Data Mining: Concepts and Techniques (2nd edition)

Jiawei Han and Micheline Kamber
Morgan Kaufmann Publishers, 2006

Bibliographic Notes for Chapter 10
Mining Object, Spatial, Multimedia, Text, and Web Data

Mining complex types of data has been a fast developing, popular research field, with many research papers and tutorials appearing in conferences and journals on data mining and database systems. This chapter covers a few important themes, including multidimensional analysis and mining of complex data objects, spatial data mining, multimedia data mining, text mining, and Web mining.

Zaniolo, Ceri, Faloutsos, et al. [ZCF+97] present a systematic introduction of advanced database systems for handling complex types of data. For multidimensional analysis and mining of complex data objects, Han, Nishio, Kawano, and Wang [HNKW98] proposed a method for the design and construction of object cubes by multidimensional generalization and its use for mining complex types of data in object-oriented and object-relational databases. A method for the construction of multiple layered databases by generalization-based data mining techniques for handling semantic heterogeneity was proposed by Han, Ng, Fu, and Dao [HNFD98]. Zaki, Lesh and Ogihara worked out a system called PlanMine, which applies sequence mining for plan failures [ZLO98]. A generalization-based method for mining plan databases by divide-and-conquer was proposed by Han, Yang, and Kim [HYK99].

Geospatial database systems and spatial data mining have been studied extensively. Some introductory materials about spatial database can be found in Maguire, Goodchild, and Rhind [MGR92], Güting [Gue94], Egenhofer [Ege89], Shekhar, Chawla, Ravada, et al. [SCR+99], Rigaux, Scholl, and Voisard [RSV01], and Shekhar and Chawla [SC03]. For geospatial data mining, a comprehensive survey on spatial data mining methods can be found in Ester, Kriegel, and Sander [EKS97] and Shekhar and Chawla [SC03]. A collection of research contributions on geographic data mining and knowledge discovery are in Miller and Han [MH01]. Lu, Han, and Ooi [LHO93] proposed a generalization-based spatial data mining method by attribute-oriented induction. Ng and Han [NH94] proposed performing descriptive spatial data analysis based on clustering results instead of on predefined concept hierarchies. Zhou, Truffet, and Han proposed efficient polygon amalgamation methods for on-line multidimensional spatial analysis and spatial data mining [ZTH99]. Stefanovic, Han, and Koperski [SHK00] studied the problems associated with the design and construction of spatial data cubes. Koperski and Han [KH95] proposed a progressive refinement method for mining spatial association rules. Knorr and Ng [KN96] presented a method for mining aggregate proximity relationships and commonalities in spatial databases. Spatial classification and trend analysis methods have been developed by Ester, Kriegel, Sander, and Xu [EKSX97] and Ester, Frommelt, Kriegel, and Sander [EFKS98]. A two-step method for classification of spatial data was proposed by Koperski, Han, and Stefanovic [KHS98].

Spatial clustering is a highly active area of recent research into geospatial data mining. For a detailed list of references on spatial clustering methods, please see the Bibliographic Notes of Chapter 7. A spatial data mining system prototype, GeoMiner, was developed by Han, Koperski, and Stefanovic [HKS97]. Methods for mining spatiotemporal patterns have been studied by Tsoukatos and Gunopulos [TG01], Hadjieleftheriou, Kollios, Gunopulos, and Tsotras [HKGT03], and Mamoulis, Cao, Kollios, Hadjieleftheriou, et al. [MCK+04]. Mining spatiotemporal information related to moving objects has been studied by Vlachos, Gunopulos, and Kollios [VGK02], and Tao, Faloutsos, Papadias, and Liu [TFPL04]. A bibliography of temporal, spatial, and spatio-temporal data mining research was compiled by Roddick, Hornsby, and Spiliopoulou [RHS01].

