Optimal Approach for Temporal Patterns Discovery

Khellaf BOUANDAS and **Aomar OSMANI**
LIPN UMR CNRS 7030
Université de Paris 13
99, avenue J.-B. Clément 93430 Villetaneuse, France
{ao, kbd}@lipn.univ-paris13.fr

Abstract

This paper presents new technique for discovering temporal patterns when considered primitives are intervals. Apriori technique is the most used one to deal with temporal patterns using point primitives. An extension of this technique is proposed by Höppner to deal with interval primitives. In this paper, we show that it is not necessary to discover all patterns, instead it is sufficient to discover the set of optimum "interesting" patterns, which is smaller than the set of all significant patterns. For this task, we will introduce a new approach proposal to reduce the combinatorial explosion of generated patterns. The resulting technique, called TPGIP (Temporal Patterns Generation with Interval Primitives), is used to discover the optimal set of interesting patterns efficiently. Then, TPGIP explores some symmetric properties of interval algebra and uses partial patterns structure to propose an efficient approach to explore the patterns set in order to generate the candidate patterns. Some experimental and comparative results are shown at the end of this paper.

Key words: Temporal Sequence, Temporal Pattern, Temporal Matrix, Interval Primitives.

Introduction

Since its introduction, Discovery of Sequential Patterns (AGRAWAL, IMIELINSKI, & SWAMI 1993) has become one of the core of data mining tasks. Most of discovering approaches have been developed essentially in the punctual context, assume static data, and they did not consider the time complexity. The primitives are considered as a single point in time. This problem has been addressed in many domains such as planning (MOIZUMI 1998), speech recognition (RABINER & JUANG 1993), telecommunication networks (DOUSSON & DUONG 1999; OSMANI & LÉVY 2000), and DNA applications (ANDRADE, CASADIO, & MASOTTI 2001; JENSSEN *et al.* 2002).

The problem of finding common characteristics of temporal data requires a notion of similarity. It has an elegantly simple problem statement, that is, to find the set of all interval primitives and their temporal relationships expressed in term of Allen's temporal logic (ALLEN 1983).

It is widely recognized that the set of temporal patterns could rapidly grow to be unwieldy: The number of discovered patterns grow exponentially with the number of primitives and the database long. In this paper, we show that it is not necessary to find all patterns to guarantee the patterns set construction. We show that it is sufficient to consider only the optimum subset of interesting patterns. Further, all other patterns may be generated from the optimum subset. Moreover, we can guarantee completeness of the optimum subset of patterns, i.e. given the optimum subset of patterns; all others can be found.

The outline of this paper is as follow: In section 2 we develop the suitable notions of temporal patterns needed to model the patterns problem when considered primitives are intervals. Next, we present the evaluation function in order to fix the observblity interval of temporal pattern. With this definition, the task of interesting pattern enumeration is well defined. In section 4, we propose an efficient technique, introducing effective and simple optimization of Apriori technique (AGRAWAL & SRIKANT 1994). This optimization reduces the number of passes, and at the same time, at each iteration pass, it reduces the number of considered patterns. In section 5, we show some results of the new technique and we give some comparative results.

Normalization

Most works where the objective is to discover similarities consider point primitives with discrete set of dates. This characteristic reduces the time and the space complexities but offer lower expressiveness. Moreover, it does not cover, for example, the primitives duration. In our approach, we consider the works presented by (HÖPPNER 2001; OSMANI & LÉVY 2000) and based on interval algebra properties (ALLEN 1983).

The interval algebra considers the possible relations (see Figure 1) between two interval primitives as the set of all possible combinations of the two intervals on a directed line. There exist thirteen relations consisting on seven basic relations: *equals (eq), before (b), meets (m), overlaps (o), contains (c), starts (s), is-finished-by (f),* plus the converses of the last six relations. The set of the interval relations is noted I and defined as follow:

$$I = \{b, b^{-1}, m, m^{-1}, o, o^{-1}, if, if^{-1}, c, c^{-1}, s, s^{-1}, eq\}$$

