu [TFPL04], Li, Han, Kim and Gonzalez [LHKG07], Lee, Han and Whang [LHW07], and Li, Ding, Han, et al. [LDH+10]. For the bibliography of temporal, spatial, and spatiotemporal data mining research, see a collection by Roddick, Hornsby, and Spiliopoulou [RHS01].

Multimedia data mining has deep roots in image processing and pattern

recognition, which has been studied extensively there, with many textbooks, such as Gonzalez and Woods [GW07], Russ [Rus06], Duda, Hart, and Stork [DHS01], and Z. Zhang and R. Zhang [ZZ09]. Searching and mining of multimedia data has been studied by many (see, *e.g.*, Fayyad and Smyth [FS93], Faloutsos and Lin [FL95], Natsev, Rastogi, and Shim [NRS99], Zaïane, Han, and Zhu [ZHZ00]). An overview of image mining methods is done by Hsu, Lee, and Zhang [HLZ02].

Text data analysis has been studied extensively in information retrieval, with many textbooks and survey articles, such as Croft, Metzler, and Strohman [CMS09], S. Buttcher, C. Clarke, G. Cormack [BCC10], Manning, Raghavan and Schutze [MRS08], Grossman and Frieder [GF04], Baeza-Yates and Riberio-Neto [BYRN11], Zhai [Zha08], Feldman and Sanger [FS06], Berry [Ber03] and Weiss, Indurkhya, Zhang, and Damerau [WIZD04]. Text mining is a fast developing field with numerous papers published in recent years, covering many topics such as topic models *see, *e.g.*, Blei and Lafferty [BL09]), sentiment analysis (see, *e.g.*, Pang and Lee [PL07], and contextual text mining (see, *e.g.*, Mei and Zhai [MZ06]).

Web mining is another focused theme, with books like Chakrabarti [Cha03a], Liu [Liu06], and Berry [Ber03]. Web mining has substantially improved web search engines with a few influential milestone works, such as Brin and Page [BP98], Kleinberg [Kle99b], Chakrabarti, Dom, Kumar, et al. [CDK+99], Kleinberg and Tomkins [KT99]. Numerous results have been generated since then, such as search log mining (see, *e.g.*, Silvestri [Sil10]), blog mining (see, *e.g.*, Mei, Liu, Su, and Zhai [MLSZ06]), and mining online forums (see, *e.g.*, Cong, Wang, Lin et al. [CWL+08]).

Books and surveys on stream data systems and stream data processing include Babu and Widom [BW01], Babcock, Babu, Datar, et al. [BBD+02], Muthukrishnan [Mut05], Aggarwal [Agg06]. **Stream data mining** research covers stream cube model, *e.g.*, Chen, Dong, Han, et al. [CDH+02], stream frequent pattern mining, *e.g.*, Manku and Motwani [MM02], and Karp, Papadimitriou and Shenker [KPS03], stream classification, *e.g.*, Domingos and Hulten [DH00], Wang, Fan, Yu and Han [WFYH03], Aggarwal, Han, Wang and Yu [AHWY04], and stream clustering, *e.g.*, Guha, Mishra, Motwani, and O'Callaghan [GMMO00], Aggarwal, Han, Wang, and Yu [AHWY03].

There are many books that discuss **applications of data mining**. For financial data analysis and financial modeling, see *e.g.*, Benninga [Ben08] and Higgins [Hig08]. For retail data mining and customer relationship management, see *e.g.*, books by Berry and Linoff [BL04] and Berson, Smith, and Thearling [BST99]. For telecommunication-related data mining, see *e.g.*, Horak [Hor08]. There are also books on scientific data analysis, such as Grossman, Kamath, Kegelmeyer, et al. [GKK+01] and Kamath [Kam09].

Issues on the **theoretical foundations of data mining** have been addressed by many researcher. For example, Mannila presents a summary of studies on the foundations of data mining in [Man00]. The data reduction view of data mining is 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 Grunwald and Rissanen [GR07]. The pattern discovery point of view of data mining is addressed in numerous machine learning and data mining studies, ranging from association mining, to decision tree induction, sequential pattern mining, clustering, and so on. The probability theory point of view is popular in the statistics and machine learning literature, such as Bayesian networks and hierarchical Bayesian models in Chapter 9, and probabilistic graph models, *e.g.*, Koller and Friedman [KF09]. Kleinberg, Papadimitriou, and Raghavan [KPR98] present a microeconomic view, treating data mining as an optimization problem. The study on inductive database view include Imielinski and Mannila [IM96] and De Raedt, Guns, and Nijssen [RGN10].