Mutlimedia data mining has deep roots in image processing and pattern recognition, which has been studied extensively in computer science, with many textbooks published, such as Gonzalez and Woods [GW02], Russ [Rus02], and Duda, Hart, and Stork [DHS01]. The theory and practice of multimedia database systems have been

introduced in many textbooks and surveys, including Subramanian [Sub98], Yu and Meng [YM97], Perner [Per02], and Mitra and Acharya [MA03]. The IBM QBIC (Query by Image and Video Content) system was introduced by Flickner, Sawhney, Niblack, Ashley, et al. [FSN+95]. Faloutsos and Lin [FL95] developed FastMap, a fast algorithm for indexing, data mining, and visualization of traditional and multimedia datasets. Natsev, Rastogi, and Shim [NRS99] developed WALRUS, a similarity retrieval algorithm for image databases that explores wavelet-based signatures with region-based granularity. Fayyad and Smyth [FS93] developed a classification method to analyze high-resolution radar images for identification of volcanoes on Venus. Fayyad, Djorgovski, and Weir [FDW96] applied decision tree methods to the classification of galaxies, stars, and other stellar objects in the Palomar Observatory Sky Survey (POSS-II) project. Stolorz and Dean [SD96] developed Quakefinder, a data mining system for detecting earthquakes from remote sensing imagery. Zaïane, Han, and Zhu [ZHZ00] proposed a progressive deepening method for mining object and feature associations in large multimedia databases. A multimedia data mining system prototype, MultiMediaMiner, was developed by Zaïane, Han, Li, et al. [ZHL+98] as an extension of the DBMiner system proposed by Han, Fu, Wang, et al. [HFW+96]. An overview of image mining methods is presented by Hsu, Lee, and Zhang [HLZ02].

Text data analysis has been studied extensively in information retrieval, with many good textbooks and survey articles, such as Salton and McGill [SM83], Faloutsos [Fal85], Salton [Sal89], van Rijsbergen [vR90], Yu and Meng [YM97], Raghavan [Rag97], Subramanian [Sub98], Baeza-Yates and Riberio-Neto [BYRN99], Kleinberg and Tomkins [KT99], Berry [Ber03] and Weiss, Indurkhya, Zhang, and Damerau [WIZD04]. The technical linkage between information filtering and information retrieval was addressed by Belkin and Croft [BC92]. The *latent semantic indexing* method for document similarity analysis was developed by Deerwester, Dumais, Furnas, et al. [DDF+90]. The *probabilistic latent semantic analysis* method was introduced to information retrieval by Hofmann [Hof98]. The *locality preserving indexing* method for document representation was developed by He, Cai, Liu, and Ma [HCLM04]. The use of signature files is described in Tsichritzis and Christodoulakis [TC83]. Feldman and Hirsh [FH98] studied methods for mining association rules in text databases. Methods for automated document classification have been studied by many researchers, such as Wang, Zhou, and Liew [WZL99], Nigam, McCallum, Thrun and Mitchell [NMTM00] and Joachims [Joa01]. An overview of text classification is given by Sebastiani [Seb02]. Document clustering by *Probabilistic Latent Semantic Analysis (PLSA)* was introduce by Hofmann [Hof98] and that using *Latent Dirichlet Allocation (LDA)* method was proposed by Blei, Ng, and Jordan [BNJ03]. Using such clustering methods to facilitate comparative analysis of documents was done by Zhai, Velivelli, and Yu [ZVY04]. A comprehensive study of using dimensionality reduction methods for document clustering can be found in Cai, He, and Han [CHH05].

Web mining started in recent years together with the development of Web search engines and Web information service systems. There has been a great deal of work on Web data modeling and Web query systems, such as W3QS by Konopnicki and Shmueli [KS95], WebSQL by Mendelzon, Mihaila, and Milo [MMM97], Lorel by Abitboul, Quass, McHugh, et al. [AQM+97], Weblog by Lakshmanan, Sadri, and Subramanian [LSS96], WebOQL by Arocena and Mendelzon [AM98], and NiagraCQ by Chen, DeWitt, Tian, and Wang [CDTW00]. Florescu, Levy, and Mendelzon [FLM98] presented a comprehensive overview of research on Web databases. An introduction to the semantic Web was presented by Berners-Lee, Hendler, and Lassila [BLHL01].