Definition 1 (Interval Primitive) *An interval primitive is a triplet (b_A, s_A, f_A) such that s_A denotes the primitive identifier and $[b_A, f_A]$ the occurrence interval of s_A, where b_A and f_A are point primitives and $b_A < f_A$.*

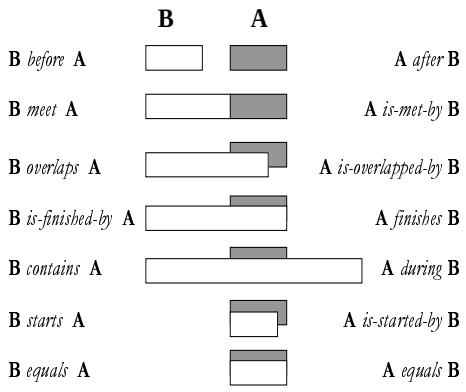

Figure 1: Allen's interval relationships

The normalization process consists to organize the interval primitives by exploiting the partial order property between intervals. However, for the cases where two interval primitives are identical, we define an additional ordering relation (for example identifier names lexicographic order). Than, the set of the ordered interval primitives forms a series of triplets known as the *sequence* of the interval primitives given as follow:

$$(b_1, s_1, f_1), (b_2, s_2, f_2), \ldots, (b_i, s_i, f_i), \ldots, (b_n, s_n, f_n)$$

where $b_i < f_i$ and $b_i \leq b_{i+1}$.

Remark 1 *When the interval limits are not required, we note the intervals sequence as $(s_1, \ldots, s_i, \ldots, s_n)$*

The second property of the normalization process is the maximality of interval primitives in the sense that if the sequence contains two primitives (b_i, s_i, f_i) and (b_j, s_j, f_j) and that $s_i = s_j$ then $s_i\{b, m\}s_j$. If this assumption is violated, the two primitives are substituted by their union $(min(b_i, b_j), s_i, max(f_i, f_j))$.

Definition 2 (Temporal Pattern) *A temporal pattern of dimension n, noted $n-TP$, is defined as a pair (s, R) where s represents the series of interval primitives and the matrix $R \in I^{n \times n}$ refers the relationships between these primitives.*

Example 1 *Figure 2 presents an example of interval primitives sequence. From this example, we can extract the following $n-TP$:*

- $1-TP$: $((s_1); (eq))$
- $2-TP$: $((s_2, s_3); (eq, m), (m^{-1}, eq))$
- $3-TP$: $((s_3, s_4, s_5); (eq, o, b), (o^{-1}, eq, o), (b^{-1}, o^{-1}, eq))$
- $4-TP$: $((s_2, s_3, s_4, s_5); (eq, m, b, b), (m^{-1}, eq, o, b), (b^{-1}, o^{-1}, eq, o), (b^{-1}, b^{-1}, o^{-1}, eq))$
- $5-TP$: $((s_1, s_3, s_4, s_5, s_6); (eq, b, b, b, b), (b^{-1}, eq, o, b, b), (b^{-1}, o^{-1}, eq, o, b), (b^{-1}, b^{-1}, o^{-1}, eq, o), (b^{-1}, b^{-1}, b^{-1}, o^{-1}, eq))$
- $6-TP$: $((s_1, s_2, s_3, s_4, s_5, s_6); (eq, b, b, b, b, b), (b^{-1}, eq, m, b, b, b), (b^{-1}, m^{-1}, eq, o, b, b), (b^{-1}, b^{-1}, o^{-1}, eq, o, b), (b^{-1}, b^{-1}, b^{-1}, o^{-1}, eq, o), (b^{-1}, b^{-1}, b^{-1}, b^{-1}, o^{-1}, eq))$

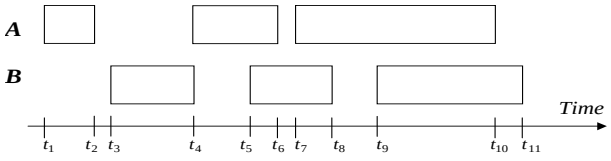

Figure 2: The interval primitives sequence

We denote by $dim(P)$ the dimension of the temporal pattern P; it represents the number of intervals in P. We define $TP(S)$, the set of temporal patterns over S, informally as the set of all temporal patterns of arbitrary dimension. Of course, many interval primitives maintain the same temporal relationships.

