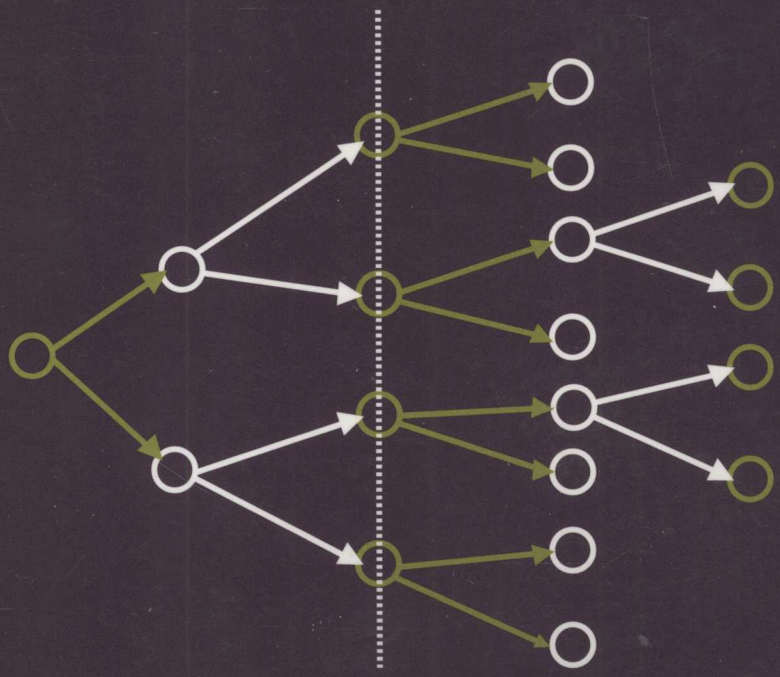
PROCEEDINGS OF THE FOURTH SIAM INTERNATIONAL CONFERENCE ON DATA MINING

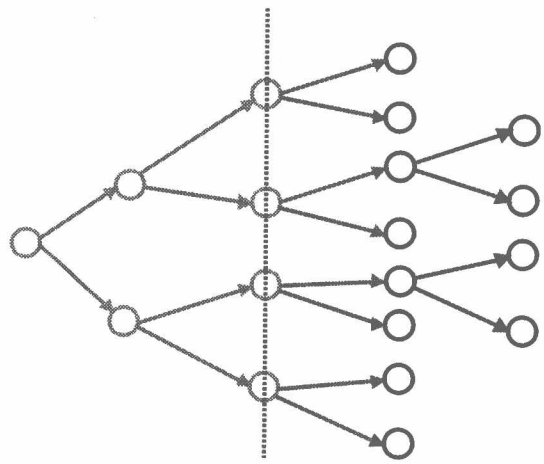


Edited by Michael W. Berry
Umeshwar Dayal
Chandrika Kamath
David Skillicorn

Lake Buena Vista, Florida 2004

siam

PROCEEDINGS OF THE FOURTH SIAM INTERNATIONAL CONFERENCE ON DATA MINING



Edited by
Michael W. Berry
University of Tennessee
Knoxville, Tennessee

Chandrika Kamath
Lawrence Livermore National Laboratory
Livermore, California

Umeshwar Dayal
Hewlett-Packard Corporation
Palo Alto, California

David Skillicorn
Queens University
Kingston, Ontario Canada

siam
Society for Industrial and Applied Mathematics
Philadelphia

PROCEEDINGS OF THE FOURTH SIAM INTERNATIONAL CONFERENCE ON DATA MINING

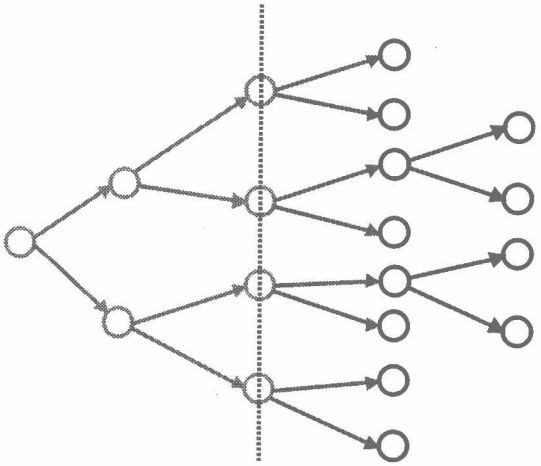
Proceedings of the Fourth SIAM International Conference on Data Mining, Lake Buena Vista, FL, April 22–24, 2004

10 9 8 7 6 5 4 3 2 1

Library of Congress Catalog Card Number: 2004102812

ISBN 0-89871-568-7

PROCEEDINGS OF THE FOURTH SIAM INTERNATIONAL CONFERENCE ON DATA MINING



SIAM PROCEEDINGS SERIES LIST

Glowinski, R., Golub, G. H., Meurant, G. A., and Periaux, J., *First International Conference on Domain Decomposition Methods for Partial Differential Equations* (1988)

Salam, Fathi M. A. and Levi, Mark L., *Dynamical Systems Approaches to Nonlinear Problems in Systems and Circuits* (1988)

