Chapter 1

Introduction

1.1 Bibliographic Notes

The book *Knowledge Discovery in Databases*, edited by Piatetsky-Shapiro and Frawley [PSF91], is an early collection of research papers on knowledge discovery from data. The book *Advances in Knowledge Discovery and Data Mining*, edited by Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy [FPSSe96], is a collection of later research results on knowledge discovery and data mining. There have been many data mining books published in recent years, including *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman [HTF09], *Introduction to Data Mining* by Tan, Steinbach and Kumar [TSK05], *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* by Witten, Frank, and Hall [WFH11], *Predictive Data Mining* by Weiss and Indurkhya [WI98], *Mastering Data Mining: The Art and Science of Customer Relationship Management* by Berry and Linoff [BL99], *Principles of Data Mining (Adaptive Computation and Machine Learning)* by Hand, Mannila, and Smyth [HMS01], *Mining the Web: Discovering Knowledge from Hypertext Data* by Chakrabarti [Cha03], *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* by Liu [Liu06], *Data Mining: Introductory and Advanced Topics* by Dunham [Dun03], and *Data Mining: Multimedia, Soft Computing, and Bioinformatics* by Mitra and Acharya [MA03]. There are also books containing collections of papers or chapters on particular aspects of knowledge discovery, such as *Relational Data Mining* edited by Dzeroski and Lavrac [De01], *Mining Graph Data* edited by Cook and Holder [CH07], *Data Streams: Models and Algorithms* edited by Aggarwal [Agg06], *Next Generation of Data Mining* edited by Kargupta, Han, Yu, et al. [KHY+08], *Multimedia Data Mining: A Systematic Introduction to Concepts and Theory* edited by Z. Zhang and R. Zhang [ZZ09], *Geographic Data Mining and Knowledge Discovery* edited by Miller and Han [MH09], and *Link Mining—Models, Algorithms and Applications* edited by Yu, Han and Faloutsos [YHF10]. There are many tutorial notes on data mining in major database, data mining, machine learning, statistics,

and web technology conferences.

KDNuggets is a regular electronic newsletter containing information relevant to knowledge discovery and data mining, moderated by Piatetsky-Shapiro since 1991. The Internet site *KDNuggets* (www.kdnuggets.com) contains a good collection of KDD-related information.

The data mining community started its first international conference on knowledge discovery and data mining in 1995. The conference evolved from the four international workshops on knowledge discovery in databases, held from 1989 to 1994. ACM-SIGKDD, a Special Interest Group on Knowledge Discovery in Databases was set up under ACM in 1998 and has been organizing the international conferences on knowledge discovery and data mining since 1999. IEEE Computer Science Society has organized its annual data mining conference, International Conference on Data Mining (ICDM), since 2001. SIAM (Society on Industrial and Applied Mathematics) has organized its annual data mining conference, SIAM Data Mining conference (SDM), since 2002. A dedicated journal, *Data Mining and Knowledge Discovery*, published by Kluwers Publishers, has been available since 1997. An ACM journal, *ACM Transactions on Knowledge discovery from Data*, started its first volume in 2007. ACM-SIGKDD also publishes a biannual newsletter, *SIGKDD Explorations*. There are a few other international or regional conferences on data mining, such as the *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (ECML PKDD), the *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (PAKDD), and the *International Conference on Data Warehousing and Knowledge Discovery* (DaWaK).

Research in data mining has also been published in books, conferences, and journals on databases, statistics, machine learning, and data visualization. References to such sources are listed below.

Popular textbooks on database systems include *Database Systems: The Complete Book* by Garcia-Molina, Ullman, and Widom [GMUW08], *Database Management Systems* by Ramakrishnan and Gehrke [RG03], *Database System Concepts* by Silberschatz, Korth, and Sudarshan [SKS10], and *Fundamentals of Database Systems* by Elmasri and Navathe [EN10]. For an edited collection of seminal articles on database systems, see *Readings in Database Systems* by Hellerstein and Stonebraker [HS05]. There are also many books on data warehouse technology, systems, and applications, such as *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* by Kimball and Ross [KR02], *The Data Warehouse Lifecycle Toolkit:* by Kimball, Ross, Thornthwaite et al. [KRTM08], *Mastering Data Warehouse Design: Relational and Dimensional Techniques* by Imhoff, Galemmo and Geiger [IGG03], and *Building the Data Warehouse* by Inmon [Inm96]. A set of research papers on materialized views and data warehouse implementations were collected in *Materialized Views: Techniques, Implementations, and Applications* by Gupta and Mumick [GM99]. Chaudhuri and Dayal [CD97] present an early comprehensive overview of data warehouse technology.