Definition 3 (Pattern Instance) *Given P and Q two $k-TP$. Q is said an instance of P if and only if there exists an injective function π from s_P to s_Q such that the relations between these primitives are preserved, i.e.:*

- $\forall i \in \{1, \ldots, k\} : \; s_P(i) = s_Q(\pi(i))$ *and*
- $\forall i, j \in \{1, \ldots, k\} : \; R_P[i, j] = R_Q[\pi(i), \pi(j)]$

We note I_P the set of instances of the pattern P. Figure 2 gives an example. If we consider two patterns P and Q such that I_P corresponds to the set of patterns respecting the relationship "$A\ before\ B$" and I_Q corresponds to the set of patterns respecting the relationship "$B\ overlaps\ A$". The instances sets are:

$$I_P = \{(s_1, s_2), (s_1, s_4), (s_1, s_6), (s_3, s_6)\}$$
$$I_Q = \{(s_4, s_5)\}$$

The instances set can be reduced to the disjointed instances (they did not shared any interval primitives). Then, we denote the set of disjoined instances of P by $I_P^d \subseteq I_P$ and the result will be $I_P^d = \{(s_1, s_2), (s_3, s_6)\}$.

The core of the proposal technique is the candidate generation. To make easier this generation, we utilize the sliding window (MANNILA, TOIVONEN, & VERKAMO 1997), and we enlarge the observability interval notion, for an effective selection of the interesting patterns.

Observability interval

The sliding window constitutes a more used tool to sweep the primitives sequence. If we consider punctual primitives, the sliding window is defined as a maximal period of time

where pattern may be observed. In the case of interval primitives, the sliding window corresponds to the maximal interval where pattern P may be partially observed (all primitives of P are relation in $I \setminus \{b,\ b^{-1}\}$ with the sliding window). In (HÖPPNER & KLAWONN 2001), the continued period in which an instance of pattern P can be observed inside the sliding window, is identified as the observability interval of this instance.

Definition 4 (Obsevability interval) *Let us consider P the temporal k-pattern and w the width of sliding window. We define the obsevability interval of an instance of P, denoted by O_P, as follow:*

$$O_P = \begin{cases} [f_{(s_P)_1}, b_{(s_P)_1} + w], & k = 1 \\ O_Q \cap [f_{(s_P)_k}, b_{(s_P)_k} + w], & k > 1 \end{cases}$$

where $Q \sqsubseteq P$ and $dim(Q) = k - 1$.

In (HÖPPNER 2001), Frank Höppner declares the pattern "interesting" if the total duration of the observability intervals of the pattern instances set, within the sliding window, exceeds an optimum threshold of the sliding window. Whereas, this adaptation is not relevant and it generates a redundancy in the observability interval. This is due to the use of the observability intervals of all instances without restrictions.

Given $P = (s_P, R_P)$ an instance of the i-pattern P such that $\{\ b,\ b^{-1},\ c,\ c^{-1}\} \not\subseteq R_P$ and $Q = (s_Q, R_Q)$ an instance of the j-pattern Q such that $\{\ b,\ b^{-1},\ c,\ c^{-1}\} \not\subseteq R_Q$. The maximality assumption guarantees that all instances of both P and I are disjoined. Let M be the temporal pattern resulting from the join of an instance of P with an other one of Q. In the sliding window, we have the following possible three scenarios.

case 1: We consider an instance of P and two instances of Q, respectively Q_1 and Q_2. M has at least two instances who shared the i-pattern P. Let M_1 and M_2 be these two instances obtained by the join of P with respectively Q_1 and Q_2. Then, the respective observablity are $[b_{Q_1}, f_P]$ and $[b_{Q_2}, f_P]$ and the duration d_M of M is:

$$\begin{aligned} d_M &= d_{M_1} + d_{M_2} \\ &= (f_P - b_{Q_1}) + (f_P - b_{Q_2}) \\ &\geq f_P - b_{Q_1} = d_{M_1} \\ &> f_P - b_{Q_2} = d_{M_2} \end{aligned}$$