Datta, B., Johnson, C., Kaashoek, M., Plemmons, R., and Sontag, E., *Linear Algebra in Signals, Systems and Control* (1988)

Ringeisen, Richard D. and Roberts, Fred S., *Applications of Discrete Mathematics* (1988)

McKenna, James and Temam, Roger, *ICIAM '87: Proceedings of the First International Conference on Industrial and Applied Mathematics* (1988)

Rodrigue, Garry, *Parallel Processing for Scientific Computing* (1989)

Caflish, Russel E., *Mathematical Aspects of Vortex Dynamics* (1989)

Wouk, Arthur, *Parallel Processing and Medium-Scale Multiprocessors* (1989)

Flaherty, Joseph E., Paslow, Pamela J., Shephard, Mark S., and Vasilakis, John D., *Adaptive Methods for Partial Differential Equations* (1989)

Kohn, Robert V. and Milton, Graeme W., *Random Media and Composites* (1989)

Mandel, Jan, McCormick, S. F., Dendy, J. E., Jr., Farhat, Charbel, Lonsdale, Guy, Parter, Seymour V., Ruge, John W., and Stüben, Klaus, *Proceedings of the Fourth Copper Mountain Conference on Multigrid Methods* (1989)

Colton, David, Ewing, Richard, and Rundell, William, *Inverse Problems in Partial Differential Equations* (1990)

Chan, Tony F., Glowinski, Roland, Periaux, Jacques, and Widlund, Olof B., *Third International Symposium on Domain Decomposition Methods for Partial Differential Equations* (1990)

Dongarra, Jack, Messina, Paul, Sorensen, Danny C., and Voigt, Robert G., *Proceedings of the Fourth SIAM Conference on Parallel Processing for Scientific Computing* (1990)

Glowinski, Roland and Lichnewsky, Alain, *Computing Methods in Applied Sciences and Engineering* (1990)

Coleman, Thomas F. and Li, Yuying, *Large-Scale Numerical Optimization* (1990)

Aggarwal, Alok, Borodin, Allan, Gabow, Harold, N., Galil, Zvi, Karp, Richard M., Kleitman, Daniel J., Odlyzko, Andrew M., Pulleyblank, William R., Tardos, Éva, and Vishkin, Uzi, *Proceedings of the Second Annual ACM-SIAM Symposium on Discrete Algorithms* (1990)

Cohen, Gary, Halpern, Laurence, and Joly, Patrick, *Mathematical and Numerical Aspects of Wave Propagation Phenomena* (1991)

Gómez, S., Hennart, J. P., and Tapia, R. A., *Advances in Numerical Partial Differential Equations and Optimization: Proceedings of the Fifth Mexico-United States Workshop* (1991)

Glowinski, Roland, Kuznetsov, Yuri A., Meurant, Gérard, Périaux, Jacques, and Widlund, Olof B., *Fourth International Symposium on Domain Decomposition Methods for Partial Differential Equations* (1991)

Alavi, Y., Chung, F. R. K., Graham, R. L., and Hsu, D. F., *Graph Theory, Combinatorics, Algorithms, and Applications* (1991)

Wu, Julian J., Ting, T. C. T., and Barnett, David M., *Modern Theory of Anisotropic Elasticity and Applications* (1991)

Shearer, Michael, *Viscous Profiles and Numerical Methods for Shock Waves* (1991)

Griewank, Andreas and Corliss, George F., *Automatic Differentiation of Algorithms: Theory, Implementation, and Application* (1991)

Frederickson, Greg, Graham, Ron, Hochbaum, Dorit S., Johnson, Ellis, Kosaraju, S. Rao, Luby, Michael, Megiddo, Nimrod, Schieber, Baruch, Vaidya, Pravin, and Yao, Frances, *Proceedings of the Third Annual ACM-SIAM Symposium on Discrete Algorithms* (1992)

Field, David A. and Komkov, Vadim, *Theoretical Aspects of Industrial Design* (1992)

Field, David A. and Komkov, Vadim, *Geometric Aspects of Industrial Design* (1992)

Bednar, J. Bee, Lines, L. R., Stolt, R. H., and Weglein, A. B., *Geophysical Inversion* (1992)

O'Malley, Robert E. Jr., *ICIAM 91: Proceedings of the Second International Conference on Industrial and Applied Mathematics* (1992)

Keyes, David E., Chan, Tony F., Meurant, Gérard, Scroggs, Jeffrey S., and Voigt, Robert G., *Fifth International Symposium on Domain Decomposition Methods for Partial Differential Equations* (1992)

Dongarra, Jack, Messina, Paul, Kennedy, Ken, Sorensen, Danny C., and Voigt, Robert G., *Proceedings of the Fifth SIAM Conference on Parallel Processing for Scientific Computing* (1992)

Corones, James P., Kristensson, Gerhard, Nelson, Paul, and Seth, Daniel L., *Invariant Imbedding and Inverse Problems* (1992)

Ramachandran, Vijaya, Bentley, Jon, Cole, Richard, Cunningham, William H., Guibas, Leo, King, Valerie, Lawler, Eugene, Lenstra, Arjen, Mulmuley, Ketan, Sleator, Daniel D., and Yannakakis, Mihalis, *Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms* (1993)

