P.O.Box 8718, Beijing 100080, China	Journal of Software, Feb. 2007,18(2):246-258
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2007 by *Journal of Software*

时态数据挖掘的相似性发现技术

潘 定, 沈钧毅

潘 定1,2, 沈钧毅2
1(暨南大学 管理学院,广东 广州 510632)
2(西安交通大学 计算机科学与技术系,陕西 西安 710049)
作者简介: 潘定(1963－),男,江苏宝应人,博士,高级工程师,主要研究领域为数据挖掘,数据 仓库.沈钧毅(1939－),男,教授,博士生导师,CCF高级会员,主要研究领域为数据库理论,数据挖掘.
联系人: 潘 定 Phn: +86-20-85220180, E-mail: pandingcn@gmail.co

Received 2005-06-19; Accepted 2006-01-11

Abstract
Temporal data mining (TDM) has been attracting more and more interest from a vast range of domains, from engineering to finance. Similarity discovery technique concentrates on the evolution and development of data, attempting to discover the similarity regularity of dynamic data evolution. The most significant techniques developed in recent researches to deal with similarity discovery in TDM are analyzed. Firstly, the definitions of three categories of temporal data, time series, event sequence, and transaction sequence are presented, and then the current techniques and methods related to various sequences with similarity measures, representations, searching, and various mining tasks getting involved are classified and discussed. Finally, some future research trends on this area are discussed.

Pan D, Shen JY. Similarity discovery techniques in temporal data mining. *Journal of Software*, 2007,18(2): 246-258.
DOI: 10.1360/jos180246
http://www.jos.org.cn/1000-9825/18/246.htm

摘要
现实世界存在着大量的时态数据,时态数据挖掘(temporal data mining,简称TDM)是近年来学术界关注的一个重要研究课题.相似性发现技术关注数据的发展变化,试图从时态数据中发现事物动态演化的相似性规律.分析和比较了近年来TDM研究中涉及的主要相似性发现技术.首先区分定义了3类时态数据:时间序列、事件序列和交易序列;然后分类并讨论了各种与序列相关的主要方法和技术,涉及相似性度量、序列抽象表示和搜索,以及各类挖掘任务及其算法操作;最后展望进一步研究的方向.

基金项目: Supported by the National Natural Science Foundation of China under Grant Nos.60173058, 70372024 (国家自然科学基金)

References:

[1] Agrawal R, Faloutsos C, Swami A. Efficient similarity search in sequence databases. In: David BL, ed. Proc. of the 4th Int'l Conf. on Foundations of Data Organization and Algorithms, FODO'93. Chicago: Springer-Verlag, 1993. 69-84.

[2] Zeng HQ. Research on mining and similarity searching in time series database [Ph.D. Thesis]. Shanghai: Fudan University, 2003 (in Chinese with English abstract).

[3] Roddick J, Spiliopoulou M. A survey of temporal knowledge discovery paradigms and methods. IEEE Trans. on Knowledge and Data Engineering, 2002,14(4):750-768.

[4] Keogh EJ, Pazzani MJ. Scaling up dynamic time warping to massive dataset. In: Zytkow JM, Rauch J, eds. Proc. of the 3rd European Conf. on Principles of Data Mining and Knowledge Discovery, PKDD'99. Prague: Springer-Verlag, 1999. 1-11.

[5] Yi B, Jagadish H, Faloutsos C. Efficient retrieval of similar time sequences under time warping. In: Sipple RS, ed. Proc. of the 4th Int'l Conf. on Data Engineering, ICDE'98. Orlando: IEEE Computer Society, 1998. 201-208.

[6] Kim SW, Park S, Chu WW. Efficient processing of similarity search under time warping in sequence databases: An index-based approach. Information Systems, 2004,29(5):405-420.

[7] Keogh EJ, Ratanamahatana CA. Exact indexing of dynamic time warping. Knowledge and Information Systems, 2005,7(3): 358-386.