case 2: Let P_1 and P_2 be two instances of P, and we consider a single instance of Q. Let M_1 and M_2 be the two instances of M corresponding respectively to the join of the instances of P with Q. The respective observability intervals $[b_Q, f_{P_1}]$ and $[b_Q, f_{P_2}]$. The duration of M is:

$$\begin{aligned} d_M &= d_{M_1} + d_{M_2} \\ &= (f_{P_1} - b_Q) + (f_{P_2} - b_Q) \\ &\geq f_{P_1} - b_Q = d_{M_1} \\ &> f_{P_2} - b_Q = d_{M_2} \end{aligned}$$

case 3: We consider two instances of P, P_1 and P_2, and two instances of Q, Q_1 and Q_2. Let M_1 and M_2 be two instances resulting from the respective join of P_1 and P_2 with Q_1 and Q_2. If R_{M_1}=R_{M_2}, the resulting patterns refer the same temporal pattern M, the respective observablity intervals are $[b_{Q_1}, f_{P_1}]$ and $[b_{Q_2}, f_{P_2}]$. The duration of M is compued as follow:

$$\begin{aligned} d_M &= d_{M_1} + d_{M_2} \\ &= (f_{P_1} - b_{Q_1}) + (f_{P_2} - b_{Q_2}) \\ &\geq f_{P_1} - b_{Q_1} = d_{M_1} \\ &> f_{P_2} - b_{Q_2} = d_{M_2} \end{aligned}$$

Consequently, we realize the following points:

- The observability interval of the first instance of given temporal pattern covers completely the observability interval of all other instances in the same sliding window i.e $O_{M_2} \subseteq O_{M_1}$.
- The observation of the first instance of given temporal pattern is sufficient to observe all other instances in the same sliding window.
- The gap between the composites interval of the first instance is optimal.

Example 2 *Let us consider the example in Figure 3. The pattern "A is-finished-by B" has only one instance (s_1, s_3) for all considered window width, whereas the pattern "$A\ before\ B$" for a suitable window (for example for window width greater than $t_9 - t_5$) has two instances not disjointed (s_1, s_4) and (s_1, s_5), who the respective observability intervals are $[t_6, t_5 + w]$ and $[t_8, t_5 + w]$. Thus, we conclude $[t_8, t_5 + w] \subseteq [t_6, t_5 + w]$.*

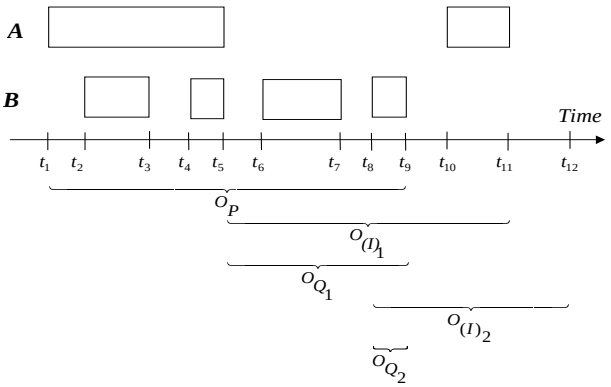

Figure 3: Pattern's observability intervals

This observability interval property is one of ideas applied in our proposed approach to reduce the combinatorial explosion of generated patterns relative to the approches using Apriori technique.

Candidate generation

The proposal technique is an alternative of Apriori technique when considered primitives are intervals. TPGIP exploits the interval algebra structure employing partial order relation on the intervals primitives. It privileges the temporal patterns that the gap between the composites interval is optimal and exploits of the neighborhood properties for a better selection of patterns two by two. Before introduce the

TPGIP technique, we start with a new notions and relations in order to facilitate the cross of the k-patterns subset, noted $TP_k(S) \subseteq TP(S)$.