Kleinman, Ralph, Angell, Thomas, Colton, David, Santosa, Fadil, and Stakgold, Ivar, *Second International Conference on Mathematical and Numerical Aspects of Wave Propagation* (1993)

Banks, H. T., Fabiano, R. H., and Ito, K., *Identification and Control in Systems Governed by Partial Differential Equations* (1993)

Sleator, Daniel D., Bern, Marshall W., Clarkson, Kenneth L., Cook, William J., Karlin, Anna, Klein, Philip N., Lagarias, Jeffrey C., Lawler, Eugene L., Maggs, Bruce, Milenkovic, Victor J., and Winkler, Peter, *Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms* (1994)

Lewis, John G., *Proceedings of the Fifth SIAM Conference on Applied Linear Algebra* (1994)

Brown, J. David, Chu, Moody T., Ellison, Donald C., and Plemmons, Robert J., *Proceedings of the Cornelius Lanczos International Centenary Conference* (1994)

Dongarra, Jack J. and Tourancheau, B., *Proceedings of the Second Workshop on Environments and Tools for Parallel Scientific Computing* (1994)

Bailey, David H., Bjørstad, Petter E., Gilbert, John R., Mascagni, Michael V., Schreiber, Robert S., Simon, Horst D., Torczon, Virginia J., and Watson, Layne T., *Proceedings of the Seventh SIAM Conference on Parallel Processing for Scientific Computing* (1995)

Clarkson, Kenneth, Agarwal, Pankaj K., Atallah, Mikhail, Frieze, Alan, Goldberg, Andrew, Karloff, Howard, Manber, Udi, Munro, Ian, Raghavan, Prabhakar, Schmidt, Jeanette, and Young, Moti, *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms* (1995)

Becache, Elaine, Cohen, Gary, Joly, Patrick, and Roberts, Jean E., *Third International Conference on Mathematical and Numerical Aspects of Wave Propagation* (1995)

Engl, Heinz W., and Rundell, W., *GAMM–SIAM Proceedings on Inverse Problems in Diffusion Processes* (1995)

Angell, T. S., Cook, Pamela L., Kleinman, R. E., and Olmstead, W. E., *Nonlinear Problems in Applied Mathematics* (1995)

Tardos, Éva, Applegate, David, Canny, John, Eppstein, David, Galil, Zvi, Karger, David R., Karlin, Anna R., Linial, Nati, Rao, Satish B., Vitter, Jeffrey S., and Winkler, Peter M., *Proceedings of the Seventh Annual ACM-SIAM Symposium on Discrete Algorithms* (1996)

Cook, Pamela L., Roytburd, Victor, and Tulin, Marshal, *Mathematics Is for Solving Problems* (1996)

Adams, Loyce and Nazareth, J. L., *Linear and Nonlinear Conjugate Gradient-Related Methods* (1996)

Renardy, Yuriko Y., Coward, Adrian V., Papageorgiou, Demetrios T., and Sun, Shu-Ming, *Advances in Multi-Fluid Flows* (1996)

Berz, Martin, Bischof, Christian, Corliss, George, and Griewank, Andreas, *Computational Differentiation: Techniques, Applications, and Tools* (1996)

Delic, George and Wheeler, Mary F., *Next Generation Environmental Models and Computational Methods* (1997)

Engl, Heinz W., Louis, Alfred, and Rundell, William, *Inverse Problems in Geophysical Applications* (1997)

Saks, Michael, Anderson, Richard, Bach, Eric, Berger, Bonnie, Blum, Avrim, Chazelle, Bernard, Edelsbrunner, Herbert, Henzinger, Monika, Johnson, David, Kannan, Sampath, Khuller, Samir, Maggs, Bruce, Muthukrishnan, S., Ruskey, Frank, Seymour, Paul, Spencer, Joel, Williamson, David P., and Williamson, Gill, *Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms* (1997)

Alexandrov, Natalia M. and Hussaini, M. Y., *Multidisciplinary Design Optimization: State of the Art* (1997)

Van Huffel, Sabine, *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling* (1997)

Ferris, Michael C. and Pang, Jong-Shi, *Complementarity and Variational Problems: State of the Art* (1997)

Bern, Marshall, Fiat, Amos, Goldberg, Andrew, Kannan, Sampath, Karloff, Howard, Kenyon, Claire, Kierstead, Hal, Kosaraju, Rao, Linial, Nati, Rabani, Yuval, Rödl, Vojta, Sharir, Micha, Shmoys, David, Spielman, Dan, Spinrad, Jerry, Srinivasan, Aravind, and Sudan, Madhu, *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* (1998)

DeSanto, John A., *Mathematical and Numerical Aspects of Wave Propagation* (1998)

Tarjan, Robert E., Warnow, Tandy, Amenta, Nina, Benham, Craig, Corneil, Derek G., Edelsbrunner, Herbert, Feigenbaum, Joan, Gusfield, Dan, Habib, Michel, Hall, Leslie, Karp, Richard, King, Valerie, Koller, Daphne, McKay, Brendan, Moret, Bernard, Muthukrishnan, S., Phillips, Cindy, Raghavan, Prabhakar, Randall, Dana, and Scheinerman, Edward, *Proceedings of the Tenth ACM-SIAM Symposium on Discrete Algorithms* (1999)