[8] Vlachos M, Hadjieleftheriou M, Gunopulos D, Keogh EJ. Indexing multi-dimensional time-series with support for multiple distance measures. In: Getoor L, Senator TE, eds. Proc. of the 9th ACM SIGKDD 2003. Washington: ACM Press, 2003. 216-225.

[9] Chen L, Ng RT. On the marriage of Lp-norms and edit distance. In: Nascimento MA, -zsu MT, Kossmann D, Miller RJ, Blakeley JA, Schiefer KB, eds. Proc. of the 30th Int'l Conf. on Very Large Data Bases, VLDB 2004. Toronto: Morgan Kaufmann Publishers, 2004. 792-804.

[10] Keogh EJ, Smyth P. A probabilistic approach to fast pattern matching in time series databases. In: Heckerman D, Mannila H, Pregibon D, eds. Proc. of the 3rd Int'l Conf. on Knowledge Discovery and Data Mining. Newport Beach: AAAI Press, 1997. 24-30.

[11] Faloutsos C, Ranganathan M, Manolopoulos Y. Fast subsequence matching in time-series databases. In: Snodgrass RT, Winslett M, eds. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Minneapolis: ACM Press, 1994. 419-429.

[12] Chan K, Fu AW. Efficient time series matching by wavelets. In: Kitsuregawa M, Maciaszek L, Papazoglou M, Pu C, eds. Proc. of the 15th Int'l Conf. on Data Engineering, ICDE'99. Sydney: IEEE Computer Society, 1999. 126-133.

[13] Popivanov I, Miller RJ. Similarity search over time series data using wavelets. In: Agrawal R, Dittrich K, Ngu AH, eds. Proc. of the 18th Int'l Conf. on Data Engineering, ICDE 2002. San Jose: IEEE Computer Society, 2002. 212-221.

[14] Korn F, Jagadish H, Faloutsos C. Efficiently supporting ad hoc queries in large datasets of time sequences. In: Peckham J, ed. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Tucson: ACM Press, 1997. 289-300.

[15] Yi B, Faloutsos C. Fast time sequence indexing for arbitrary Lp norms. In: Abbadi AE, Brodie ML, Chakravarthy S, Dayal U, Kamel N, Schlageter G, Whang KY, eds. Proc. of the 26th Int'l Conf. on Very Large Data Bases, VLDB 2000. Cairo: Morgan Kaufmann Publishers, 2000. 385-394.

[16] Keogh EJ, Chakrabarti K, Pazzani M, Mehrotra S. Dimensionality reduction for fast similarity search in large time series databases. Knowledge and Information Systems, 2000,3(3):263-286.

[17] Keogh EJ, Chakrabarti K, Mehrotra S, Pazzani MJ. Locally adaptive dimensionality reduction for indexing large time series databases. In: Aref WG, ed. Proc. of the SIGMOD Int'l Conf. on Management of Data. Santa Barbara: ACM Press, 2001. 151-162.

[18] Perng CS, Wang H, Zhang S, Parker DS. Landmark: A new model for similarity-based pattern querying in time series database. In: Young DC, ed. Proc. of the 16th Int'l Conf. on Data Engineering, ICDE 2000. San Diego: IEEE Computer Society, 2000. 33-42.

[19] Das G, Lin K, Mannila H, Renganathan G, Smyth P. Rule discovery from time series. In: Agrawal R, Stolorz PE, Piatetsky-Shapiro G, eds. Proc of the 4th Int'l Conf. on Knowledge Discovery and Data Mining, KDD'98. New York: AAAI Press, 1998. 16-22.

[20] Huang Y, Yu PS. Adaptive query processing for time-series data. In: Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Diego: ACM Press, 1999. 282-286.

[21] Giles CL, Lawrence S, Tsoi A. Noisy time series prediction using recurrent neural networks and grammatical inference. Machine Learning, 2001,44(1):161-184.

[22] Lin J, Keogh EJ, Lonardi S, Chiu BY. A symbolic representation of time series, with implications for streaming algorithms. In: Zaki MJ, Aggarwal CC, eds. Proc. of the 8th SIGMOD Workshop on DMKD 2003. San Diego: ACM Press, 2003. 2-11.

