

Reference

- R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, Automatic subspace clustering of high dimensional data for data mining applications, in Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98), pages 94–105, Seattle,WA, June 1998.
- R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'93), pages 207–216, Washington, DC, May 1993.
- R. Agrawal and R. Srikant, Fast algorithm for mining association rules in large databases, in Research Report RJ 9839, IBM Almaden Research Center, San Jose, CA, June 1994a.
- R. Agrawal and R. Srikant, Fast algorithms for mining association rules, in Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94), pp. 487–499, Santiago, Chile, Sept. 1994b.
- E. Alpaydin, Introduction to Machine Learning (Adaptive Computation and Machine Learning), MIT Press, 2004
- E. Alpaydin, Introduction to Machine Learning (Adaptive Computation and Machine Learning), 2nd. ed., Adaptive Computation and Machine Learning series, The MIT Press, 2009.
- M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander, OPTICS: Ordering points to identify the clustering structure, in Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99), pages 49–60, Philadelphia, PA, June 1999.
- P. Arabie, L. J. Hubert, and G. De Soete. Clustering and Classification. World Scientific, 1996.
- M. Asahara and Y. Matsumoto, Japanese unknown word identification by character-based chunking, Proceedings of the 20th international conference on Computational Linguistics (COLING-2004), Geneva, Switzerland, pp.459–465, August 2004.
- J. L. Balcázar, F. Bonchi, A. Gionis and M. Sebag, Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Lecture Notes in Artificial Intelligence, 2010.
- D. Barbara, W. Du Mouchel, C. Faloutsos, P. J. Haas, J. H. Hellerstein, Y. Ioannidis, H. V., Jagadish, T. Johnson, R. Ng, V. Poosala, K. A. Ross, and K. C. Servcik. The New Jersey data reduction report. Bull. Technical Committee on Data Engineering, 20:3–45, Dec. 1997.
- S. D. Bay, Nearest-neighbor classification from multiple feature subsets, Intelligent Data Analysis 3(3), pp. 191–209, 1999.
- R. J. Bayardo, Efficiently mining long patterns from databases, in Proc. 1998 ACM SIGMOD Int. Conf. Management of Data (SIGMOD'98), pp. 85–93, Seattle, WA, June 1998.
- M. J. A. Berry, and G. Linoff, Data mining techniques for marketing, sales, and customer support. New York: John Wiley, 1997.
- C. M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.

- C. M. Bishop, *Pattern Recognition and Machine*, Springer, 2006.
- A. Blum, Empirical support for winnow and weighted majority algorithms: Results on a calendar scheduling domain, *Machine Learning*, vol.26, no.1, pp.5–23, 1997.
- A. Blum, and T. Mitchell, Combining labeled and unlabeled data with co-training, in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, Madison, WI, San Francisco, Morgan Kaufmann, pp. 92–100, 1998.
- B. Boser, I. Guyon, and V. N. Vapnik, A training algorithm for optimal margin classifiers, In *Proc. Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152, ACM Press: San Mateo, CA, 1992.
- P. Bradley, U. Fayyad, and C. Reina, Scaling clustering algorithms to large databases, in *Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98)*, pp. 9–15, New York, NY, Aug. 1998.
- U. Brefeld, and T. Scheffer, Co-EM support vector learning, in R. Greiner and D. Schuurmans, editors, *Proceedings of the Twenty-First International Conference on Machine Learning*, Banff, Alberta, Canada. New York: ACM, pp. 121–128, 2004.
- L. Breiman, Stacked regression, *Machine Learning* 24(1), pp. 49–64, 1996.
- L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth International Group, 1984.
- L. Breiman and J. Friedman, Predicting multivariate responses in multiple linear regression (with discussion), *J. Roy. Statist. Soc. B.*, Vol. 59, pp. 3-37, 1997.
- S. Brin, R. Motwani, J.D. Ullman, and S. Tsur, Dynamic itemset counting and implication rules for market basket analysis, in *Proc. 1997 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'97)*, pages 255–264, Tucson, AZ, May 1997.
- C. J. C. Burges, A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, Vol. 2, pp. 121–168, 1998.
- F. Camastra and A. Vinciarelli, *Machine Learning for Audio, Image and Video Analysis: Theory and Applications (Advanced Information and Knowledge Processing)*, Springer, 2008.
- J. Cendrowska, PRISM: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies* 27(4):349–370, 1998.
- S. Chakrabarti, *Mining the Web: Analysis of Hypertext and Semi Structured Data*. Morgan Kaufmann, 2002.
- J. S. Chang and K. Su, Automatic construction of a chinese electronic dictionary, *Proceedings of the Third Workshop on Very Large Corpora*, D. Yarovsky and K. Church, Somerset (Ed.), New Jersey, pp.107–120, Association for Computational Linguistics, 1995.
- P. Charoenpornasawat, B. Kijirikul, and S. Meknavin, Feature-based thai unknown word boundary identification using winnow, *Proceedings of the IEEE Asia-Pacific Conference on Circuits and Systems (APCCAS-98)*, Chiang Mai, Thailand, pp.547–550, November 1998.