Chakrabarti [Cha02] presented a comprehensive coverage of data mining for hypertext and the Web. Mining the Web's link structures to recognize authoritative Web pages was introduced by Chakrabarti, Dom, Kumar, et al. [CDK+99] and Kleinberg and Tomkins [KT99]. The HITS algorithm was developed by Kleinberg [Kle99]. The PageRank algorithm was developed by Brin and Page [BP98]. Embley, Jiang and Ng [EJN99] developed some heuristic rules based on the DOM structure to discover record boundaries within a page, which assist data extraction from the Web page. Wong and Fu [WF00] defined tag types for page segmentation and gave a label to each part of the Web page for assisting classification. Chakrabarti et al. [Cha01, CJT01] addressed the fine-grained topic distillation and disaggregated hubs into regions by analyzing DOM structure as well as intrapage text distribution. Lin and Ho [LH02] considered ⟨TABLE⟩ tag and its offspring as a content block and used an entropy-based approach to discover informative ones. Bar-Yossef and Rajagopalan [BYR02] proposed the template detection problem and presented an algorithm based on the DOM structure and the link information. Cai et al. [CYWM03, CHWM04] proposed the Vision-based Page Segmentation algorithm and developed the block-level link analysis techniques. They have also successfully applied the block-level link analysis on Web search [CYWM04] and Web image organizing and mining [CHM+04, CHL+04].

Web page classification was studied by Chakrabarti, Dom, and Indyk [CDI98] and Wang, Zhou, and Liew [WZL99]. A multilayer database approach for constructing a Web warehouse was studied by Zaïane and Han [ZH95]. Web usage mining has been promoted and implemented by many industry firms. Automatic construction of adaptive Web sites based on learning from Weblog user access patterns was proposed by Perkowitz and Etzioni [PE99]. The use of Weblog access patterns for exploring Web usability was studied by Tauscher and Greenberg [TG97]. A research prototype system, WebLogMiner, was reported by Zaïane, Xin, and Han [ZXH98]. Srivastava, Cooley, Deshpande, and Tan [SCDT00] presented a survey of Web usage mining and its applications. Shen, Tan, and Zhai used Weblog search history to facilitate context-sensitive information retrieval and personalized Web search [STZ05].

Bibliography

[AM98] G. Arocena and A. O. Mendelzon. WebOQL: Restructuring documents, databases, and webs. In *Proc. 1998 Int. Conf. Data Engineering (ICDE'98)*, pages 24–33, Orlando, FL, Feb. 1998.

[AQM+97] S. Abitboul, D. Quass, J. McHugh, J. Widom, and J. Wiener. The Lorel query language for semi-structured data. *Int. J. Digital Libraries*, 1:68–88, 1997.

[BC92] N. Belkin and B. Croft. Information filtering and information retrieval: Two sides of the same coin? *Comm. ACM*, 35:29–38, 1992.

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

[BLHL01] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284:34–43, May 2001.

[BNJ03] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *J. Machine Learning Research*, 3:993–1022, 2003.

[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.

[BYR02] Z. Bar-Yossef and S. Rajagopalan. Template detection via data mining and its applications. In *Proc. 2002 Int. World Wide Web Conf. (WWW'02)*, pages 580–591, Honolulu, HI, May 2002.

[BYRN99] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press/Addison-Wesley, 1999.

[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.

[CDTW00] J. Chen, D. DeWitt, F. Tian, and Y. Wang. NiagraCQ: A scalable continuous query system for internet databases. In *Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00)*, pages 379–390, Dallas, TX, May 2000.

[Cha01] S. Chakrabarti. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In *Proc. 2001 Int. World Wide Web Conf. (WWW'01)*, pages 211–220, Hong Kong, China, May 2001.

[Cha02] S. Chakrabarti. *Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data.* Morgan Kaufmann, 2002.

[CHH05] D. Cai, X. He, and J. Han. Document clustering using locality preserving indexing. *IEEE Trans. Knowledge and Data Engineering*, 17:1624–1637, 2005.

[CHL+04] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. Hierarchical clustering of WWW image search results using visual, textual and link analysis. In *Proc. ACM Multimedia 2004*, pages 952–959, New York, NY, Oct. 2004.