Hendrickson, Bruce, Yelick, Katherine A., Bischof, Christian H., Duff, Iain S., Edelman, Alan S., Geist, George A., Heath, Michael T., Heroux, Michael H., Koelbel, Chuck, Schrieber, Robert S., Sincovec, Richard F., and Wheeler, Mary F., *Proceedings of the Ninth SIAM Conference on Parallel Processing for Scientific Computing* (1999)

Henderson, Michael E., Anderson, Christopher R., and Lyons, Stephen L., *Object Oriented Methods for Interoperable Scientific and Engineering Computing* (1999)

Shmoys, David, Brightwell, Graham, Cohen, Edith, Cook, Bill, Eppstein, David, Gerards, Bert, Irani, Sandy, Kenyon, Claire, Ostrovsky, Rafail, Peleg, David, Pevzner, Pavel, Reed, Bruce, Stein, Cliff, Tetali, Prasad, and Welsh, Dominic, *Proceedings of the Eleventh ACM-SIAM Symposium on Discrete Algorithms* (2000)

Bermúdez, Alfredo, Gómez, Dolores, Hazard, Christophe, Joly, Patrick, and Roberts, Jean E., *Fifth International Conference on Mathematical and Numerical Aspects of Wave Propagation* (2000)

Kosaraju, S. Rao, Bellare, Mihir, Buchsbaum, Adam, Chazelle, Bernard, Graham, Fan Chung, Karp, Richard, Lovász, László, Motwani, Rajeev, Myrvold, Wendy, Pruhs, Kirk, Sinclair, Alistair, Spencer, Joel,Stein, Cliff, Tardos, Eva, Vempala, Santosh, *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms* (2001)

Koelbel, Charles and Meza, Juan, *Proceedings of the Tenth SIAM Conference on Parallel Processing for Scientific Computing* (2001)

Grossman, Robert, Kumar, Vipin, and Han, Jiawei, *Proceedings of the First SIAM International Conference on Data Mining* (2001)

Berry, Michael, *Computational Information Retrieval* (2001)

Eppstein, David, Demaine, Erik, Doerr, Benjamin, Fleischer, Lisa, Goel, Ashish, Goodrich, Mike, Khanna, Sanjeev, King, Valerie, Munro, Ian, Randall, Dana, Shepherd, Bruce, Spielman, Dan, Sudakov, Benjamin, Suri, Subhash, and Warnow, Tandy, *Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (2002)

Grossman, Robert, Han, Jiawei, Kumar, Vipin, Mannila, Heikki, and Motwani, Rajeev, *Proceedings of the Second SIAM International Conference on Data Mining* (2002)

Estep, Donald and Tavener, Simon, *Collected Lectures on the Preservation of Stability under Discretization* (2002)

Ladner, Richard E., *Proceedings of the Fifth Workshop on Algorithm Engineering and Experiments* (2003)

Barbará, Daniel and Kamath, Chandrika, *Proceedings of the Third SIAM International Conference on Data Mining* (2003)

Olshevsky, Vadim, *Fast Algorithms for Structured Matrices: Theory and Applications* (2003)

Munro, Ian, Albers, Susanne, Arge, Lars, Brodal, Gerth, Buchsbaum, Adam, Cowen, Lenore, Farach-Colton, Martin, Frieze, Alan, Goldberg, Andrew, Hershberger, John, Jerrum, Mark, Johnson, David, Kosaraju, Rao, López-Ortiz, Alejandro, Mosca, Michele, Muthukrishnan, S., Rote, Günter, Ruskey, Frank, Spinrad, Jeremy, Stein, Cliff, and Suri, Subhash, *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (2004)

Arge, Lars and Italiano, Giuseppe F., *Proceedings of the Sixth Workshop on Algorithm Engineering and Experiments and the First Workshop on Analytic Algorithms and Combinatorics* (2004)

Hill, James M. and Moore, Ross, *Applied Mathematics Entering the 21st Century: Invited Talks from the ICIAM 2003 Congress* (2004)

Berry, Michael W., Dayal, Umeshwar, Kamath, Chandrika and Skillicorn, David, *Proceedings of the Fourth SIAM International Conference on Data Mining* (2004)

MESSAGE FROM THE PROGRAM CO-CHAIRS

We are pleased to present the proceedings of the 2004 SIAM International Conference on Data Mining. The pervasiveness of data mining in research and industry continues to grow, especially in disciplines such as bioinformatics and homeland security. We were excited to have a record number of paper submissions (161) this year as well as a record number of program committee members (90). We hope that the research and experiences captured in these proceedings are insightful to both expert and novice users and practitioners of data mining approaches.

We received 161 paper submissions from 9 countries. Each submitted paper was reviewed by 5 members of the program committee. The reviewing period was followed by a discussion phase. Finally 26 papers (16.1%) were selected to appear in the program as regular papers, another 12 papers (7.4%) were accepted as student papers, and 23 (14.2%) were accepted as poster presentations. The student papers were reviewed under the same guidelines as regular papers this year and hence were allocated the same number of pages (12) as regular papers. Poster papers received 5 pages in the proceedings.