- A. Chaturvedi, P. Green, and J. Carroll, K-means, k-medians and k-modes: Special cases of partitioning multiway data, in The Classification Society of North America (CSNA) Meeting Presentation, Houston, TX, 1994.
- A. Chaturvedi, P. Green, and J. Carroll, K-modes clustering, *J. Classification*, 18:35–55, 2001.
- P. Cheeseman and J. Stutz, Bayesian classification (AutoClass): Theory and results. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 153–180, AAAI/MIT Press, 1996.
- K. Cheng, G. Young, and K. Wong, A study on wordbased and integral-bit chinese text compression algorithm, *Journal of the American Society for Information Science*, vol.50, no.3, pp.218–228, 1999.
- W.T. Chuang, A. Tiyyagura, J. Yang, G. Giuffrida, A fast algorithm for hierarchical text classification, in: Y. Kambayashi, M. Mohania, A. Tjoa (Eds.), *Proceedings of DaWaK-00, 2nd International Conference on Data Warehousing and Knowledge Discovery*, Springer Verlag, Heidelberg, Germany, London, UK, 2000, pp. 409–418, published in the “Lecture Notes in Computer Science” series, number 1874.
- P. Clark and T. Niblett, The CN2 induction algorithm, *Machine Learning*, 3:261–283, 1989.
- W. Cleveland, *Visualizing Data*. Hobart Press, 1993.
- W. W. Cohen, Fast effective rule induction. In A. Prieditis and S. Russell, editors, *Proceedings of the Twelfth International Conference on Machine Learning*, Tahoe City, CA. San Francisco: Morgan Kaufmann, pp. 115–123, 1995.
- D. J. Cook and L. B. Holder. Graph-based data mining. *IEEE Intelligent Systems*, 15(2), 32-41, 2000.
- T. Cover and P. Hart. Nearest neighbor pattern classification, *IEEE Trans. Information Theory*, 13:21–27, 1967.
- J. Cristianini and N. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.
- B. V. Dasarathy, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, IEEE Computer Society Press, 1991.
- A.DasGupta, *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics*, Springer, 2011.
- M. Dash, H. Liu, and J. Yao. Dimensionality reduction of unsupervised data, in *Proceedings of the 1997 IEEE Int. Conf. Tools with AI (ICTAI'97)*, pages 532–539, IEEE Computer Society, 1997.
- T. Dasu and T. Johnson, *Exploratory Data Mining and Data Cleaning*. JohnWiley & Sons, 2003.
- I. Daubechies, The wavelet transform, time-frequency localization and signal analysis, *IEEE Trans. Inf. Theory*, 36 (5), pp. 961-1005, 1990.
- I. Daubechies. *Ten Lectures on Wavelets*. Capital City Press, 1992.

- W. H. E. Day and H. Edelsbrunner, Efficient algorithms for agglomerative hierarchical clustering methods, *J. Classification*, Vol. 1, pp. 7–24, 1984.
- F. Debole and F. Sebastiani, Supervised term weighting for automated text categorization, in: *Proceedings of SAC-03, 18th ACM Symposium on Applied Computing*, ACM Press, New York, US, Melbourne, US, forthcoming, 2003.
- A. Dempster, N. Laird, and D. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society*, 39:1–38, 1977.
- Z.-H. Deng, S.-W. Tang, D.-Q. Yang, M. Zang, X.-B. Wu and M. Yang, A linear text classification algorithm based on category relevance factors, in: *Proceedings of ICADL-02, 5th International Conference on Asian Digital Libraries*, ACM Press, New York, US, Singapore, 2002, pp. 88–98.
- T. G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization, *Machine Learning* 40(2), pp. 139–158, 2000.
- A. J. Dobson, *An Introduction to Generalized Linear Models* (2nd ed.), Chapman and Hall, 2001.
- P. Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple Bayesian classifier, In *Proc. 1996 Int. Conf. Machine Learning (ML'96)*, pp. 105–112, Bari, Italy, July 1996.
- H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. N. Vapnik, Support vector regression machines, in M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems Vol. 9*, pp. 155–161, MIT Press, 1997.
- R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification* (2nd. ed.), John Wiley and Sons, 2001.
- S. Dzeroski and N. Lavrac, editors. *Relational Data Mining*. Kluwer, Berlin, 2001.
- B. Efron, and R. Tibshirani, *An introduction to the bootstrap*. London: Chapman and Hall, 1993.
- J. P. Egan, *Signal detection theory and ROC analysis*. Series in Cognition and Perception. New York: Academic Press, 1975.
- C. Elkan, Boosting and naïve Bayesian learning, in *Technical Report CS97-557, Dept. Computer Science and Engineering, Univ. Calif. at San Diego*, Sept. 1997.
- M. Ester, H.-P. Kriegel, and X. Xu. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification, in *Proc. 1995 Int. Symp. Large Spatial Databases (SSD'95)*, pages 67–82, Portland, ME, Aug. 1995.
- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, A density-based algorithm for discovering clusters in large spatial databases, in *Proc. 1996 Int. Conf. Knowledge Discovery and Data Mining (KDD'96)*, pages 226–231, Portland, OR, Aug. 1996.
- R. Feldman. Link analysis: Current state of the art. In *KDD-02 Tutorial*, 2002.
- D. Fisher. Improving inference through conceptual clustering, in *Proc. 1987 Nat. Conf. Artificial Intelligence (AAAI'87)*, pages 461–465, Seattle, WA, July 1987.

- E. Fix and J. L. Hodges Jr. Discriminatory analysis, non-parametric discrimination: consistency properties, In Technical Report 21-49-004 (4), USAF School of Aviation Medicine, Randolph Field, Texas, 1951.
- R. Fletcher, Practical Methods of Optimization, John Wiley & Sons, 1987.
- I. K. Fodor, "A survey of dimension reduction techniques," LLNL technical report, June 2002, UCRL-ID-148494.
- Y. Freund, and R. E. Schapire, Experiments with a new boosting algorithm, in L. Saitta, editor, Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Italy. San Francisco, Morgan Kaufmann, pp. 148–156, 1996.
- Y. Freund and R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Computer and System Sciences, 55, pp. 119–139, 1997.
- Y. Freund and R. E. Schapire, A short introduction to boosting, Journal of Japanese Society for Artificial Intelligence, vol.14, no.5, pp.771–780, 1999.
- J. H. Friedman, Greedy function approximation: A gradient boosting machine, Annals of Statistics 29(5), pp. 1189–1232, 2001.
- J. H. Friedman, Another approach to polychotomous classification, Technical Report, Department of Statistics, Stanford University, Stanford, CA, 1996.
- J. H. Friedman, J. L. Bentley, and R. A. Finkel, An algorithm for finding best matches in logarithmic expected time, ACM Transactions on Math Software, 3:209–226, 1977.
- J. H. Friedman, T. Hastie, and R. Tibshirani, Additive logistic regression: A statistical view of boosting, Annals of Statistics 28(2), pp. 337–374, 2000.
- Michael Friendly, Visualizing Categorical Data, SAS Publishing, 2000.
- Ben Fry, Visualizing Data: Exploring and Explaining Data with the Processing Environment, O'Reilly Media, 1st edition, 2008.
- N. Fuhr, Models for retrieval with probabilistic indexing, Information Processing and Management 1 (25) (1989) 55–72.
- K. Fukunaga and D. Hummels, Bayes error estimation using parzen and k-nn procedure, In IEEE Trans. Pattern Analysis and Machine Learning, pages 634–643, 1987.
- J. Fürnkranz, Round robin classification, Journal of Machine Learning Research, Vol. 2, pp. 721–747, 2002.
- J. Fürnkranz, and G. Widmer, Incremental reduced-error pruning. In H. Hirsh and W. Cohen, editors, Proceedings of the Eleventh International Conference on Machine Learning, New Brunswick, NJ. San Francisco: Morgan Kaufmann, pp. 70–77, 1994.
- J. Gama, E. Bradley and J. Hollmen, Advances in Intelligent Data Analysis X: 10th International Symposium, IDA 2011, Porto, Portugal, October 29-31, 2011, Lecture Notes in Artificial Intelligence, 2011.