Statistical methods for data analysis are described in many books, such as Hastie, Tibshirani, Friedman [HTF09], Freedman, Pisani and Purves [FPP07], Devore [Dev03], Kutner, Nachtsheim, Neter, and Li [KNNL04], Dobson [Dob01], Breiman, Friedman, Olshen, and Stone [BFOS84], Pinheiro and Bates [PB00], Johnson and Wichern [JW02], Huberty [Hub94], Shumway and Stoffer [SS05], and 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]. A dedicated visual data mining book, *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*, is by Soukup and Davidson [SD02]. 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.

Ubiquitous and invisible data mining have been discuss in many occasions, such as John [Joh99], and some articles in a book edited by Kargupta, Joshi, Sivakumar, and Yesha [KJSY04]. 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. 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 addressed popularly in literature. Books on privacy and security in data mining include Thuraisingham [Thu04], Aggarwal and Yu [AY08], Vaidya, Clifton and Zhu [VCZ10], and Fung, Wang, Fu and Yu [FWFY10]. Research articles include Agrawal and Srikant [AS00], Evfimievski, Srikant, Agrawal and Gehrke [ESAG02], Vaidya and Clifton [VC03]. Differential privacy was introduced by Dwork [Dwo06] and studied by many, such as Hay, Rastogi, Miklau and Suciu [HRMS10].

There have been lots of discussions on **trend and research directions of data mining** in various forums and occasions. Several books are collections of articles on such issues, such as Kargupta, Han, Yu, et al. [KHY+08].

Bibliography

[AB99] R. Albert and A.-L. Barabasi. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[Agg06] C. C. Aggarwal. *Data Streams: Models and Algorithms*. Kluwer Academic, 2006.

[AHWY03] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. In *Proc. 2003 Int. Conf. Very Large Data Bases (VLDB'03)*, pages 81–92, Berlin, Germany, Sept. 2003.

[AHWY04] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. On demand classification of data streams. In *Proc. 2004 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'04)*, pages 503–508, Seattle, WA, Aug. 2004.

[ALSS95] R. Agrawal, K.-I. Lin, H. S. Sawhney, and K. Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In *Proc. 1995 Int. Conf. Very Large Data Bases (VLDB'95)*, pages 490–501, Zurich, Switzerland, Sept. 1995.

[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.

[AY08] C. C. Aggarwal and P. S. Yu. *Privacy-Preserving Data Mining: Models and Algorithms*. Springer, 2008.

[BB01] P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach* (2nd ed.). MIT Press, 2001.

[BB02] C. Borgelt and M. R. Berthold. Mining molecular fragments: Finding relevant substructures of molecules. In *Proc. 2002 Int. Conf. Data Mining (ICDM'02)*, pages 211–218, Maebashi, Japan, Dec. 2002.

[BBD+02] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and issues in data stream systems. In *Proc. 2002 ACM Symp.*

Principles of Database Systems (PODS'02), pages 1–16, Madison, WI, June 2002.

[BCC10] S. Buettcher, C. L. A. Clarke, and G. V. Cormack. *Information Retrieval: Implementing and Evaluating Search Engines.* MIT Press, 2010.

[BD02] P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting* (2nd ed.). Springer, 2002.

[BDF+97] D. Barbará, W. DuMouchel, C. Faloutos, 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.

[Ben08] S. Benninga. *Financial Modeling, 3rd. ed.* MIT Press, 2008.

[Ber03] M. W. Berry. *Survey of Text Mining: Clustering, Classification, and Retrieval.* Springer, 2003.

[BFOS84] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth International Group, 1984.

[BG04] I. Bhattacharya and L. Getoor. Iterative record linkage for cleaning and integration. In *Proc. SIGMOD 2004 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'04)*, pages 11–18, Paris, France, June 2004.

[BJR08] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control* (4th ed.). Prentice-Hall, 2008.

[BL04] M. J. A. Berry and G. S. Linoff. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* John Wiley & Sons, 2004.

[BL09] D. Blei and J. Lafferty. *Topic Models.* Taylor and Francis, 2009.