[CHM+04] D. Cai, X. He, W.-Y. Ma, J.-R. Wen, and H.-J. Zhang. Organizing WWW images based on the analysis of page layout and web link structure. In *Proc. 2004 IEEE Int. Conf. Multimedia and EXPO (ICME'04)*, pages 113–116, Taipei, Taiwan, June 2004.

[CHWM04] D. Cai, X. He, J.-R. Wen, and W.-Y. Ma. Block-level link analysis. In *Proc. Int. 2004 ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'04)*, pages 440–447, Sheffield, UK, July 2004.

[CJT01] S. Chakrabarti, M. Joshi, and V. Tawde. Enhanced topic distillation using text, markup tags, and hyperlinks. In *Proc. Int. 2001 ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'01)*, pages 208–216, New Orleans, LA, Sept. 2001.

[CYWM03] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Vips: A vision based page segmentation algorithm. In *MSR-TR-2003-79*, Microsoft Research Asia, 2003.

[CYWM04] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Block-based web search. In *Proc. 2004 Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'04)*, pages 456–463, Sheffield, UK, July 2004.

[DDF+90] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. American Society for Information Science*, 41:391–407, 1990.

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

[EFKS98] M. Ester, A. Frommelt, H.-P. Kriegel, and J. Sander. Algorithms for characterization and trend detection in spatial databases. In *Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98)*, pages 44–50, New York, NY, Aug. 1998.

[Ege89] M. J. Egenhofer. *Spatial Query Languages*. UMI Research Press, 1989.

[EJN99] D. W. Embley, Y. Jiang, and Y.-K. Ng. Record-boundary discovery in web documents. In *Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99)*, pages 467–478, Philadelphia, PA, June 1999.

[EKS97] M. Ester, H.-P. Kriegel, and J. Sander. Spatial data mining: A database approach. In *Proc. 1997 Int. Symp. Large Spatial Databases (SSD'97)*, pages 47–66, Berlin, Germany, July 1997.

[EKSX97] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. Density-connected sets and their application for trend detection in spatial databases. In *Proc. 1997 Int. Conf. Knowledge Discovery and Data Mining (KDD'97)*, pages 10–15, Newport Beach, CA, Aug. 1997.

[Fal85] C. Faloutsos. Access methods for text. *ACM Comput. Surv.*, 17:49–74, 1985.

[FDW96] U. M. Fayyad, S. G. Djorgovski, and N. Weir. Automating the analysis and cataloging of sky surveys. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 471–493. AAAI/MIT Press, 1996.

[FH98] R. Feldman and H. Hirsh. Finding associations in collections of text. In R. S. Michalski, I. Bratko, and M. Kubat, editors, *Machine Learning and Data Mining: Methods and Applications*, pages 223–240. John Wiley Sons, 1998.

[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.

[FLM98] D. Florescu, A. Y. Levy, and A. O. Mendelzon. Database techniques for the world-wide web: A survey. *SIGMOD Record*, 27:59–74, 1998.

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

[FSN+95] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, B. Dom, Q. Huang, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, S. Steele, and P. Yanker. Query by image and video content: The QBIC system. *IEEE Computer*, 28:23–32, 1995.

[Gue94] R. H. Gueting. An introduction to spatial database systems. *The VLDB Journal*, 3:357–400, 1994.

[GW02] R. C. Gonzalez and R. E. Woods. *Digital Image Processing* (2nd ed.). Prentice Hall, 2002.

[HCLM04] X. He, D. Cai, H. Liu, and W.-Y. Ma. Locality preserving indexing for document representation. In *Proc. 2004 Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'04)*, pages 96–103, Sheffield, UK, July 2004.

[HFW+96] J. Han, Y. Fu, W. Wang, J. Chiang, W. Gong, K. Koperski, D. Li, Y. Lu, A. Rajan, N. Stefanovic, B. Xia, and O. R. Zaïane. DBMiner: A system for mining knowledge in large relational databases. In *Proc. 1996 Int. Conf. Data Mining and Knowledge Discovery (KDD'96)*, pages 250–255, Portland, OR, Aug. 1996.