- M. Ganiz, W.M. Pottenger, and C.D. Janneck, Recent Advances in Literature based Discovery, *Journal of the American Society for Information Science and Technology*, 2006.
- X. Ge, W. Pratt, and P. Smyth, Discovering chinese words from unsegmented text, *Research and Development in Information Retrieval*, pp.271–272, 1999.
- J. Gennari, P. Langley, and D. Fisher, Models of incremental concept formation, *Artificial Intelligence*, 40:11–61, 1989.
- R. Ghani, Combining labeled and unlabeled data for multiclass text categorization, in C. Sammut and A. Hoffmann, editors, *Proceedings of the Nineteenth International Conference on Machine Learning*, Sydney, Australia. San Francisco: Morgan Kaufmann, pp. 187–194, 2002.
- G. Grahne and J. Zhu, Efficiently using prefix-trees in mining frequent itemsets, In *Proc. ICDM'03 Int. Workshop on Frequent Itemset Mining Implementations (FIMI'03)*, Melbourne, FL, Nov. 2003.
- P. Grünwald, *the Minimum Description Length principle*, MIT Press, June 2007.
- S. Guha, R. Rastogi, and K. Shim, CURE: An efficient clustering algorithm for large databases, in *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98)*, pp. 73–84, Seattle,WA, June 1998.
- S. Guha, R. Rastogi, and K. Shim, ROCK: A robust clustering algorithm for categorical attributes, in *Proc. 1999 Int. Conf. Data Engineering (ICDE'99)*, pp. 512–521, Sydney, Australia, Mar. 1999.
- D. Gunopulos, T. Hofmann, D. Malerba and M. Vazirgiannis, *Machine Learning and Knowledge Discovery in Databases, Part II: European Conference, ECML PKDD 2011*, Athens, Greece, September 5-9, 2011, *Lecture Notes in Artificial Intelligence*, 2011.
- A.Haar, "Zur Theorie der orthogonalen Funktionensysteme", *Mathematische Annalen* 69 (3), 1910, pp. 331–371, doi:10.1007/BF01456326.
- S.K. Halgamuge and L. Wang (Eds.), *Classification and Clustering for Knowledge Discovery*, springer, 2005.
- J. Han and Y. Fu, Discovery of multiple-level association rules from large databases, in *Proc. 1995 Int. Conf. Very Large Data Bases (VLDB'95)*, pp. 420–431, Zurich, Switzerland, Sept. 1995.
- J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd edition, The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor, Morgan Kaufmann Publishers, 2004.
- E.-H. Han, G. Karypis, Centroid-based document classification: analysis and experimental results, in: *Principles of Data Mining and Knowledge Discovery*, 2000, pp. 424–431, 2000.
- J. Han, J. Pei, and Y. Yin, Mining frequent patterns without candidate generation, in *Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00)*, pp. 1–12, Dallas, TX, May 2000.
- P. E. Hart. The condensed nearest neighbor rule, *IEEE Transactions on Information Theory*, Vol. 14, pp. 515–516, 1968.
- J. A. Hartigan, *Clustering Algorithms*, John Wiley & Sons, 1975.

- C. Haruechaiyasak, S. Kongyoung, and M.N. Dailey, A comparative study on thai word segmentation approaches, Proceedings of 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2008), Krabi, Thailand, pp.125–128, May 2008.
- C. Haruechaiyasak, C. Sangkeettrakarn, P. Palingoon, S. Kongyoung, and C. Damrongrat, A collaborative framework for collecting thai unknown words from the web, Proceedings of the COLING/ACL on Main conference poster sessions (COLING/ACL-2006), Sydney, Australia, pp.345–352, July 2006.
- T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning, New York: Springer-Verlag, 2001.
- T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd. ed.), Springer, 2009.
- S. S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall, 1999.
- R. Hecht-Nielsen, Neurocomputing, Addison-Wesley, 1990.
- D. Heckerman. Bayesian networks for knowledge discovery. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, pp. 273–305, MIT Press, 1996.
- J. Hertz, A. Krogh, and R. G. Palmer, Introduction to the Theory of Neural Computation, Addison-Wesley, 1991.
- J. Hipp, A. Myka, R. Wirth, and U. Guntzer, A new algorithm for faster mining of generalized association rules”, Proceedings of the 2nd European Conference on Principles of Data Mining and Knowledge Discovery (PKDD '98), (Nantes, France), pp. 74–82, Sep 1998.
- J. Hong, I. Mozetic, and R. S. Michalski. AQ15: Incremental learning of attribute-based descriptions from examples, the method and user's guide, in Report ISG 85-5, UIUCDCSF- 86-949, Dept. Comp. Science, University of Illinois at Urbana-Champaign, 1986.
- Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery, Vol. 2, pp. 283–304, 1998.
- B. B. Hubbard. The World According to Wavelets. A. K. Peters, 1996.
- D.A. Hull, Improving text retrieval for the routing problem using latent semantic indexing, in: W.B. Croft, C.J. van Rijsbergen (Eds.), Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval, Springer Verlag, Heidelberg, Germany, Dublin, Ireland, 1994, pp. 282–289.
- S.-Y. Hwang and E.-P. Lim, A data mining approach to new library book recommendations, Lecture Notes in Computer Science ICADL 2002, (Singapore), pp. 229–240, Dec 2002.
- D. J. Ittner, D.D. Lewis, D.D. Ahn, Text categorization of low quality images, in: Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US, 1995, pp. 301–315.