[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] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. 7th Int. World Wide Web Conf. (WWW'98)*, pages 107–117, Brisbane, Australia, April 1998.

[BST99] A. Berson, S. J. Smith, and K. Thearling. *Building Data Mining Applications for CRM.* McGraw-Hill, 1999.

[BW01] S. Babu and J. Widom. Continuous queries over data streams. *SIGMOD Record*, 30:109–120, 2001.

[BYRN11] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval, 2nd ed.* Addison-Wesley, 2011.

[CDH+02] Y. Chen, G. Dong, J. Han, B. W. Wah, and J. Wang. Multi-dimensional regression analysis of time-series data streams. In *Proc. 2002 Int. Conf. Very Large Data Bases (VLDB'02)*, pages 323–334, Hong Kong, China, Aug. 2002.

[CDI98] S. Chakrabarti, B. E. Dom, and P. Indyk. Enhanced hypertext classification using hyper-links. In *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98)*, pages 307–318, Seattle, WA, June 1998.

[CDK+99] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. M. Kleinberg. Mining the web's link structure. *COMPUTER*, 32:60–67, 1999.

[Cha03a] S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data.* Morgan Kaufmann, 2003.

[Cha03b] C. Chatfield. *The Analysis of Time Series: An Introduction* (6th ed.). Chapman and Hall, 2003.

[CKT06] D. Chakrabarti, R. Kumar, and A. Tomkins. Evolutionary clustering,. In *Proc. 2006 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'06)*, pages 554–560, Philadelphia, PA, Aug. 2006.

[Cle93] W. Cleveland. *Visualizing Data.* Hobart Press, 1993.

[CMS09] B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice.* Addison Wesley, 2009.

[CSZ+07] Y. Chi, X. Song, D. Zhou, K. Hino, and B. L. Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. In *Proc. 2007 ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining (KDD'07)*, San Jose, CA, Aug. 2007.

[CWL+08] G. Cong, L. Wang, C.-Y. Lin, Y.-I. Song, and Y. Sun. Finding question-answer pairs from online forums. In *Proc. 2008 Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'08)*, pages 467–474, Singapore, July 2008.

[CYZ+08] C. Chen, X. Yan, F. Zhu, J. Han, and P. S. Yu. Graph OLAP: Towards online analytical processing on graphs. In *Proc. 2008 Int. Conf. Data Mining (ICDM'08)*, Pisa, Italy, Dec. 2008.

[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.

[DH00] P. Domingos and G. Hulten. Mining high-speed data streams. In *Proc. 2000 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'00)*, pages 71–80, Boston, MA, Aug. 2000.

[DHS01] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification* (2nd ed.). John Wiley & Sons, 2001.

[Dob01] A. J. Dobson. *An Introduction to Generalized Linear Models* (2nd ed.). Chapman and Hall, 2001.

[DP07] G. Dong and J. Pei. *Sequence Data Mining*. Springer, 2007.

[Dwo06] C. Dwork. Differential privacy. In *Proc. 2006 Int. Col. Automata, Languages and Programming (ICALP)*, Venice, Italy, July 2006.

[EK10] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge Univ. Press, 2010.

[ESAG02] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *Proc. 2002 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'02)*, pages 217–228, Edmonton, Canada, July 2002.

[FFF99] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proc. ACM SIGCOMM'99 Conf. Applications, Technologies, Architectures, and Protocols for Computer Communication*, pages 251–262, Cambridge, MA, Aug. 1999.

[FGW01] U. Fayyad, G. Grinstein, and A. Wierse. *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann, 2001.

[FL95] C. Faloutsos and K.-I. Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proc. 1995 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'95)*, pages 163–174, San Jose, CA, May 1995.

[FPP07] D. Freedman, R. Pisani, and R. Purves. *Statistics (4th ed.)*. W. W. Norton & Co., 2007.

[FRM94] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. In *Proc. 1994 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'94)*, pages 419–429, Minneapolis, MN, May 1994.

[FS93] U. Fayyad and P. Smyth. Image database exploration: Progress and challenges. In *Proc. AAAI'93 Workshop on Knowledge Discovery in Databases (KDD'93)*, pages 14–27, Washington, DC, July 1993.

[FS06] R. Feldman and J. Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge Uni. Press, 2006.