[23] Bagnall AJ, Janacek GJ. Clustering time series from ARMA models with clipped data. In: Kim W, Kohavi R, Gehrke J, DuMouchel W, eds. Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Seattle: ACM Press, 2004. 49-58.

[24] Gaffney SJ, Smyth P. Curve clustering with random effects regression mixtures. In: Bishop CM, Frey BJ, eds. Proc. of the Workshop on Artificial Intelligence and Statistics. Florida: Society for Artificial Intelligence and Statistics, 2003.

[25] Xiong Y, Yeung DY. Time series clustering with ARMA mixtures. Pattern Recognition, 2004,37(8):1675-1689.

[26] Kalpakis K, Gada D, Puttagunta V. Distance measures for effective clustering of ARIMA time-series. In: Cercone N, Lin TY, Wu X, eds. Proc. of the IEEE Int'l Conf. on Data Mining, ICDM 2001. San Jose: IEEE Computer Society, 2001. 273-280.

[27] Ramoni M, Sebastiani P, Cohen P. Bayesian clustering by dynamics. Machine Learning, 2002,47(1):91-121.

[28] Smyth P. Clustering sequences with hidden Markov models. In: Mozer M, Jordan MI, Petsche T, eds. Proc. of the Advances in Neural Information Processing Systems 9, NIPS'96. Cambridge: MIT Press, 1997. 648-654.

[29] Oates T, Firoiu L, Cohen PR. Using dynamic time warping to bootstrap HMM-based clustering of time series. In: Sun R, Giles CL, eds. Sequence Learning-Paradigms, Algorithms, and Applications. LNAI 1828, Heidelberg: Springer-Verlag, 2001. 35-52.

[30] Ge X, Smyth P. Deformable Markov model templates for time-series pattern matching. In: Proc. of the 6th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000. 81-90.

[31] Mannila H, Toivonen H, Verkamo AI. Discovery of frequent episode in event sequences. Data Mining and Knowledge Discovery, 1997,1 (3):259-289.

[32] Mannila H, Meek C. Global partial orders from sequential data. In: Proc. of the 6th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000. 161-168.

[33] Oliveira AL, Silva JPM. Efficient algorithms for the inference of minimum size DFAs. Machine Learning, 2001,44(7):93-119.

[34] Nakamura K, Matsumoto M. Incremental learning of context free grammars. In: Adriaans P, Fernau H, van Zaanen M, eds. Proc. of the 6th Int'l Colloquium Grammatical Inference. ICGI 2002, Amsterdam: Springer-Verlag, 2002. 174-184.

[35] Smyth P. Probabilistic model-based clustering of multivariate and sequential data. In: Heckerman D, Whittaker J, eds. Proc. of the 7th Int'l Workshop on AI and Statistics. Los Gatos: Morgan Kaufmann Publishers, 1999. 299-304.

[36] Park S, Chu WW, Yoon J, Won J. Similarity search of time-warped subsequences via a suffix tree. Information Systems, 2003, 28 (7):867-883.

[37] Keogh EJ, Pazzani MJ. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In: Agrawal R, Stolorz PE, eds. Proc. of the 4th Int'l Conf. on KDD. New York: AAAI Press, 1998. 239-241.

[38] Keogh EJ, Pazzani MJ. An indexing scheme for fast similarity search in large time series databases. In: -zsoyoglu ZM, ed. Proc. of the 11th Int'l Conf. on Scientific and Statistical Database Management. Cleveland: IEEE Computer Society, 1999. 56-67.

[39] Morinaka Y, Yoshikawa M, Amagasa T, Uemura S. The L-index: An indexing structure for efficient subsequence matching in time sequence databases. In: Industrial Track and Workshops Proc. of the 5th PAKDD 2001. Hong Kong, 2001. 51-60.

[40] Povinelli RJ, Feng X. A new temporal pattern identification method for characterization and prediction of complex time series events. IEEE Trans. on Knowledge and Data Engineering, 2003,15(2):339-352.