- A. J. Izenman, *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, Springer, 2008.
- A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, 1988.
- A. K. Jain, M. N. Murty, and P. J. Flynn, Data clustering: A survey, *ACM Comput. Surv.*, 31:264–323, 1999.
- M. James, *Classification Algorithms*, John Wiley & Sons, 1985.
- F. V. Jensen. *An Introduction to Bayesian Networks*. Springer Verlag, 1996.
- D. Jensen and H. Goldberg. *AAAI Fall Symposium on AI and Link Analysis*. AAAI Press, 1998.
- C. Jing-Shin and K. Su, An unsupervised iterative method for chinese new lexicon extraction, *International Journal of Computational Linguistics and Chinese Language Processing*, vol.2, no.2, 1997.
- T. Joachims, A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization, in: D.H. Fisher (Ed.), *Proceedings of ICML-97, 14th International Conference on Machine Learning*, Morgan Kaufmann Publishers, San Francisco, US, Nashville, US, 1997, pp. 143–151.
- G. H. John and P. Langley. Static versus dynamic sampling for data mining. In *Proceedings of 1996 Int. Conf. Knowledge Discovery and Data Mining (KDD'96)*, pages 367–370, Portland, OR, Aug. 1996.
- R. A. Johnson and D. A. Wichern, *Applied Multivariate Statistical Analysis (5th ed.)*, Prentice Hall, 2002.
- G. Karypis, E.-H. Han, and V. Kumar, CHAMELEON: A hierarchical clustering algorithm using dynamic modeling, *COMPUTER*, 32:68–75, 1999.
- L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, 1990.
- A. Kawtrakul, C. Thumkanon, Y. Poovorawan, P. Varasrai, and M. Suktarachan, Automatic thai unknown word recognition, *Proceedings of the 4th Natural Language Processing Pacific Rim Symposium (NLPRS-97)*, Phuket, Thailand, pp.341–346, October 1997.
- V. Kecman, *Learning and Soft Computing*, MIT Press, 2001.
- M.M. Kessler, Bibliographic Coupling between Scientific Papers, *American Documentation*, Vol. 14, pp.10–25, 1963.
- J. Kivinen and H. Mannila. The power of sampling in knowledge discovery. In *Proceedings of the 13th ACM Symp. Principles of Database Systems*, pages 77–85, Minneapolis, MN, May 1994.
- J. M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, *Journal of the ACM*, Vol. 46, No. 5, pp.604–632, 1999.
- Y. Kodratoff and R. S. Michalski, *Machine Learning, An Artificial Intelligence Approach*, Vol. 3. Morgan Kaufmann, 1990.

- R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in Proc. 14th Joint Int. Conf. Artificial Intelligence (IJCAI'95), vol. 2, pages 1137–1143, Montreal, Canada, Aug. 1995.
- R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montreal, Canada. San Francisco, Morgan Kaufmann, pp. 1137–1143, 1995.
- R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.
- K. Laskey and S. Mahoney, Network fragments: Representing knowledge for constructing probabilistic models, In Proc. 13th Annual Conf. on Uncertainty in Artificial Intelligence, pages 334–341, San Francisco, CA, August 1997.
- M. Last, A. Kandel and H. Bunke, *Data Mining In Time Series Databases*, World Scientific, 2004.
- S. L. Lauritzen, The EM algorithm for graphical association models with missing data, *Computational Statistics and Data Analysis*, 19:191–201, 1995.
- M. Law, *Introduction to Support Vector Machines*, Department of Computer Science and Engineering, Michigan State University, 2005
- C. L. Lawson and R. J. Hanson, *Solving least-squares problems*, Philadelphia: SIAM Publications, 1995.
- B. Lee, and Tarng, Y. S., "Application of the discrete wavelet transform to the monitoring of tool failure in end milling using the spindle motor current". *International Journal of Advanced Manufacturing Technology* 15 (4), 1999, 238–243. doi:10.1007/s001700050062.
- V. Lertnattee and T. Theeramunkong, Effect of Term Distributions on Centroid-based Text Categorization, *Information Sciences*, Vol. 158, No. 1, pp. 89–115, Elsevier Science, January 2004a.
- V. Lertnattee and T. Theeramunkong, Multi-dimensional Text Classification for Drug Information, *IEEE Transactions on Information Technology in Biomedicine*, Vol. 8, No. 3, pp. 306–312, IEEE Computer Society Press, September 2004b.
- V. Lertnattee and T. Theeramunkong, Parallel Text Categorization for Multi-Dimensional Data, *Lecture Notes in Computer Science*, LNCS-3283, Springer-Verlag, 2004c, pp. 38–41.
- V. Lertnattee and T. Theeramunkong, Improving Thai educational Web page classification using inverse class frequency, In Proceedings of International Symposium on Communication and Information Technology (ISCIT-2005), 12–14 October 2005, Beijing, pp. 817–820, 2005.
- V. Lertnattee and T. Theeramunkong, Class Normalization in Centroid-based Text Categorization, *Information Sciences*, Elsevier Science, Vol. 176, No. 12, pp. 1712–1738, June 2006a.
- V. Lertnattee and T. Theeramunkong, Effects of Term Distributions on Binary Classification, In Proceedings of the First International Conference on Knowledge, Information and Creativity Support Systems (KICSS 2006), 1–4 August 2006, Ayutthaya, Thailand, 2006b, pp. 147–154.
- V. Lertnattee and T. Theeramunkong, Text Classification for Thai Medicinal Web Pages, *Lecture Notes in Computer Science*, LNCS-4426, Springer-Verlag, pp. 631–638, 2007a.