[FWFY10] B.C.M. Fung, K. Wang, A.W.-C. Fu, and P. S. Yu. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques.* Chapman & Hall/CRC, 2010.

[Gat00] B. Gates. *Business @ the Speed of Thought: Succeeding in the Digital Economy.* Warner Books, 2000.

[GF04] D. A. Grossman and O. Frieder. *Information Retrieval: Algorithms and Heuristics.* Springer, 2004.

[GFKT01] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of relational structure. In *Proc. 2001 Int. Conf. Machine Learning (ICML'01)*, pages 170–177, Williamstown, MA, 2001.

[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.

[GMMO00] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams. In *Proc. 2000 Symp. Foundations of Computer Science (FOCS'00)*, pages 359–366, Redondo Beach, CA, 2000.

[GR07] P. D. Grunwald and J. Rissanen. *The Minimum Description Length Principle.* The MIT Press, 2007.

[GW07] R. C. Gonzalez and R. E. Woods. *Digital Image Processing* (3rd ed.). Prentice Hall, 2007.

[Ham94] J. Hamilton. *Time Series Analysis.* Princeton Univ. Press, 1994.

[HCD94] L. B. Holder, D. J. Cook, and S. Djoko. Substructure discovery in the subdue system. In *Proc. AAAI'94 Workshop on Knowledge Discovery in Databases (KDD'94)*, pages 169–180, Seattle, WA, July 1994.

[Hig08] R. C. Higgins. *Analysis for Financial Management with S&P Bind-In Card.* Irwin/McGraw-Hill, 2008.

[HKGT03] M. Hadjieleftheriou, G. Kollios, D. Gunopulos, and V. J. Tsotras. On-line discovery of dense areas in spatio-temporal databases. In *Proc. 2003 Int. Symp. Spatial and Temporal Databases (SSTD'03)*, pages 306–324, Santorini Island, Greece, July 2003.

[HLW07] W. Hsu, M. L. Lee, and J. Wang. *Temporal and Spatio-Temporal Data Mining.* IGI Publishing, 2007.

[HLZ02] W. Hsu, M. L. Lee, and J. Zhang. Image mining: Trends and developments. *J. Intelligent Information Systems*, 19:7–23, 2002.

[Hor08] R. Horak. *Telecommunications and Data Communications Handbook, 2nd ed.* Wiley-Interscience, 2008.

[HRMS10] M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially-private queries through consistency. In *Proc. 2010 Int. Conf. Very Large Data Bases (VLDB'10)*, Singapore, Sept. 2010.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer Verlag, 2009.

[Hub94] C. H. Huberty. *Applied Discriminant Analysis.* New York, 1994.

[HWB+04] J. Huan, W. Wang, D. Bandyopadhyay, J. Snoeyink, J. Prins, and A. Tropsha. Mining spatial motifs from protein structure graphs. In *Proc. 8th Int. Conf. Research in Computational Molecular Biology (RECOMB)*, pages 308–315, San Diego, CA, March 2004.

[IM96] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. *Comm. ACM*, 39:58–64, 1996.

[IWM98] A. Inokuchi, T. Washio, and H. Motoda. An apriori-based algorithm for mining frequent substructures from graph data. In *Proc. 2000 European Symp. Principles of Data Mining and Knowledge Discovery (PKDD'00)*, pages 13–23, Lyon, France, Sept. 1998.

[JBD05] X. Ji, J. Bailey, and G. Dong. Mining minimal distinguishing subsequence patterns with gap constraints. In *Proc. 2005 Int. Conf. Data Mining (ICDM'05)*, pages 194–201, Houston, TX, Nov. 2005.

[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.

[JSD+10] M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao. Graph regularized transductive classification on heterogeneous information networks. In *Proc. 2010 European Conf. Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD'10)*, Barcelona, Spain, Sept. 2010.

[JW02] R. A. Johnson and D. A. Wichern. *Applied Multivariate Statistical Analysis* (5th ed.). Prentice Hall, 2002.

[Kam09] C. Kamath. *Scientific Data Mining: A Practical Perspective*. Society for Industrial and Applied Mathematic (SIAM), 2009.

[KBDM09] B. Kulis, S. Basu, I. Dhillon, and R. Mooney. Semi-supervised graph clustering: a kernel approach. *Machine Learning*, 74:1–22, 2009.