[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.

[HKS97] J. Han, K. Koperski, and N. Stefanovic. GeoMiner: A system prototype for spatial data mining. In *Proc. 1997 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'97)*, pages 553–556, Tucson, AZ, May 1997.

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

[HNFD98] J. Han, R. T. Ng, Y. Fu, and S. Dao. Dealing with semantic heterogeneity by generalization-based data mining techniques. In M. P. Papazoglou and G. Schlageter, editors, *Cooperative Information Systems: Current Trends Directions*, pages 207–231. Academic Press, 1998.

[HNKW98] J. Han, S. Nishio, H. Kawano, and W. Wang. Generalization-based data mining in object-oriented databases using an object-cube model. *Data and Knowledge Engineering*, 25:55–97, 1998.

[Hof98] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. 1999 Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'99)*, pages 50–57, Berkeley, CA, Aug. 1998.

[HYK99] J. Han, Q. Yang, and E. Kim. Plan mining by divide-and-conquer. In *Proc. 1999 SIGMOD Workshop Research Issues on Data Mining and Knowledge Discovery (DMKD'99)*, pages 8:1–8:6, Philadelphia, PA, May 1999.

[Joa01] T. Joachims. A statistical learning model of text classification with support vector machines. In *Proc. Int. 2001 ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'01)*, pages 128–136, New Orleans, LA, Sept. 2001.

[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.

[KHS98] K. Koperski, J. Han, and N. Stefanovic. An efficient two-step method for classification of spatial data. In *Proc. 8th Symp. Spatial Data Handling*, pages 45–55, Vancouver, Canada, 1998.

[Kle99] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:604–632, 1999.

[KN96] E. Knorr and R. Ng. Finding aggregate proximity relationships and commonalities in spatial data mining. *IEEE Trans. Knowledge and Data Engineering*, 8:884–897, 1996.

[KS95] D. Konopnicki and O. Shmueli. W3QS: A query system for the world-wide-web. In *Proc. 1995 Int. Conf. Very Large Data Bases (VLDB'95)*, pages 54–65, Zurich, Switzerland, Sept. 1995.

[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.

[LH02] S.-H. Lin and J.-M. Ho. Discovering informative content blocks from web documents. In *Proc. 2002 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'02)*, pages 588–593, Edmonton, Canada, July 2002.

[LHO93] W. Lu, J. Han, and B. C. Ooi. Knowledge discovery in large spatial databases. In *Proc. Far East Workshop Geographic Information Systems*, pages 275–289, Singapore, June 1993.

[LSS96] L. V. S. Lakshmanan, F. Sadri, and S. Subramanian. A declarative query language for querying and restructuring the web. In *Proc. Int. Workshop Research Issues in Data Engineering*, pages 12–21, Tempe, AZ, 1996.

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

[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.

[MGR92] D. J. Maguire, M. Goodchild, and D. W. Rhind. *Geographical Information Systems: Principles and Applications*. Longman, 1992.

[MH01] H. Miller and J. Han. *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, 2001.

[MMM97] A. O. Mendelzon, G. A. Mihaila, and T. Milo. Querying the world-wide web. *Int. J. Digital Libraries*, 1:54–67, 1997.

[NH94] R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. In *Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94)*, pages 144–155, Santiago, Chile, Sept. 1994.

[NMTM00] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39:103–134, 2000.

[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.

[PE99] M. Perkowitz and O. Etzioni. Adaptive web sites: Conceptual cluster mining. In *Proc. 1999 Joint Int. Conf. Artificial Intelligence (IJCAI'99)*, pages 264–269, Stockholm, Sweden, 1999.

[Per02] P. Perner. *Data Mining on Multimedia Data*. Springer Verlag, 2002.

[Rag97] P. Raghavan. Information retrieval algorithms: A survey. In *Proc. 1997 ACM-SIAM Symp. Discrete Algorithms*, pages 11–18, New Orleans, LA, 1997.

[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.

[RSV01] P. Rigaux, M. O. Scholl, and A. Voisard. *Spatial Databases: With Application to GIS*. Morgan Kaufman, 2001.