- V. Lertnattee and T.Theeramunkong, Effects of Term Distributions on Binary Classification, IEICE Transaction on Information and Systems, Vol. E90-D, No.10, The Institute of Electronics, Information and Communication Engineers (IEICE), pp. 1592-1600, 2007b.
- V.Lertnattee and T.Theeramunkong, Improving Thai Academic Web Page Classification Using Inverse Class Frequency and Web Link Information, in Proceedings of the IEEE 22nd International Conference on Advanced Information Networking and Applications (AINA 2008), pp. 1144-1149, 25-28 March 2008, Okinawa, Japan.
- V.Lertnattee, S. Chomya, T. Theeramunkong and V. Sornlertlamvanich, Applying collective intelligence for search improvement on Thai herbal information, in Proceedings of 2009 Ninth IEEE International Conference on Computer and Information Technology, Vol. 2, October 11-October 14, 2009 Xiamen, China, pp.178-183, ISBN: 978-0-7695-3836-5.
- Lexitron version 2.2, 2008, NECTEC, Thai-English, English-Thai Dictionary.
- M. Lima, Visual Complexity: Mapping Patterns of Information, Princeton Architectural Press, 2011.
- P. Limcharoen, C. Nattee, and T. Theeramunkong, Two-phase candidate generation for thai word segmentation using GLR parsing technique, Proceedings of the 3rd International Conference on Knowledge, Information and Creativity Support Systems (KICSS-2008), Hanoi, Vietnam, pp. 98-103, December 2008.
- R. Lindsay and M.D. Gordon, Literature-based Discovery by Lexical Statistics, Journal of the American Society for Information Science, Vol.50, No.7, pp.574-587, 1999.
- G.C. Ling, M. Asahara, and Y. Matsumoto, Chinese unknown word identification using character-based tagging and chunking, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL-2003), Vol.2, Sapporo, Japan, pp.197-200, July 2003.
- J. Liu, Y. Pan, K. Wang, and J. Han, Mining frequent itemsets by opportunistic projection, in Proc. 2002 ACM SIGKDD Int.Conf.KnowledgeDiscovery
- S. P. Lloyd, Least Squares Quantization in PCM. IEEE Trans. Information Theory, Vol. 28, pp. 128-137, 1982, (original version: Technical Report, Bell Labs, 1957.)
- D. Loshin, Enterprise Knowledge Management: TheData Quality Approach.MorganKaufmann, 2001.
- C. L. Lui and F. L. Chung, Discovery of generalized association rules with multiple minimum supports, in Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD '2000), (Lyon, France), pp. 510-515, Sep 2000.
- J. MacQueen, Some methods for classification and analysis ofmultivariate observations, in Proc. 5th Berkeley Symp. Math. Statist. Prob., Vol. 1, pp. 281-297, 1967.
- A.K. McCallum, Bow: A toolkit for statistical lan-guage modeling, text retrieval, classification and clustering, 1996.
- A. Michail, Data mining library reuse patterns using generalized association rules, International Conference on Software Engineering, pp. 167-176, 2000.

- R. S. Michalski, On the quasiminimal solution of the covering problem, in Proceedings of the 5th International Symposium on Information Processing (FCIP-69), volume A3 (Switching Circuits), pp. 125–128, Bled, Yugoslavia, 1969.
- R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, Machine Learning, An Artificial Intelligence Approach, Vol. 1, Morgan Kaufmann, 1983.
- R. S. Michalski, J. G. Carbonell, and T. M. Mitchell. Machine Learning, An Artificial Intelligence Approach, Vol. 2. Morgan Kaufmann, 1986.
- R. S. Michalski, I. Brakto, and M. Kubat. Machine Learning and Data Mining: Methods and Applications. JohnWiley & Sons, 1998.
- R. S. Michalski and R. E. Stepp, Learning from observation: Conceptual clustering, in R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, Machine Learning: An Artificial Intelligence Approach (Vol. 1), Morgan Kaufmann, 1983.
- R. S. Michalski and G. Tecuci. Machine Learning, A Multistrategy Approach, Vol. 4, Morgan Kaufmann, 1994.
- D. Michie, D. J. Spiegelhalter, and C. C. Taylor. Machine Learning, Neural and Statistical, Classification. Ellis Horwood, 1994.
- T. M. Mitchell, Machine Learning. McGraw-Hill, 1997.
- S. K. Murthy, Automatic construction of decision trees from data: A multi-disciplinary survey, Data Mining and Knowledge Discovery, 2:345–389, 1998.
- M. Nagao and S. Mori, A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of japanese, Proceedings of the 15th International Conference on Computational Linguistics (COLING-94), Kyoto, Japan, pp.611–615, August 1994.
- H. Nanba, N. Kando, and M. Okumura, Classification of Research Papers using Citation Links and Citation Types: Towards Automatic Review Article Generation, Proceedings of the American Society for Information Science(ASIS)/ the 11th SIG Classification Research Workshop, Classification for User Support and Learning, Chicago, USA, pp.117–134, Morgan Kaufmann Publishers, San Francisco, US, 2000.
- J. Neter, M. H. Kutner, C. J. Nachtsheim, and L. Wasserman. Applied Linear Statistical Models (4th ed.) Irwin, 1996.
- 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.
- K. Nigam, and R. Ghani, Analyzing the effectiveness and applicability of co-training, Proceedings of the Ninth International Conference on Information and Knowledge Management, McLean, VA. New York, ACM, pp. 86–93, 2000.
- K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell, Text classification from labeled and unlabeled documents using EM, Machine Learning 39(2/3), pp. 103–134, 2000.
- N. J. Nilsson, Principles of Artificial Intelligence, Morgan Kaufmann, 1965.