[KF09] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[KH95] K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Proc. 1995 Int. Symp. Large Spatial Databases (SSD'95)*, pages 47–66, Portland, ME, Aug. 1995.

[KH09] M.-S. Kim and J. Han. A particle-and-density based evolutionary clustering method for dynamic networks. In *Proc. 2009 Int. Conf. Very Large Data Bases (VLDB'09)*, Lyon, France, Aug. 2009.

[KHY+08] H. Kargupta, J. Han, P. S. Yu, R. Motwani, and V. Kumar. *Next Generation of Data Mining*. Chapman & Hall/CRC, 2008.

[KJSY04] H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha. *Data Mining: Next Generation Challenges and Future Directions*. AAAI/MIT Press, 2004.

[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.

[Kle99a] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:604–632, 1999.

[Kle99b] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:604–632, 1999.

[KMS03] J. Kubica, A. Moore, and J. Schneider. Tractable group detection on large link data sets. In *Proc. 2003 Int. Conf. Data Mining (ICDM'03)*, pages 573–576, Melbourne, FL, Nov. 2003.

[KNNL04] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models with Student CD*. Irwin, 2004.

[KPR98] J. M. Kleinberg, C. Papadimitriou, and P. Raghavan. A microeconomic view of data mining. *Data Mining and Knowledge Discovery*, 2:311–324, 1998.

[KPS03] R. M. Karp, C. H. Papadimitriou, and S. Shenker. A simple algorithm for finding frequent elements in streams and bags. *ACM Trans. Database Systems*, 28, 2003.

[KR03] D. Krane and R. Raymer. *Fundamental Concepts of Bioinformatics*. Benjamin Cummings, 2003.

[Kre02] V. Krebs. Mapping networks of terrorist cells. *Connections*, 24:43–52, Winter 2002.

[KRR+00] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proc. 2000 IEEE Symp. Foundations of Computer Science (FOCS'00)*, pages 57–65, Redondo Beach, CA, Nov. 2000.

[KT99] J. M. Kleinberg and A. Tomkins. Application of linear algebra in information retrieval and hypertext analysis. In *Proc. 18th ACM Symp. Principles of Database Systems (PODS'99)*, pages 185–193, Philadelphia, PA, May 1999.

[KYB03] I. Korf, M. Yandell, and J. Bedell. *BLAST*. O'Reilly Media, Sebastopol, CA, 2003.

[LDH+10] Z. Li, B. Ding, J. Han, R. Kays, and P. Nye. Mining periodic behaviors for moving objects. In *Proc. 2010 ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD'10)*, Washington D.C., July 2010.

[LHKG07] X. Li, J. Han, S. Kim, and H. Gonzalez. Roam: Rule- and motif-based anomaly detection in massive moving object data sets. In *Proc. 2007 SIAM Int. Conf. Data Mining (SDM'07)*, Minneapolis, MN, April 2007.

[LHW07] J.-G. Lee, J. Han, and K. Whang. Clustering trajectory data. In *Proc. 2007 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'07)*, Beijing, China, June 2007.

[Liu06] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, 2006.

[LKF05] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proc. 2005 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'05)*, pages 177–187, Chicago, IL, Aug. 2005.

[LNK03] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proc. 2003 Int. Conf. Information and Knowledge Management (CIKM'03)*, pages 556–559, New Orleans, LA, Nov. 2003.

[Man00] H. Mannila. Theoretical frameworks of data mining. *SIGKDD Explorations*, 1:30–32, 2000.

[MCK+04] N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao, and D. Cheung. Mining, indexing, and querying historical spatiotemporal data. In *Proc. 2004 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'04)*, pages 236–245, Seattle, WA, Aug. 2004.

[Men03] J. Mena. *Investigative Data Mining with Security and Criminal Detection*. Butterworth-Heinemann, 2003.

[MH09] H. Miller and J. Han. *Geographic Data Mining and Knowledge Discovery (2nd ed.)*. Chapman & Hall/CRC, 2009.

[Mil98] R. G. Miller. *Survival Analysis*. Wiley-Interscience, 1998.

[MLSZ06] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proc. 15th Int. Conf. on World Wide Web (WWW'06)*, pages 533–542, Edinburgh, Scotland, May 2006.

[MM02] G. Manku and R. Motwani. Approximate frequency counts over data streams. In *Proc. 2002 Int. Conf. Very Large Data Bases (VLDB'02)*, pages 346–357, Hong Kong, China, Aug. 2002.