Definition 5 (Prefix and Suffix Pattern) *Let P and Q be respectively the $n-TP$ and the $k-TP$.*

1. *Q is the k-prefix pattern of P iff:*
   $s_Q = \{(s_P)_i / (s_P)_i \in s_P \wedge 1 \leq i \leq k\}$
2. *Q is the k-suffix of P iff:*
   $s_Q = \{(s_P)_i / (s_P)_i \in s_P \wedge (n-k) < i \leq n\}$

Definition 6 (Precedence Relation) *Given two temporal k-patterns P and Q. P precedes Q, noted $P \preceq Q$, iff:*

$\exists i = max\{1, min\{j/(s_P)_j \neq (s_Q)_j\}\} : (s_P)_i \leq (s_Q)_i$

Theorem 1 *The precedence relation defines a partial order on the patterns subset with the same size.*

Proof: Let us suppose $M, P, Q \in TP_k(S)$. Then, the precedence relation is:

1. *reflexive: $P \preceq P$ for i=1.*
2. *transitive: $M \preceq P \wedge P \preceq Q \Rightarrow M \preceq Q$.*
   $M \preceq P \Leftrightarrow \exists i = max\{1, min\{j/(s_M)_j \neq (s_P)_j\}\} : (s_M)_i \leq (s_P)_i$
   $P \preceq Q \Leftrightarrow \exists r = max\{1, min\{j/(s_P)_j \neq (s_Q)_j\}\} : (s_P)_r \leq (s_Q)_r$
   For $l = min(i, r) : (s_M)_l \leq (s_P)_l \leq (s_Q)_l$
   then $M \preceq Q$.
3. *anti-symmetric: $P \preceq Q \wedge Q \preceq P \Rightarrow P = Q$*
   $P \preceq Q \Leftrightarrow \exists i = max\{1, min\{j/(s_P)_j \neq (s_Q)_j\}\} : (s_P)_i \leq (s_Q)_i$,
   $Q \preceq P \Leftrightarrow \exists r = max\{1, min\{j/(s_Q)_j \neq (s_P)_j\}\} : (s_Q)_r \leq (s_P)_r$,
   Therefore $i = r$ *and* $(s_P)_i \leq (s_Q)_i \wedge (s_Q)_r \leq (s_P)_r$
   Then $\forall i \leq dim(P) : (s_P)_i = (s_Q)_i$, *i.e.* $P = Q$.

We will exploit the Theorem 1 to structure the patterns set. We are interested exclusively by the neighbourhood patterns, said *adjacent patterns*.

Definition 7 (Adjacent Patterns) *Let P and Q be two temporal k-patterns. P is adjacent with Q iff:*

- *$P \preceq Q$ and*
- *the $(k-1)$-suffix of P is the same as the $(k-1)$-prefix of Q.*

The new technique is an optimal and iterative traversing of the patterns set. At iteration k, corresponding to a level, a subset of candidate patterns is created by the join of the adjacent interesting patterns discovered during the previous iteration. This generation supports exclusively the ordered patterns two by two such that the gap between those intervals composite is optimum. It takes in entry P and Q, two adjacent interesting $k-TP$, and generates, C a single candidate $(k+1)-TP$.

Figure 4 illustrates how to build the matrix R_C out of R_P and R_Q. As consequence, R_C maintains the matrix R_P like prefix matrix and envelopes them by a new line and a new column. The principal task of this construction is to generate this envelope. As the relations of each column are the opposite relations of the corresponding line, TPGIP computes the relations of the last line of R_C. Thus, it recovers the relations of the last line of R_Q to extract the last $k-1$ relations from the last line of R_C. The first relation of this line is the temporal relation between $(s_Q)_{k-1}$ and $(s_P)_1$, which is respectively the first interval primitive of P and the last interval primitive of Q.

R_P	$(s_P)_1$	$(s_P)_2 \ldots (s_P)_{k-1}$
$(s_P)_1$	$=$	B
$(s_P)_2$ $\vdots$ $(s_P)_{k-1}$	C	A

(a) Pattern P

R_Q	$(s_Q)_1 \ldots (s_Q)_{k-2}$	$(s_Q)_{k-1}$
$(s_Q)_1$ $\vdots$ $(s_Q)_{k-2}$	A	D
$(s_Q)_{k-1}$	E	$=$