- N. J. Nilsson, *Learning Machines: Foundations of trainable pattern-classifying systems*, McGraw-Hill, New York, 1965.
- J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer-Verlag, 1999.
- O. Okun, G. Valentini and M. Re, *Ensembles in Machine Learning Applications (Studies in Computational Intelligence)*, Springer, 2011.
- L. Page, S. Brin, R. Motwani, and T. Winograd, *The Pagerank Citation Ranking: Bringing Order to the Web*, tech. rep., Stanford Digital Library Technologies Project, 1998.
- J. S. Park, M. S. Chen, and P. S. Yu, An effective hash-based algorithm for mining association rules, in *Proc. 1995 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'95)*, pp. 175–186, San Jose, CA, May 1995.
- L. Parsons, E. Haque, and H. Liu, Subspace clustering for high dimensional data: A review, *SIGKDD Explorations*, Vol. 6, pp. 90–105, 2004.
- N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, Discovering frequent closed itemsets for association rules, in *Proc. 7th Int. Conf. Database Theory (ICDT'99)*, pages 398–416, Jerusalem, Israel, Jan. 1999.
- J. Pei, J. Han, and R. Mao, CLOSET: An efficient algorithm for mining frequent closed itemsets, in *Proc. 2000 ACM-SIGMOD Int. Workshop Data Mining and Knowledge Discovery (DMKD'00)*, pp. 11–20, Dallas, TX, May 2000.
- J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu, PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth, In *Proc. 2001 Int. Conf. Data Engineering (ICDE'01)*, pages 215–224, Heidelberg, Germany, April 2001.
- 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 prefix-span approach, *IEEE Trans. Knowledge and Data Engineering*, 16:1424–1440, 2004.
- G. Piatetsky-Shapiro. Notes of IJCAI'89 Workshop Knowledge Discovery in Databases (KDD'89). Detroit, MI, July 1989.
- G. Piatetsky-Shapiro and W. J. Frawley. *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991.
- Y. Poowarawan, Dictionary-based thai syllable separation, *Proceedings of the Ninth Electronics Engineering Conference (EECON-86)*, Thailand, pp.409–418, 1986.
- W. Pratt, M. Hearst, and L. Fagan, A Knowledge-based Approach to Organizing Retrieved Documents, in *Proc. of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, Orlando, pp.80–85, 1999.
- F. P. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*, Springer-Verlag, 1985.
- W. H. Press, S. A. Teukolosky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1996.

- F. Provost, and T. Fawcett. 1997. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In D. Heckerman, H. Mannila, D. Pregibon, and R. Uthurusamy, editors, Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, Huntington Beach, CA. Menlo Park, CA: AAAI Press.
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999.
- J. R. Quinlan, Induction of decision trees. *Machine Learning*, 1:81–106, 1986.
- J. R. Quinlan, Simplifying decision trees, *Int. J. Man-Machine Studies*, 27:221–234, 1987.
- J. R. Quinlan, An empirical comparison of genetic and decision-tree classifiers, in Proc. 1988 Int. Conf. Machine Learning (ML'88), pages 135–141, San Mateo, CA, 1988.
- J. R. Quinlan, Unknown attribute values in induction, in Proc. 6th Int. Workshop Machine Learning, pages 164–168, Ithaca, NY, June 1989.
- J. R. Quinlan, Learning logic definitions from relations, *Machine Learning*, 5:139–166, 1990.
- J. R. Quinlan, Learning with continuous classes, in Proc. 1992 Australian Joint Conf. on Artificial Intelligence, pages 343–348, Hobart, Tasmania, 1992.
- J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.
- J. R. Quinlan, Bagging, boosting, and C4.5, in Proc. 1996 Nat. Conf. Artificial Intelligence (AAAI'96), volume 1, pages 725–730, Portland, OR, Aug. 1996.
- J. R. Quinlan and R. M. Cameron-Jones, FOIL: A midterm report, In Proc. 1993 European Conf. Machine Learning, pages 3–20, Vienna, Austria, 1993.
- T. Redman. Data Quality: Management and Technology. Bantam Books, 1992.
- T. Redman, Data Quality: The Field Guide. Digital Press (Elsevier), 2001.
- C. J. van Rijsbergen. Information Retrieval. Butterworth, 1990.
- B. D. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, 1996.
- S. E. Robertson, K. Sparck-Jones, Relevance weighting of search terms, *Journal of American Society for Information Science* 3 (27) (1976) 129–146.
- J. J. Rocchio, Relevance feedback in information retrieval, in: G. Salton (Ed.), *The SMART REtrieval System: Experiments in Automatic Document Processing*, Prentice-Hall, Englewood Cliffs, New Jersey, 1971, pp. 313–323.
- S. Rogers and M. Girolami, *A First Course in Machine Learning*, Chapman & Hall/CRC, 2011.
- E. Rosch, *Principles of Categorization*, pp.27–48, John Wiley & Sons Inc., 1978.
- R. Rousseau and A. Zuccala, A Classification of Author Co-citations: Definitions and Search Strategies, *J. Am. Soc. Inf. Sci. Technol.*, Vol. 55, No. 6, pp. 513–529, 2004.
- S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*, Prentice Hall, 1995.
- G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, *Information Processing and Management* 24 (5), pp. 513–523, 1988.