[MRS08] C. D. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[Mut05] S. Muthukrishnan. *Data Streams: Algorithms and Applications*. Now Publishers, 2005.

[MZ06] Q. Mei and C. Zhai. A mixture model for contextual text mining. In *Proc. 2006 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'06)*, pages 649–655, Philadelphia, PA, Aug. 2006.

[NBW06] M. Newman, A.-L. Barabasi, and D. J. Watts. *The Structure and Dynamics of Networks*. Princeton Univ. Press, 2006.

[New10] M. Newman. *Networks: An Introduction*. Oxford Univ. Press, 2010.

[NG04] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69, 2004.

[NGER09] J. Neville, B. Gallaher, and T. Eliassi-Rad. Evaluating statistical tests for within-network classifiers of relational data. In *Proc. 2009 Int. Conf. Data Mining (ICDM'09)*, pages 397–406, Miami, FL, Dec. 2009.

[NK04] S. Nijssen and J. Kok. A quickstart in frequent structure mining can make a difference. In *Proc. 2004 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'04)*, pages 647–652, Seattle, WA, Aug. 2004.

[NRS99] A. Natsev, R. Rastogi, and K. Shim. Walrus: A similarity retrieval algorithm for image databases. In *Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99)*, pages 395–406, Philadelphia, PA, June 1999.

[PB00] J. C. Pinheiro and D. M. Bates. *Mixed Effects Models in S and S-PLUS*. Springer Verlag, 2000.

[PHMA+04] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. Mining sequential patterns by pattern-growth: The PrefixSpan approach. *IEEE Trans. Knowledge and Data Engineering*, 16:1424–1440, 2004.

[PL07] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135, 2007.

[Rab89] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77:257–286, 1989.

[RGN10] L. De Raedt, T. Guns, and S. Nijssen. Constraint programming for data mining and machine learning. In *Proc. 2010 AAAI Conf. Artificial Intelligence (AAAI'10)*, Atlanta, GA, July 2010.

[RHS01] J. F. Roddick, K. Hornsby, and M. Spiliopoulou. An updated bibliography of temporal, spatial, and spatio-temporal data mining research. In *Lecture Notes in Computer Science 2007*, pages 147–163, Springer, 2001.

[Rus06] J. C. Russ. *The Image Processing Handbook* (5th ed.). CRC Press, 2006.

[SA96] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *Proc. 5th Int. Conf. Extending Database Technology (EDBT'96)*, pages 3–17, Avignon, France, Mar. 1996.

[SC03] S. Shekhar and S. Chawla. *Spatial Databases: A Tour*. Prentice Hall, 2003.

[SD02] T. Soukup and I. Davidson. *Visual Data Mining: Techniques and Tools for Data Visualization and Mining.* Wiley, 2002.

[SHK00] N. Stefanovic, J. Han, and K. Koperski. Object-based selective materialization for efficient implementation of spatial data cubes. *IEEE Trans. Knowledge and Data Engineering*, 12:938–958, 2000.

[SHZ+09] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. RankClus: Integrating clustering with ranking for heterogeneous information network analysis. In *Proc. 2009 Int. Conf. Extending Data Base Technology (EDBT'09)*, Saint-Petersburg, Russia, Mar. 2009.

[Sil10] F. Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 4:1–174, 2010.

[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.

[STH+10] Y. Sun, J. Tang, J. Han, M. Gupta, and B. Zhao. Community evolution detection in dynamic heterogeneous information networks. In *Proc. 2010 KDD Workshop on Mining and Learning with Graphs (MLG'10)*, Washington D.C., July 2010.

[SZ04] D. Shasha and Y. Zhu. *High Performance Discovery In Time Series: Techniques and Case Studies.* Springer, 2004.

[TFPL04] Y. Tao, C. Faloutsos, D. Papadias, and B. Liu. Prediction and indexing of moving objects with unknown motion patterns. In *Proc. 2004 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'04)*, Paris, France, June 2004.

[TG01] I. Tsoukatos and D. Gunopulos. Efficient mining of spatiotemporal patterns. In *Proc. 2001 Int. Symp. Spatial and Temporal Databases (SSTD'01)*, pages 425–442, Redondo Beach, CA, July 2001.