The program of SIAM DM 2004 includes four keynote lectures, four tutorials, and one industry/government laboratory session. The industry/government laboratory session is an addition to the conference, and we hope that it will be an important forum for discussing data mining practices and experiences in those communities. In addition, we have six workshops on topics including bioinformatics, counterterrorism and privacy, clustering high-dimensional data, mining scientific and engineering datasets, high-performance and distributed mining, and data mining in resource constrained environments.

We would like to thank our impressive program committee members, whose dedication and diligence made the selection of papers for these proceedings possible. We also thank the members of the steering committee for their help and guidance. Special thanks go to the conference co-chairs, Chandrika Kamath and David Skillicorn, who supervised the conference planning and deadlines. Thanks also to the tutorial chair, Srinivasan Parthasarathy, who organized an excellent set of tutorials; to the workshop chair, Hillol Kargupta, for doing a superb job arranging the six workshops; and to the sponsorship chair, Sanjay Ranka, for his help in organizing the industry/government laboratory session.

We are grateful to Microsoft Corporation for providing the Conference Management Tool (CMT) that facilitated the collection and management of paper submissions. Special thanks to Tim Olson for his help in training us to use the CMT and for troubleshooting when needed. We also thank Murray Browne and Marcy Copeland at the University of Tennessee for their help in assembling the conference program and the staff at SIAM for their help in the production of this proceedings and in all the necessary arrangements for the conference.

Of course, this conference would not be possible without the excellent papers and presentations represented by this proceedings. We thank all the authors for their participation in SIAM DM 2004.

Michael Berry and Umeshwar Dayal, Program Co-Chairs

PREFACE

The Fourth SIAM International Conference on Data Mining continues the tradition of providing an open forum for the presentation and discussion of innovative algorithms as well as novel applications of data mining. This is reflected in the talks by the four keynote speakers, who will discuss data usability issues in systems for data mining in science and engineering (Graves), issues raised by new technologies that generate biological data (Page), ways to find complex structured patterns in linked data (Senator), and advances in Bayesian inference techniques (Bishop).

In addition to the keynote talks, the conference also features four tutorials and six workshops on a range of subjects. The tutorials will provide the participants an in-depth exposition on selected topics of current interest, including data mining in computer security, analysis of medical patient data, and ways of avoiding common mistakes in data mining. The workshops are a forum for discussing new ideas, brainstorming on work in progress, and identifying new algorithms and application areas for data mining. The topics of the six workshops include data mining in resource-constrained environments; bioinformatics; clustering high-dimensional data and its applications; link analysis, counterterrorism, and privacy; high-performance and distributed mining; and mining scientific and engineering datasets. These workshops and tutorials, in addition to the papers and the poster session, provide an exciting environment in which the participants can interact with each other.

We would like to thank the entire organizing committee for the terrific job they have done in putting together a strong technical program: Michael Berry and Umeshwar Dayal for assembling a well-rounded program committee and for overseeing the paper selection process; Srinivasan Parthasarathy for soliciting and assembling a top-notch tutorial program; Hillol Kargupta for selecting workshops on a diverse range of subjects, all of current interest; Sanjay Ranka for identifying sponsors for the conference; and, finally, the publicity team of Aleksandar Lazarevic, Saso Dzeroski, and John Roddick for their tireless efforts in publicizing the conference. We would also like to thank Morgan C. Wang for the local arrangements. Finally, we would like to acknowledge our sponsors: IBM Research, NASA, and the Center for Applied Scientific Computing at the Lawrence Livermore National Laboratory, for their generous support, in particular the funding of student travel grants.

Finally, we thank the authors and the participants who are the primary reason for the success of the conference. We hope you all enjoy the conference.

Chandrika Kamath and David Skillicorn, Conference Co-Chairs

Vipin Kumar, Steering Committee Chair

CONTENTS

Mining Relationships Between Interacting Episodes

Carl H. Mooney and John F. Roddick
School of Informatics and Engineering
Flinders University of South Australia,
PO Box 2100, Adelaide, South Australia 5001,
Email: {carl.mooney, roddick}@infoeng.flinders.edu.au

Abstract

The detection of recurrent episodes in long strings of tokens has attracted some interest and a variety of useful methods have been developed. The temporal relationship between discovered episodes may also provide useful knowledge of the phenomenon but as yet has received little investigation. This paper discusses an approach for finding such relationships through the proposal of a robust and efficient search strategy and effective user interface both of which are validated through experiment.

Keywords: Temporal Sequence Mining.

1 Introduction and Related Work

While the mining of frequent episodes is an important capability, the manner in which such episodes interact can provide further useful knowledge in the search for a description of the behaviour of a phenomenon. For example, discovering a relationship between input data which was hitherto thought to be independent might lead to the discovery of physical correlations between the environments within which the sensors generating the input data operate. Moreover, any temporal relationships might be used to predict future values in the sensor providing the later input.