[41] Agrawal R, Srikant R. Mining sequential patterns. In: Yu PS, Chen ALP, eds. Proc. of the 11th Int'l Conf. on Data Engineering, ICDE'95. Taipei: IEEE Computer Society, 1995. 3-14.

[42] Srikant R, Agrawal R. Mining sequential patterns: Generalizations and performance improvements. In: Apers PMG, Bouzeghoub M, Gardarin G, eds. Proc. of the 5th Int'l Conf. on Extending Database Technology, EDBT'96. Avignon: Springer-Verlag, 1996. 3-17.

[43] Han J, Pei J, Yin Y, Mao R. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery, 2004,8(1):53-87.

[44] Rainsford CP, Roddick JF. Adding temporal semantics to association rules. In: Zytkow JM, Rauch J, eds. Proc. of the 3rd European Conf. on Principles of Data Mining and Knowledge Discovery, PKDD'99. Prague: Springer-Verlag, 1999. 504-509.

[45] Chen X, Petrounias I. Mining temporal features in association rules. In: Zytkow JM, Rauch J, eds. Proc. of the 3rd European Conf. on Principles of Data Mining and Knowledge Discovery, PKDD'99. Prague: Springer-Verlag, 1999. 295-300.

[46] Ale JM, Rossi GH. An approach to discovering temporal association rules. In: Carroll J, Damiani E, Haddad H, Oppenheim D, eds. Proc. of the 2000 ACM Symp. on Applied Computing. New York: ACM Press, 2000. 294-300.

[47] -zden B, Ramaswamy S, Silberschatz A. Cyclic association rules. In: Sipple RS, ed. Proc. of the 14th Int'l Conf. on Data Engineering, ICDE'98. Orlando: IEEE Computer Society, 1998. 412-421.

[48] Ramaswamy S, Mahajan S, Silberschatz A. On the discovery of interesting patterns in association rules. In: Gupta A, Shmueli O, Widom J, eds. Proc. of the 24th Int'l Conf. on Very Large Data Bases, VLDB'98. New York: Morgan Kaufmann Publishers, 1998. 368-379.

[49] Li Y, Ning P, Wang XS, Jajodia S. Discovering calendar-based temporal association rules. Data and Knowledge Engineering, 2003, 44 (2):193-218.

[50] Han J, Pei J, Yan X. From sequential pattern mining to structured pattern mining: A pattern-growth approach. Journal of Computer Science and Technology, 2004,19(3):257-279.
[51] Zaki MJ. SPADE: An efficient algorithm for mining frequent sequences. Machine Learning, 2001,42(1-2):31-60.

[52] Harms SK, Deogun JS. Sequential association rule mining with time lags. Journal of Intelligent Information Systems, 2004,22(1): 7-22.

[53] Han JW, Pei J, Mortazavi-Asl B, Chen QM, Dayal U, Hsu MC. FreeSpan: Frequent pattern-projected sequential pattern mining. In: Proc. of the 6th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000. 355-359.

[54] Pei J, Han JW, Mortazavi-Asl B,Pinto H. PrefixSpan: Mining sequential patterns efficiently by prefix projected pattern growth. In: Young DC, ed. Proc. of the 17th Int'l Conf. on Data Engineering, ICDE 2001. Heidelberg: IEEE Computer Society, 2001. 215-226.

[55] Garofalakis M, Rastogi R, Shim K. Mining sequential patterns with regular expression constraints. IEEE Trans. on Knowledge and Data Engineering, 2002,14(3):530-552.

[56] Antunes CM, Oliveira AL. Inference of sequential association rules guided by context-free grammars. In: Adriaans P, Fernau H, van Zaanen M, eds. Proc. of the 6th Int'l Colloquium Grammatical Inference. ICGI 2002, Amsterdam: Springer-Verlag, 2002. 1-13.

[57] Pei J, Han J, Wang W. Mining sequential patterns with constraints in large databases. In: Proc. of the 2002 Int'l Conf. on Information and Knowledge Management, CIKM 2002. McLean: ACM Press, 2002. 18-25.

