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/.
CHAPTER 13. TRENDS AND RESEARCH FRONTIERS IN DATA MINING

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 Barabasi [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 [YHY04], 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, Gallagher, 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 [LKN03], 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, Kollias, Hadjieleftheriou, et al. [MCK+04], Tsoukatos and Gunopulos [TG01], and Hadjieleftheriou, Kollias, Gunopulos, and Tzotras [HKG03]. Mining moving object data has been studied by many, such as Vlachos, Gunopulos, and Kollias [VGK02], Tao, Faloutsos, Papadias, and Liu [TFPL04], Li, Han, Kim and Gonzalez [LHK07], 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 [ZH09]). 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. Butcher, 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 [W FY03], Aggarwal, Han, Wang and Yu [AHWY04], and stream clustering, e.g., Guha, Mishra, Motwani, and O’Callaghan [GMM00], 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, Kegelmeier, 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, Faloutsos, 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 [BPOS84], 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


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


BIBLIOGRAPHY

economic view of data mining. *Data Mining and Knowledge Dis-

gorithm for finding frequent elements in streams and bags. *ACM

is*. Benjamin Cummings, 2003.


[KRR+00] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar,

[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,

[KYB03] I. Korf, M. Yandell, and J. Bedell. *BLAST*. O’Reilly Media,

[LDH+10] Z. Li, B. Ding, J. Han, R. Kays, and P. Nye. Mining periodic
Knowledge Discovery and Data Mining (KDD’10)*, Washington
D.C., July 2010.

[LHK07] 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.

MOD’07)*, Beijing, China, June 2007.


[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


[NK04] S. Nijssen and J. Kok. A quickstart in frequent structure min-
ing can make a difference. In Proc. 2004 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD’04), pages 647–

Int. Conf. Management of Data (SIGMOD’99), pages 395–406,


[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


[Rab89] L. R. Rabiner. A tutorial on hidden Markov models and selected

Artificial Intelligence (AAAI’10), Atlanta, GA, July 2010.

bibliography of temporal, spatial, and spatio-temporal data mining
research. In Lecture Notes in Computer Science 2007, pages 147–163,

2006.

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

BIBLIOGRAPHY