- G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, Pennsylvania, 1989.
- G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.
- G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, NY, USA, 1986.
- A. Savasere, E. Omiecinski, and S. Navathe, An efficient algorithm for mining association rules in large databases, in *Proc. 1995 Int. Conf. Very Large Data Bases (VLDB'95)*, pp. 432–443, Zurich, Switzerland, Sept. 1995.
- R. E. Schapire, Y. Singer, A. Singhal, Boosting and Rocchio applied to text filtering, in: W.B. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson, J. Zobel (Eds.), *Proceedings of SIGIR- 98, 21st ACM International Conference on Research and Development in Information Retrieval*, ACM Press, New York, US, Melbourne, Australia, 1998, pp. 215–223.
- J. C. Schlimmer and D. Fisher, A case study of incremental concept induction, In *Proc. 1986 Nat. Conf. Artificial Intelligence (AAAI'86)*, pp. 496–501, Philadelphia, PA, 1986.
- B. Schölkopf, P. L. Bartlett, A. Smola, and R. Williamson, Shrinking the tube: a new support vector regression algorithm, in M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems Vol. 11*, pp. 330–336, MIT Press, 1999.
- F. Sebastiani, Machine learning in automated text categorization, *ACM Computing Surveys* 34 (1) (2002) 1–47.
- A. K. Seewald, How to make stacking better and faster while also taking care of an unknown weakness, *Proceedings of the Nineteenth International Conference on Machine Learning*, Sydney, Australia, San Francisco: Morgan Kaufmann, pp. 54–561, 2002.
- W. Siedlecki and J. Sklansky. On automatic feature selection. *Int. J. Pattern Recognition and Artificial Intelligence*, 2:197–220, 1988.
- A. Singhal, G. Salton, C. Buckley, Length normalization in degraded text collections, *Tech. Rep. TR95-1507*, 1995.
- A. Singhal, C. Buckley, M. Mitra, Pivoted document length normalization, in: *Research and Development in Information Retrieval*, 1996, pp. 21–29.
- J. W. Shavlik and T. G. Dietterich. *Readings in Machine Learning*. Morgan Kaufmann, 1990.
- G. Sheikholeslami, S. Chatterjee, and A. Zhang, WaveCluster: A multi-resolution clustering approach for very large spatial databases, in *Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98)*, pages 428–439, New York, NY, Aug. 1998.
- T. Shintani and M. Kitsuregawa, Parallel mining algorithms for generalized association rules with classification hierarchy, *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pp. 25–36, 1998.
- D. B. Skalak, Prototype and feature selection by sampling and random mutation hill climbing algorithms, in: *International Conference on Machine Learning*, pp. 293–301, 1994.

- H. Small, Co-Citation in the Scientific Literature: a New Measure of the Relationship between Documents, *Journal of the American Society for Information Science*, Vol.42, pp.676–684,1973.
- O. Sornil and P. Chaiwanarom, Combining prediction by partial matching and logistic regression for thai word segmentation, *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, Morristown, NJ, USA, p.1208, Association for Computational Linguistics, 2004.
- V. Sornlertlamvanich, T. Potipiti, and T. Charoenporn, Automatic corpus-based thai word extraction with the c4.5 learning algorithm, *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, pp.802–807, 2000.
- V. Sornlertlamvanich and H. Tanaka, The automatic extraction of open compounds from text, *Proceedings of 16th International Conference on Computational Linguistics (COLING-96)*, Vol.2, Copenhagen, Denmark, pp.1143–1146, August 1996.
- R. Srikant and R. Agrawal, Mining generalized association rules, In *Proc. 1995 Int. Conf. Very Large Data Bases (VLDB'95)*, pages 407–419, Zurich, Switzerland, Sept. 1995.
- R. Srikant and R. Agrawal, Mining generalized association rules, *Future Generation Computer Systems*, vol. 13, no. 2–3, pp. 161–180, 1997.
- K. Sriphaew and T. Theeramunkong, A new method for finding generalized frequent itemsets in generalized association rule mining, *Proceedings of the Seventh International Symposium on Computers and Communications (A. Corradi and M. Daneshmand, eds.)*, (Taormina-Giardini Naxos, Italy), pp. 1040–1045, Jul 2002.
- K. Sriphaew and T. Theeramunkong, Mining Generalized Closed Frequent Itemsets of Generalized Association Rules, *Lecture Notes in Computer Science*, LNCS-2773, Springer-Verlag, 2003, pp. 476-484.
- K. Sriphaew and T. Theeramunkong, Fast Algorithms for Mining Generalized Frequent Patterns of Generalized Association Rules, *IEICE Transaction on Information and Systems*, Vol. E87-D, No. 3, pp. 761-770, The Institute of Electronics, Information and Communication Engineers (IEICE), March 2004.
- K. Sriphaew and T. Theeramunkong, Revealing Topic-based Relationship among Documents using Association Rule Mining, *Artificial Intelligence and Applications*, pp.112– 117, 2005.
- K. Sriphaew and T. Theeramunkong, Quality Evaluation for Document Relation Discovery using Citation Information, *IEICE Transaction on Information and Systems*, Vol. E90-D, No.8, pp. 1225-1234, The Institute of Electronics, Information and Communication Engineers (IEICE), August 2007a.
- K. Sriphaew and T. Theeramunkong, Thanaruk, Measuring the Validity of Document Relations Discovered from Frequent Itemset Mining, in *Proceedings of IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)*, Honolulu, Hawaii, April 1-5, pp. 112-117, 2007b.
- M. Stone, Cross-validatory choice and assessment of statistical predictions, *J. Royal Statistical Society*, 36:111–147, 1974.