[THP08] Y. Tian, R. A. Hankins, and J. M. Patel. Efficient aggregation for graph summarization. In *Proc. 2008 ACM SIGMOD Int. Conf. Management of Data (SIGMOD'08)*, pages 567–580, Vancouver, BC, Canada, June 2008.

[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.

[TLZN08] L. Tang, H. Liu, J. Zhang, and Z. Nazeri. Community evolution in dynamic multi-mode networks. In *Proc. 2008 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'08)*, Las Vegas, NV, Aug. 2008.

[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.

[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.

[VCZ10] J. Vaidya, C. W. Clifton, and Y. M. Zhu. *Privacy Preserving Data Mining*. Springer, 2010.

[VGK02] M. Vlachos, D. Gunopulos, and G. Kollios. Discovering similar multidimensional trajectories. In *Proc. 2002 Int. Conf. Data Engineering (ICDE'02)*, pages 673–684, San Fransisco, CA, April 2002.

[Wat95] M. S. Waterman. *Introduction to Computational Biology: Maps, Sequences, and Genomes (Interdisciplinary Statistics)*. CRC Press, 1995.

[Wat03] D. J. Watts. *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company, 2003.

[WF94] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[WFYH03] H. Wang, W. Fan, P. S. Yu, and J. Han. Mining concept-drifting data streams using ensemble classifiers. In *Proc. 2003 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'03)*, pages 226–235, Washington, DC, Aug. 2003.

[WHJ+10] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining advisor-advisee relationships from research publication networks. In *Proc. 2010 ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD'10)*, Washington D.C., July 2010.

[WIZD04] S. Weiss, N. Indurkhya, T. Zhang, and F. Damerau. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer, 2004.

[XPK10] Z. Xing, J. Pei, and E. Keogh. A brief survey on sequence classification. *SIGKDD Explorations*, 12:40–48, 2010.

[XYFS07] X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger. SCAN: A structural clustering algorithm for networks. In *Proc. 2007 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'07)*, San Jose, CA, Aug. 2007.

[XZYL08] T. Xu, Z. M. Zhang, P. S. Yu, and B. Long. Evolutionary clustering by hierarchical Dirichlet process with hidden Markov state. In *Proc. 2008 Int. Conf. Data Mining (ICDM'08)*, Pisa, Italy, Dec. 2008.

[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.

[YH03] X. Yan and J. Han. CloseGraph: Mining closed frequent graph patterns. In *Proc. 2003 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'03)*, pages 286–295, Washington, DC, Aug. 2003.

[YHA03] X. Yan, J. Han, and R. Afshar. CloSpan: Mining closed sequential patterns in large datasets. In *Proc. 2003 SIAM Int. Conf. Data Mining (SDM'03)*, pages 166–177, San Fransisco, CA, May 2003.

[YHF10] P. S. Yu, J. Han, and C. Faloutsos. *Link Mining: Models, Algorithms and Applications*. Springer, 2010.

[YHY05] X. Yin, J. Han, and P. S. Yu. Cross-relational clustering with user's guidance. In *Proc. 2005 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'05)*, pages 344–353, Chicago, IL, Aug. 2005.

[YHY07] X. Yin, J. Han, and P. S. Yu. Object distinction: Distinguishing objects with identical names by link analysis. In *Proc. 2007 Int. Conf. Data Engineering (ICDE'07)*, Istanbul, Turkey, April 2007.

[YHY08] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the Web. *IEEE Trans. Knowledge and Data Engineering*, 20:796–808, 2008.

[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.

[YK09] L. Ye and E. Keogh. Time series shapelets: A new primitive for data mining. In *Proc. 2009 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'09)*, Paris, France, June 2009.

[Zak01] M. Zaki. SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning*, 40:31–60, 2001.

[Zha08] C. Zhai. *Statistical Language Models for Information Retrieval.* Morgan and Claypool, 2008.

[ZHZ00] O. R. Zaïane, J. Han, and H. Zhu. Mining recurrent items in multimedia with progressive resolution refinement. In *Proc. 2000 Int. Conf. Data Engineering (ICDE'00)*, pages 461–470, San Diego, CA, Feb. 2000.

[ZZ09] Z. Zhang and R. Zhang. *Multimedia Data Mining: A Systematic Introduction to Concepts and Theory.* Chapman & Hall, 2009.