Research results relating to data mining and data warehousing have been published in the proceedings of many international database conferences, in-

cluding the *ACM-SIGMOD International Conference on Management of Data (SIGMOD)*, the *International Conference on Very Large Data Bases (VLDB)*, the *ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, the *International Conference on Data Engineering (ICDE)*, the *International Conference on Extending Database Technology (EDBT)*, the *International Conference on Database Theory (ICDT)*, the *International Conference on Information and Knowledge Management (CIKM)*, the *International Conference on Database and Expert Systems Applications (DEXA)*, and the *International Symposium on Database Systems for Advanced Applications (DASFAA)*. Research in data mining is also published in major database journals, such as *IEEE Transactions on Knowledge and Data Engineering (TKDE), ACM Transactions on Database Systems (TODS), Information Systems, The VLDB Journal, Data and Knowledge Engineering*, *International Journal of Intelligent Information Systems (JIIS)*, and *Knowledge and Information Systems (KAIS)*.

Many effective data mining methods have been developed by statisticians, and introduced in a rich set of textbooks. An overview of classification from a statistical pattern recognition perspective can be found in *Pattern Classification* by Duda, Hart, Stork [DHS01]. There are also many textbooks covering different topics in statistical analysis, such as *Mathematical Statistics: Basic Ideas and Selected Topics* by Bickel and Doksum [BD01], *The Statistical Sleuth: A Course in Methods of Data Analysis* by Ramsey and Schafer [RS01], *Applied Linear Statistical Models* by Neter, Kutner, Nachtsheim, and Wasserman [NKNW96], *An Introduction to Generalized Linear Models* by Dobson [Dob90], *Applied Statistical Time Series Analysis* by Shumway [Shu88], and *Applied Multivariate Statistical Analysis* by Johnson and Wichern [JW92].

Research in statistics is published in the proceedings of several major statistical conferences, including *Joint Statistical Meetings*, *International Conference of the Royal Statistical Society*, and *Symposium on the Interface: Computing Science and Statistics*. Other sources of publication include the *Journal of the Royal Statistical Society*, *The Annals of Statistics*, *Journal of American Statistical Association*, *Technometrics*, and *Biometrika*.

Textbooks and reference books on machine learning and pattern recognition include *Machine Learning* by Mitchell [Mit97], *Pattern Recognition and Machine Learning* by Bishop [Bis06], *Pattern Recognition* by Theodoridis and Koutroumbas [TK08], *Introduction to Machine Learning* by Alpaydin [Alp11], *Probabilistic Graphical Models: Principles and Techniques* by Koller and Friedman [KF09], and *Machine Learning: An Algorithmic Perspective* by Marsland [Mar09]. For an edited collection of seminal articles on machine learning, see *Machine Learning, An Artificial Intelligence Approach*, Vols 1–4, edited by Michalski et al. [MCM83, MCM86, KM90, MT94], and *Readings in Machine Learning* by Shavlik and Dietterich [SD90].

Machine learning and pattern recognition research is published in the proceedings of several major machine learning, artificial intelligence, and pattern recognition conferences, including the *International Conference on Machine Learning (ML)*, the *ACM Conference on Computational Learning Theory (COLT)*, the *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,

the *International Conference on Pattern Recognition (ICPR)*, the *International Joint Conference on Artificial Intelligence (IJCAI)*, and the *American Association of Artificial Intelligence Conference (AAAI)*. Other sources of publication include major machine learning, artificial intelligence, pattern recognition, and knowledge system journals, some of which have been mentioned above. Others include *Machine Learning (ML), Pattern Recognition (PR), Artificial Intelligence Journal (AI)*, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, and *Cognitive Science*.

Textbooks and reference books on information retrieval include *Introduction to Information Retrieval* by Manning, Raghavan, and Schutz [MRS08], *Information Retrieval: Implementing and Evaluating Search Engines* by Bttcher, Clarke, and Cormack [BCC10], *Search Engines: Information Retrieval in Practice* by Croft, Metzler, and Strohman [CMS09], *Modern Information Retrieval: The Concepts and Technology behind Search* by Baeza-Yates and Ribeiro-Neto [BYRN11], and *Information Retrieval: Algorithms and Heuristics* by Grossman and Frieder [GF04].

Information retrieval research is published in the proceedings of several information retrieval and Web search and mining conferences, including the *International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR)*, the *International World Wide Web Conference (WWW)*, the *ACM International Conference on Web Search and Data Mining (WSDM)*, the *ACM Conference on Information and Knowledge Management (CIKM)*, the *European Conference on Information Retrieval (ECIR)*, the *Text Retrieval Conference (TREC)*, and the *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Other sources of publication include major information retrieval, information systems, and Word-Wide-Web journals, such as *Journal of Information Retrieval, ACM Transactions on Information Systems (TOIS)*, *Information Processing and Management*, *Knowledge and Information Systems (KAIS)*, and *IEEE Transactions on Knowledge and Data Engineering (TKDE)*.