General association mining algorithms are used to generate frequent itemsets from which *intra-transaction* association rules are generated. The *sequence* mining task is the discovery of *inter-transaction* associations – sequential patterns – across the same, or similar data. This problem was first addressed by Agrawal and Srikant [2] for mining transactional databases. The solution was based on extending the Apriori algorithm [1] and introduced the AprioriAll, AprioriSome, and DynamicSome algorithms. In order to address identified shortcomings, notably the time between associated sequences, single transaction constraints and to a lesser extent user-defined taxonomies, the GSP (Generalised Sequential Patterns) algorithm [23] was developed. GSP incorporated time constraints (minimum and maximum gap between episodes), and sliding windows, and proved to be more efficient than its predecessors.

Typically improvements in performance have come about by employing a depth-first approach to the mining, as opposed to the more traditional breadth-first approach, and it has been recognised by Yang *et al.* [25] that these methods generally perform better when the data is memory-resident and when the patterns are long. As a result, algorithms based on a depth-first traversal of the search space were introduced and there was an increased focus on incorporating constraints into the mining process. Among these algorithms are SPADE (Sequential PAttern Discovery using Equivalence classes) [27] and its variant cSPADE (constrained SPADE) [26] which relies on combinatorial properties and lattice based search techniques and allow constraints to be placed on the mined sequences, and the SPIRIT (Sequential Pattern mIning with Regular expressIon consTraints) algorithm [11] using regular expression constraints. All of these algorithms use a candidate generation and prune method requiring multiple passes over the data that has inherent problems with large datasets and long sequences, and as a result pattern-growth algorithms have appeared. PrefixSpan [21] being one representative of this type of algorithm.

The ever increasing amount of data being collected has also introduced the problems of how to handle the addition of new data within an existing ruleset and the possible non-relevance of older data, and in association rule mining methods have been proposed to deal with this [8]. With respect to these problems sequence mining is no different and similar techniques have also been developed [18, 20, 28].

The data used for sequence mining is not limited to data stored in overtly temporal or longitudinally maintained datasets – examples include genome searching, web logs, alarm data in telecommunications networks, population health data etc. In such domains data can be viewed as a series of events occurring at specific times

and therefore the problem becomes a search for collections of events (episodes) that occur frequently together. Solving this problem requires a different approach, and several types of algorithm have been proposed for different domains. Manilla *et al.* [17] developed the WINEPI algorithm and evaluated it on alarm detection data; regular expressions have been used to develop string matching algorithms [3, 7] and in web mining [22]. In addition, Yang *et al.* [25] developed a method that uses a *compatibility matrix*[1] in conjunction with a *match* metric ("aggregated amount of occurrences") to discover long sequential patterns within (primarily) gene sequence data.

The purpose of generating frequent sequences, irrespective of the method, is to be able to infer some *rules*, and thus potential *knowledge* about behaviour. For example, given the sequence $A \ldots B \ldots C$ occurring in a string multiple times, rules such as: token (or event) $A \overset{before}{\longrightarrow} B \overset{before}{\longrightarrow} C$, can be expressed. It has been argued by Padmanabhan *et al.* [19] that these types of inference, and hence the temporal patterns, have limited expressive power. For this reason mining for sequences and the generation of rules based on first-order temporal logic (FOTL) [19], has extended the previous work by Manilla *et al.* [16] to include inferences of the type *Since, Until, Next, Always, Sometimes, Before, After* and *While*, by searching the database for specific patterns that satisfy a particular temporal logic formula. One disadvantage to this approach is that no intermediate results are obtained and thus any mining for a different temporal logic pattern must be conducted on the complete database, incurring significant overhead.

Höppner *et al.* [13, 14] use a modified rule semantic, J-Measure and rule specialisation to find temporal rules from a set of frequent patterns in a state sequence. The method described uses a windowing approach, similar to that which is used to discover frequent episodes, and then imposes Allen's interval logic [4] to describe rules that exist within these temporal patterns. Kam *et al.* [15] deal with temporal data for events that last over a period of time and introduce the concept of temporal representation and foster the view that this can be used to express relationships between interval based events, also using Allen's temporal logic.

This type of temporal inference is aligned with the type of interactions we are mining, although in this paper we are limiting the scope of our search to a subset of Allen's temporal relationships, and we propose a method that is based on *point relationships* that may exist within the frequent episodes, not between the frequent episodes. We also take the approach of mining the frequent episodes and incrementally mining the results, in situ, to obtain the interactions. By proceeding in this manner we are able to save the intermediary results (frequent episodes) and the interactions in order that both may be mined at a future time, using different constraints, with a significantly reduced overhead.

This paper investigates the mining of temporal relationships within frequent episodes in a potentially very long string of tokens. Sections 2 and 3 presents the problem formally and presents a methodology for solving it. Since visualisation techniques have not been well examined, Section 4 discusses the user interface constructed to view discovered interactions, and Section 5 discusses our experiments using real-world data. Section 6 concludes with some discussion of future work.

2 Frequent Episode Discovery

The problem of discovering interacting episodes is one that consists of two distinct parts. First (phase 1), the mining of the frequent episodes and second (phase 2), using the discovered frequent episodes as input for the discovery of the interactions. However, since the frequent episodes are available after each pass of the input sequence, phase 2 can be performed in parallel. Note that the two phases are sufficiently different, as is the terminology used, and as such in this section we define the sequence mining problem and the following section defines the interacting episode mining problem. Each section begins by defining the notation that will be used and is followed by detailing the method and the algorithms designed for the tasks.