[58] Geurts P. Pattern extraction for time series classification. In: Raedt L, Siebes A, eds. Proc. of the 5th European Conf. on Principles of Data Mining and Knowledge Discovery, PKDD 2001. Freiburg: Springer-Verlag, 2001. 115-127.

[59] Povinelli RJ, Johnson MT, Lindgren AC, Ye J. Time series classification using Gaussian mixture models of reconstructed phase spaces. IEEE Trans. on Knowledge and Data Engineering, 2004,16(6):779-783.

[60] Lesh N, Zaki MJ, Ogihara M. Mining features for sequence classification. In: Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Diego: ACM Press, 1999. 342-346.

[61] Lin J, Vlachos M, Keogh EJ, Gunopulos D. Iterative incremental clustering of time series. In: Bertino E, Christodoulakis S, eds. Proc. of the 9th Int'l Conf. on Extending Database Technology, EDBT 2004. Crete: Springer-Verlag, 2004. 106-122.

[62] Yang J, Wang W. CLUSEQ: Efficient and effective sequence clustering. In: Dayal U, Ramamritham K, Vijayaraman TM, eds. Proc. of the 19th Int'l Conf. on Data Engineering, ICDE 2003. Bangalore: IEEE Computer Society, 2003. 101-112.

[63] Morzy T, Wojciechowski M, Zakrzewicz M. Scalable hierarchical clustering method for sequences of categorical values. In: Cheung DW, ed. Proc. of the 5th Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Hong Kong: Springer-Verlag, 2001. 282-293.

[64] Wang K, Xu C, Liu B. Clustering transactions using large items. In: Proc. of the 1999 Int'l Conf. on Information and Knowledge Management, CIKM'99. Kansas: ACM Press, 1999. 483-490.

[65] Yang Y, Guan X, You J. CLOPE: A fast and effective clustering algorithm for transactional data. In: Hand D, Keim D, Ng R, Za-ane OR, Goebel R, eds. Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Edmonton: ACM Press, 2002. 682-687.

[66] Cheung DW, Han J, Ng VT, Wong CY. Maintenance of discovered association rules in large databases. In: Su SYW, ed. Proc. of the 12th Int'l Conf. on Data Engineering, ICDE'96. New Orleans: IEEE Computer Society, 1996. 106-114.

[67] Parthasarathy S, Zaki MJ, Ogihara M, Dwarkadas S. Incremental and interactive sequence mining. In: Proc. of the 1999 Int'l Conf. on Information and Knowledge Management, CIKM'99. Kansas: ACM Press, 1999. 251-258.

[68] Masseglia F, Poncelet P, Teisseire M. Incremental mining of sequential patterns in large databases. Data and Knowledge Engineering, 2003,46(1):97-121.

[69] Cheng H, Yan X, Han J. IncSpan: Incremental mining of sequential patterns in large database. In: Kim W, Kohavi R, Gehrke J, eds. Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Seattle: ACM Press, 2004. 527-532.

[70] Spiliopoulou M, Roddick JF. Higher order mining: Modeling and mining the results of knowledge discovery. In: Ebecken N, Brebbia CA, eds. Proc. of the 2nd Int'l Conf. on Data Mining Methods and Databases. Cambridge: WIT Press, 2000. 309-320.

[71] Cotofrei P, Stoffel K. From temporal rules to temporal meta-rules. In: Kambayashi Y, Mohania MK, W-- B, eds. Proc. of the 6th Int'l Conf. Data Warehousing and Knowledge Discovery, DaWaK 2004. Zaragoza: Springer-Verlag, 2004. 169-178.

[72] Keogh EJ, Kasetty S. On the need for time series data mining benchmarks: A survey and empirical demonstration. Data Mining and Knowledge Discovery, 2003,7(4):349-371.

附中文参考文献:
[2] 曾海权.时间序列挖掘与相似性查找技术研究[博士学位论文].上海:复旦大学,2003.