Bibliography

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

[Alp11] E. Alpaydin. *Introduction to Machine Learning (2nd ed.).* MIT Press, 2011.

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

[BD01] P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Vol. 1.* Prentice Hall, 2001.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[BL99] M. J. A. Berry and G. Linoff. *Mastering Data Mining: The Art and Science of Customer Relationship Management.* John Wiley & Sons, 1999.

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

[CD97] S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. *SIGMOD Record*, 26:65–74, 1997.

[CH07] D. J. Cook and L. B. Holder. *Mining Graph Data.* John Wiley & Sons, 2007.

[Cha03] S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data.* Morgan Kaufmann, 2003.

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

[De01] S. Dzeroski and N. Lavrac (eds.). *Relational Data Mining.* Springer, 2001.

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

[Dob90] A. J. Dobson. *An Introduction to Generalized Linear Models*. Chapman and Hall, 1990.

[Dun03] M. Dunham. *Data Mining: Introductory and Advanced Topics*. Prentice Hall, 2003.

[EN10] R. Elmasri and S. B. Navathe. *Fundamental of Database Systems* (6th ed.). Addison Wesley, 2010.

[FPSSe96] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.). *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.

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

[GM99] A. Gupta and I. S. Mumick. *Materialized Views: Techniques, Implementations, and Applications*. MIT Press, 1999.

[GMUW08] H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database Systems: The Complete Book* (2nd ed.). Prentice Hall, 2008.

[HMS01] D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, 2001.

[HS05] J. M. Hellerstein and M. Stonebraker. *Readings in Database Systems* (4th ed.). MIT Press, 2005.

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

[IGG03] C. Imhoff, N. Galemmo, and J. G. Geiger. *Mastering Data Warehouse Design : Relational and Dimensional Techniques*. John Wiley & Sons, 2003.

[Inm96] W. H. Inmon. *Building the Data Warehouse*. John Wiley & Sons, 1996.

[JW92] R. A. Johnson and D. A. Wichern. *Applied Multivariate Statistical Analysis* (3rd ed.). Prentice Hall, 1992.

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

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

[KM90] Y. Kodratoff and R. S. Michalski. *Machine Learning, An Artificial Intelligence Approach, Vol. 3.* Morgan Kaufmann, 1990.

[KR02] R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (2nd ed.). John Wiley & Sons, 2002.

[KRTM08] R. Kimball, M. Ross, W. Thornthwaite, and J. Mundy. *The Data Warehouse Lifecycle Toolkit.* John Wiley & Sons, 2008.

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

[MA03] S. Mitra and T. Acharya. *Data Mining: Multimedia, Soft Computing, and Bioinformatics.* John Wiley & Sons, 2003.

[Mar09] S. Marsland. *Machine Learning: An Algorithmic Perspective.* Chapman and Hall/CRC, 2009.

[MCM83] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell. *Machine Learning, An Artificial Intelligence Approach, Vol. 1.* Morgan Kaufmann, 1983.

[MCM86] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell. *Machine Learning, An Artificial Intelligence Approach, Vol. 2.* Morgan Kaufmann, 1986.

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

[Mit97] T. M. Mitchell. *Machine Learning.* McGraw-Hill, 1997.

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

[MT94] R. S. Michalski and G. Tecuci. *Machine Learning, A Multistrategy Approach, Vol. 4.* Morgan Kaufmann, 1994.

[NKNW96] J. Neter, M. H. Kutner, C. J. Nachtsheim, and L. Wasserman. *Applied Linear Statistical Models* (4th ed.). Irwin, 1996.

[PSF91] G. Piatetsky-Shapiro and W. J. Frawley. *Knowledge Discovery in Databases.* AAAI/MIT Press, 1991.

[RG03] R. Ramakrishnan and J. Gehrke. *Database Management Systems* (3rd ed.). McGraw-Hill, 2003.

[RS01] F. Ramsey and D. Schafer. *The Statistical Sleuth: A Course in Methods of Data Analysis.* Duxbury Press, 2001.

[SD90] J. W. Shavlik and T. G. Dietterich. *Readings in Machine Learning.* Morgan Kaufmann, 1990.

[Shu88] R. H. Shumway. *Applied Statistical Time Series Analysis*. Prentice Hall, 1988.

[SKS10] A. Silberschatz, H. F. Korth, and S. Sudarshan. *Database System Concepts* (6th ed.). McGraw-Hill, 2010.

[TK08] S. Theodoridis and K. Koutroumbas. *Pattern Recognition, 4th ed.* Academic Press, 2008.

[TSK05] P. N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, 2005.

[WFH11] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmann, 2011.

[WI98] S. M. Weiss and N. Indurkhya. *Predictive Data Mining*. Morgan Kaufmann, 1998.

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

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