Let the set of available input tokens (the alphabet), denoted T, be defined as $T = \langle t_1, \ldots, t_k \rangle \mid t_i \neq t_j, i \neq j, 1 \leq i, j \leq k$. A sequence S is then defined as a time ordered ($<$) sequence of input tokens and is denoted $S = \langle s_1, s_2, \ldots, s_m \rangle \mid s_i \in T, 1 \leq i \leq m$. An *episode*, denoted E, is a sequence of tokens, $\langle s_n, s_{n+1}, \ldots, s_{n+k} \rangle$, where $E \subseteq S$. The time at which an event occurs can either be inherent in the input data (as would be the case with alarm detection data), or be implied from the ordering of the input sequence. In the context of our current work this ordering must be strictly *less than* ($<$), that is the relationships being discovered do not require the consideration of events that occur simultaneously. Future work will address the case where this constraint is relaxed.

The user defined *lookahead*, l (similar to Mannila *et al.*'s window concept[17]), defines the maximum length episode to be mined, where $|E| \leq l \leq |S|$. A window, denoted w, is defined as the length of E, where $|E| \leq l$, at any point during the mining process. Therefore the maximum number of windows, *max_win* is given by $|S| - w + 1$ and the *frequency* of E in S is

[1] a conditional probability matrix that specifies the likelihood of symbol substitution

defined as the number of windows in which E appears. The minimum frequency required for an episode to be reported, *min_freq* denoted δ, is calculated using a support, σ (user defined), multiplied by *max_win* at any given point in the mining run. Calculating the minimum frequency in this manner allows for potentially more interesting longer episodes to be reported at a lower threshold since there are fewer windows for longer episodes.

The values for both the *lookahead* l and support σ can be varied on each successive mining run, refer to Section 4 for details. The *tuning* of support for different datasets poses quite a challenge and often requires a detailed knowledge of the domain so that the mining does not produce either too many or too few results.

Thus the problem for this phase becomes: find all episodes

$$E_i \text{ on } \{\mathbf{S} \mid E_i \leq l, freq(E_i) \geq \delta,\ \delta = (|S| - w + 1) \times \sigma\}$$

The algorithm we use for finding the frequent episodes, see Algorithm 2.1, is a breadth-first search of the input sequence starting with single token episodes. These are pruned according to *min_freq*, δ, and added to the set $\mathcal{F}_1$ of frequent episodes. The classic generation of frequent k length episodes, where $k \geq 2$, involves the self-join of the frequent *(k -1* episodes, subsequent pruning in accordance with the *anti-monotone* Apriori heuristic (downward closure principle)[2] [1], and finally frequency derivation through a scan the dataset. The algorithm used is a modified version of the WINEPI algorithm [17] for finding serial episodes, which uses a combination of this principle (downward closure) and a prefix lookup. In our case a similar technique is used, but since only episodes that consist of contiguous tokens are presently of interest, the window width can be increased by one and then a check can be made to see if the first k tokens of the *(k+1)* episode occur in the current window. In order to minimize the size of the candidate sets on subsequent passes of the algorithm we take advantage of this and maintain a set of the k-prefixes of frequent episodes which can then be used to improve valid candidate generation. Thus on each subsequent pass of the algorithm the window width is increased by one and the first k tokens of the generated *(k+1)* candidates are checked against the k-prefixes and retained if there is a match. This candidate pool is then pruned and those candidates that meet the *min_freq* requirements are stored in $\mathcal{F}_{k+1}$. The algorithm terminates when either the *lookahead*, l, is reached or $\mathcal{F}_{k+1} = \emptyset$.

[2] if any length k pattern is not frequent in the database, then its length (k+1) super-pattern can never be frequent

Algorithm 2.1 Main algorithm for finding frequent episodes and frequent interactions

Input: a *sequence* S, of *tokens* $\mathbf{t} \in T$, a *lookahead* l and a *support* σ

Output: the collection $\mathcal{F}(S, l, \sigma)$ of frequent episodes E and the collection F_r of frequent interactions.

```
1:  find C_1 := { α ∈ T | |α| = 1 } ;
2:  i := 0 ; found := true
3:  while i^++ < l and found do
4:    for j := 0; j < |S| − i + 1; j^++ do
5:      α := S_j, ..., S_{i+j}
6:      δ := (|S| − |α| + 1) × σ;
7:      if i > 1 then
8:        if α_k ∈ F_k | α_k = ⟨t_1 ... t_{i−1}⟩ then
9:          add α to C_j
10:       end if
11:     else
12:       add α to C_j
13:     end if
14:   end for
15:   found := F_i ≠ ∅ where
      F_i := { ∀ α ∈ C_j | frequency (α, S, l) ≥ δ }
      /* prune */
16:   if i > 2 and found then
17:     findCurrentRelationships(F_{1..i}, i)
        /* Algorithm 3.1 & 3.2 */
18:   end if
19: end while
20: F_r := pruneCandidateInteractions(I_r)
    /* Algorithm 3.3 */
21: return F(S, l, σ) , F_r
```