- K. Su, M. Wu, and J. Chang, A corpus-based approach to automatic compound extraction, Proceedings of the ACL'94, New Mexico State University, pp.242–247, 1994.
- D. Swanson, Fish oil, Raynaud's Syndrome, and Undiscovered Public Knowledge, Perspectives in Biology and Medicine, Vol.30, No.1, pp.7–18, 1986.
- D. Swanson, Medical Literature as a Potential Source of New Knowledge, Bulletin of the Medical Library Association, Vol.78, No.1, pp. 29–37, 1990.
- J. TeCho, C. Nattee, and T. Theeramunkong, A corpus-based approach for keyword identification using supervised learning techniques, Proceedings of the 3rd International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2008), Krabi, Thailand, pp.33–36, May 2008a.
- J. TeCho, C. Nattee, and T. Theeramunkong, A corpus-based approach for automatic thai unknown word recognition using group-based evaluation by ranking, Proceedings of the 3rd International Conference on Knowledge, Information and Creativity Support Systems (KICSS-2008), Hanoi, Vietnam, pp.90–97, December 2008b.
- J. TeCho, C. Nattee, and T. Theeramunkong, A corpus-based approach for automatic thai unknown word recognition using ensemble learning techniques, PAKDD, T. Theeramunkong, B. Kijirikul, N. Cercone, and T.B. Ho (Ed.), Lecture Notes in Computer Science, vol.5476, pp.533–540, Springer, 2009a.
- J. TeCho, C. Nattee, and T. Theeramunkong, A Corpus-based Approach for Automatic Thai Unknown Word Recognition using Ensemble Learning Techniques, IEICE Transaction on Information and Systems, Vol. E92-D, No.12, The Institute of Electronics, Information and Communication Engineers (IEICE), pp. 2321-2333, December 2009b.
- T. Theeramunkong, P. Iamтана-anan, C. Nattee, A. Suriyawongkul, E. Nantajeewarawat, and P. Aimmanee, A framework for constructing a thai medical knowledge base, Proceedings of the 2nd International Conference on Knowledge, Information and Creativity Support Systems (KICSS-2007), JAIST, Ishikawa, Japan, pp.45–50, November 2007.
- T. Theeramunkong and T. Tanhermhong, Pattern-based features vs. statistical-based features in decision trees for word segmentation, IEICE Transactions on Information and Systems, vol.E87-D, no.5, pp.1254–1260, May 2004.
- T. Theeramunkong, S. Usanavasin, T. Machomsomboon, and B. Opananont, Thai word segmentation without a dictionary by using decision trees, Proceedings of the 4th Symposium on Natural Language Processing (SNLP-2000), Chaingmai, Thailand, pp.165–175, May 2000.
- T. Theeramunkong and S. Usanavasin, Non-dictionarybased thai word segmentation using decision trees, HLT '01: Proceedings of the first international conference on Human language technology research, Morristown, NJ, USA, pp.1–5, Association for Computational Linguistics, 2001.
- T. Theeramunkong and V. Lertnattee, Applying Text Classification for Thai Medicinal Web Pages, in Proceedings of the 2007 NSTDA Annual Conference Science (Science and Technology for National Productivity and Happiness) (NAC-2007), NSTDA, Pathumthani, Thailand, March 28-30, 2007.

- S. Theodoridis and K. Koutroumbas, *Pattern Recognition (4th Ed.)*, Academic Press, 2008.
- K. M. Ting, I. H. Witten, Stacking Bagged and Dagged Models. In Proc. Of Fourteenth international Conference on Machine Learning, San Francisco, CA, 367-375, 1997.
- H. Toivonen, Sampling large databases for association rules, In Proc. 1996 Int. Conf. Very Large Data Bases (VLDB'96), pp. 134–145, Bombay, India, Sept. 1996.
- R. Tong (Eds.), *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, ACM Press, New York, US, Berkeley, US, 1999, pp. 42–49.
- P. E. Utgoff, An incremental ID3. In Proc. Fifth Int. Conf. Machine Learning, pp. 107–120, San Mateo, CA, 1988.
- V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.
- V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.
- V. N. Vapnik and A. Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, *Theory of Probability and its Applications* Vol. 16, pp. 264–280, 1971.
- K. Viriyayudhakorn, C. Prayoonsri, C. Silpasuwanchai, C. Nattee, and T. Theeramunkong, A statistical approach to classify nationality of name, *Proceedings of the 7th International Symposium on Natural Language Processing (SNLP-2007)*, Pattaya, Thailand, pp.7–11, December 2007.
- J. Wang, J. Han, and J. Pei, CLOSET+: Searching for the best strategies for mining frequent closed itemsets, in Proc. 2003 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'03), pp. 236–245, Washington, DC, Aug. 2003.
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. *IEEE Transaction on Knowledge and Data Engineering*, 7:623–640, 1995.
- W. Wang, J. Yang, and R. Muntz, STING: A statistical information grid approach to spatial data mining, in Proc. 1997 Int. Conf. Very Large Data Bases (VLDB'97), pp. 186–195, Athens, Greece, Aug. 1997.
- Philip D. Wasserman, *Neural computing: theory and practice*, Van Nostrand Reinhold, 1989.
- S. Weisberg, *Applied Linear Regression*, Wiley, New York, 1980.
- S. M. Weiss and C. A. Kulikowski, *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, 1991.
- H. White and K. McCain, *Bibliometrics, Annual Review on Information Science and Technology*, ed. M. Williams, Amsterdam, Netherlands, pp.119–186, Elsevier Science Publishers, 1989.
- I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, (2nd ed.). Morgan Kaufmann, 2005.
- I. H. Witten, E. Frank and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd Edition, Morgan Kaufmann, Burlington, MA, 3 edition, 2011.

- Y. Yang and X. Liu, A re-examination of text categorization methods, in: M.A. Hearst, F. Gey,
- Y. Yang, An evaluation of statistical approaches to text categorization, *Information Retrieval* 1 (1/2) (1999) 69–90.
- M. J. Zaki, Scalable algorithms for association mining, *IEEE Trans. Knowledge and Data Engineering*, 12, pp. 372–390, 2000.
- M. J. Zaki and C. J. Hsiao, CHARM: An efficient algorithm for closed itemset mining, in *Proc. 2002 SIAM Int. Conf. Data Mining (SDM'02)*, pp. 457–473, Arlington, VA, April 2002.
- J. Zhang, J. Gao, and M. Zhou, Extraction of chinese compound words: an experimental study on a very large corpus, *Proceedings of the second workshop on Chinese language processing*, Morristown, NJ, USA, pp.132–139, Association for Computational Linguistics, 2000.
- T. Zhang, R. Ramakrishnan, and M. Livny, BIRCH: an efficient data clustering method for very large databases, In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'96)*, pages 103–114, Montreal, Canada, June 1996.
- W. Zhang, T. Yoshida, and X. Tang, A study on multiword extraction from chinese documents, *Advanced Web and Network Technologies, and Applications: APWeb 2008 International Workshops: BIDM, IWHDM, and DeWeb Shenyang, China, April 26-28, 2008. Revised Selected Papers*, Berlin, Heidelberg, pp.42–53, Springer-Verlag, 2008.
- C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*, Springer, 2012.