[Rus02] J. C. Russ. *The Image Processing Handbook* (4th ed.). CRC Press, 2002.

[Sal89] G. Salton. *Automatic Text Processing*. Addison-Wesley, 1989.

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

[SCDT00] J. Srivastava, R. Cooley, M. Deshpande, and P. N. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1:12–23, 2000.

[SCR+99] S. Shekhar, S. Chawla, S. Ravada, A. Fetterer, X. Liu, and C.-T. Lu. Spatial databases—accomplishments and research needs. *IEEE Trans. Knowledge and Data Engineering*, 11:45–55, 1999.

[SD96] P. Stolorz and C. Dean. Quakefinder: A scalable data mining system for detecting earthquakes from space. In *Proc. 1996 Int. Conf. Data Mining and Knowledge Discovery (KDD'96)*, pages 208–213, Portland, OR, Aug. 1996.

[Seb02] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47, 2002.

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

[SM83] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[STZ05] X. Shen, B. Tan, and C. Zhai. Context-sensitive information retrieval with implicit feedback. In *Proc. 2005 Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'05)*, pages 43–50, Salvador, Brazil, Aug. 2005.

[Sub98] V. S. Subrahmanian. *Principles of Multimedia Database Systems*. Morgan Kaufmann, 1998.

[TC83] D. Tsichritzis and S. Christodoulakis. Message files. *ACM Trans. Office Information Systems*, 1:88–98, 1983.

[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.

[TG97] L. Tauscher and S. Greenberg. How people revisit web pages: Empirical findings and implications for the design of history systems. *Int. J. Human Computer Studies, Special issue on World Wide Web Usability*, 47:97–138, 1997.

[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.

[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.

[vR90] C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1990.

[WF00] W. Wong and A. W. Fu. Finding structure and characteristics of web documents for classification. In *Proc. 2000 ACM-SIGMOD Int. Workshop Data Mining and Knowledge Discovery (DMKD'00)*, pages 96–105, Dallas, TX, May 2000.

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

[WZL99] K. Wang, S. Zhou, and S. C. Liew. Building hierarchical classifiers using class proximity. In *Proc. 1999 Int. Conf. Very Large Data Bases (VLDB'99)*, pages 363–374, Edinburgh, UK, Sept. 1999.

[YM97] C. T. Yu and W. Meng. *Principles of Database Query Processing for Advanced Applications*. Morgan Kaufmann, 1997.

[ZCF+97] C. Zaniolo, S. Ceri, C. Faloutsos, R. T. Snodgrass, C. S. Subrahmanian, and R. Zicari. *Advanced Database Systems*. Morgan Kaufmann, 1997.

[ZH95] O. R. Zaïane and J. Han. Resource and knowledge discovery in global information systems: A preliminary design and experiment. In *Proc. 1995 Int. Conf. Knowledge Discovery and Data Mining (KDD'95)*, pages 331–336, Montreal, Canada, Aug. 1995.

[ZHL+98] O. R. Zaïane, J. Han, Z. N. Li, J. Y. Chiang, and S. Chee. MultiMedia-Miner: A system prototype for multimedia data mining. In *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98)*, pages 581–583, Seattle, WA, June 1998.

[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.

[ZLO98] M. J. Zaki, N. Lesh, and M. Ogihara. PLANMINE: Sequence mining for plan failures. In *Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98)*, pages 369–373, New York, NY, Aug. 1998.

[ZTH99] X. Zhou, D. Truffet, and J. Han. Efficient polygon amalgamation methods for spatial OLAP and spatial data mining. In *Proc. 1999 Int. Symp. Large Spatial Databases (SSD'99)*, pages 167–187, Hong Kong, China, July 1999.

[ZVY04] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *Proc. 2004 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'04)*, pages 743–748, Seattle, WA, Aug. 2004.

[ZXH98] O. R. Zaïane, M. Xin, and J. Han. Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs. In *Proc. Advances in Digital Libraries Conf. (ADL'98)*, pages 19–29, Santa Barbara, CA, April 1998.