3 Discovering Interacting Episodes

During the discovery of the frequent episodes (Algorithm 2.1, line 16), the second phase, finding the interactions that exist within them begins. Let an interaction θ be a temporal relationship between sub-episodes $(e_i, e_j) \mid |e_i| + |e_j| \leq l$, denoted $\theta_r(e_i, e_j) \in \mathcal{R}$, where $\mathcal{R}$ is the set of temporal relationships as described by Allen [4]. To allow for varying length interactions to be discovered, or simply to minimise the interaction length, a *min_interaction_length*, γ, can be supplied by the user, and to allow for varying levels of support a user-defined *min_interaction_supp*, φ, can also be supplied. The exact nature of φ will be discussed later. Thus the problem for this phase is: find all

$$\theta_r(e_i, e_j) \text{ on } \{E_i \mid e_i, e_j \geq \gamma\,,\ \theta_r(e_i, e_j) \geq \varphi\}$$

The following two examples serve to illustrate the nature of the problem.

Figure 1: Section of an input string showing varying window widths.

EXAMPLE 1.

Given the frequent episodes E_1, E_2, E_3 where:

$E_1 = \langle B, R, I, J, A, V, E \rangle$,

$E_2 = \langle B, I, R, A, J, V, E \rangle$, and

$E_3 = \langle B, R, A, I, J, V, E \rangle$.

By inspection it can be seen that if $e_1 = \langle I, J \rangle$ and $e_2 = \langle B, R, A, V, E \rangle$ then the temporal relationship IJ *during* BRAVE exists.

A more complex example can be shown using the frequent episodes from Figure 1 as a source for the discovery of the interactions.

EXAMPLE 2.

Given the following frequent episodes:

$A_1 = \langle G, L, A, T, I, N, R, E, E, K \rangle$

$B_1 = \langle E, N, G, C, A, N, T, O, N, E, S, E, L, I, S, H \rangle$

$C_1 = \langle G, E, F, R, E, R, M, A, N, N, C, H \rangle$

$C_2 = \langle G, E, R, M, A, N, F, R, E, N, C, H \rangle$

$D_1 = \langle L, D, U, T, C, H, A, T, I, N \rangle$

$D_2 = \langle L, A, D, U, T, C, H, T, I, N \rangle$

- Episodes A_1, B_1, D_1, and D_2 are all examples of the **during** relation, denoted $\theta_d(e_1, e_2)$ – LATIN *during* GREEK, CANTONESE *during* ENGLISH and DUTCH *during* LATIN respectively,
- Episode C_1 is an example of an **overlap** relation, denoted $\theta_o(e_1, e_2)$ – GERMAN *overlaps* FRENCH, and
- Episode C_2 is an example of a **meets** relation, denoted $\theta_m(e_1, e_2)$ – GERMAN *meets* FRENCH.

From our experience, while in simple examples, one can easily detect the relationship, as the episodes get longer or more numerous, this task becomes increasingly difficult quite quickly. A further feature in the discovery is the point at which embedded noise becomes part of the dominant relationship. This depends on a number of aspects:

- Whether the sub-episode is frequent in its own right or whether it is only frequent with its noise. For example, given the frequent episodes $\alpha_1\beta\,\alpha_2$ and β and the non-frequent episode $\alpha_1\alpha_2$ we need to decide whether $\alpha_1\alpha_2$ is reportable and/or whether $\alpha_1\beta\,\alpha_2$ is a separate reportable episode from β.
- The decision of how to deal with common tokens within both a dominant and an embedded sub-episode, as in DUTCH *during* LATIN in Example 2.
- The decision of how to handle noise that interrupts an episode at different locations. Given frequent episodes $\alpha_1\alpha_2\alpha_3$, $\alpha_1\beta\alpha_2\alpha_3$ and $\alpha_1\alpha_2\beta\alpha_3$, and infrequent episodes $\alpha_1\alpha_2\alpha_4\beta\alpha_3$ and $\alpha_1\beta\alpha_2\alpha_4\alpha_3$, how can it be recognised (simply) that α_4 is noise and that β is during $\alpha_1\alpha_2\alpha_3$?

The rest of this section demonstrates how we discover the interactions and deal with these problems.

3.1 Algorithms for Interaction Discovery

Time can be viewed as both discrete and linear in nature and, with the exception of Allen [5], a logic of intervals can be constructed using points rather than from intervals themselves [12]. We also take this view in our approach to discovering interactions and as such the algorithm for discovering candidate interactions within the discovered frequent episodes is based on the set of point relationships, Figure 2, between two episodes **a** and **b**.

i.	**a**.start and **b**.start
ii.	**a**.start and **b**.end
iii.	**a**.end and **b**.start
iv.	**a**.end and **b**.end

Figure 2: Point relationships

These relationships can be used to express the complete set of Allen's [4] temporal relations, those of which are immediately relevant *during*, *overlaps*, *meets* and their inverses, are summarised in Table 1. To handle the remaining seven of Allen's relationships (*starts*, *finishes*, *before* and their inverses, and *equal*),