(b) Pattern Q

R_C	$(s_P)_1$	$(s_P)_2 \ldots (s_P)_{k-1}$	$(s_Q)_{k-1}$
$(s_P)_1$	$=$	B	r
$(s_P)_2$ $\vdots$ $(s_P)_{k-1}$	C	A	D
$(s_Q)_{k-1}$	ir	E	$=$

(c) Candidate pattern C

Figure 4: Candidate pattern generation

Evaluation and Discussion

This section gives a comparison between our technique and the Höppner's one (HÖPPNER 2001). We have treated a meteorological file application containing 1000 interval primitives divided on three variables. We have tested these two techniques by comparing the size of temporal set. For the experiments, we have chosen to vary the window width and to fix the minimal threshold of the support to 10% of the width of sliding window. For the output data, we have captured the size of the temporal subsets associated to the width of sliding window.

Figure 5 shows the evolution of the pattern set for the Höppner's technique with the variation of the width of sliding window. For every width of the sliding window, the evolution follows two trajectories: First the number of patterns grows exponentially with the growth of algorithm pass until generating the maximum pattern with the primitives bordering met the sliding window. Beyond this point, the decreasing trajectory starts. This trend corresponds to add the intervals not yet incorporated in the maximum pattern and who are between the primitives bordering the pattern. Then, the Höppner's approach illustrates the combinatorial explosion realative to Apriori technique.

For our proposal technique TPGIP, the problem of combinatorial explosion does not appear. Figure 6 contrary to Figure 5, shows a single decreasing trend for the evolution of the number of patterns according to the width of sliding window. Thus, the number of patterns generated by TPGIP,

at each iteration, is strictly lower than the number of the interesting patterns at previous iteration. For the intervals sequence, the iterations number is limited to the sequence dimension n. At the iteration k, the number of candidate patterns, independently with the width of the sliding window, is at least $n - k + 1$. Therefore, the pattern number generated by our approach is little than $\sum_{k=1}^{n} k = n(n-1) \setminus 2$.

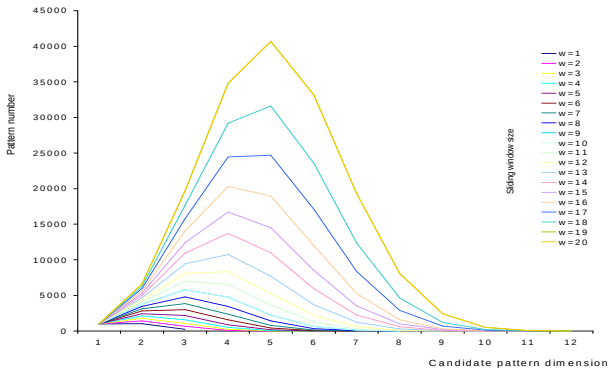

Figure 5: Höppner's technique

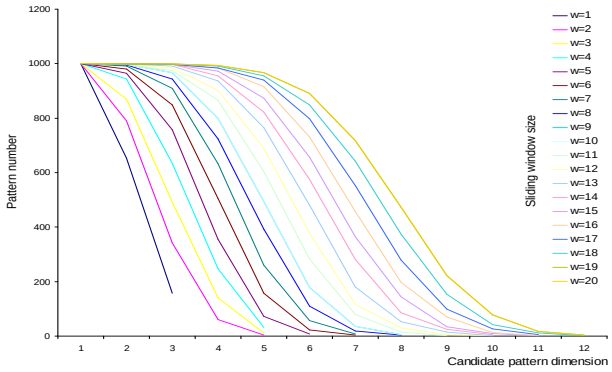

Figure 6: TPGIP technique

Conclusion

This paper adresses an approach for the discovery of an interesting patterns in temporal sequences according to the observability interval properties. It proposes a new technique, called TPGIP (Temporal Patterns Generation with Interval Primitives), to find a minimal subset of interesting temporal patterns from what it is possible to generate all other interesting patterns. This technique, applied to interval primitives, explores some properties of interval algebra relations (ex. symmetric property) to compact generated patterns. Then, it combines the partial oreder and the locality patterns over the pattern subset with the same size.

TPGIP performances are compared to Apriori technique. Encouraging and promising results are observed and reported for the proposed technique to the meteorological application data. It shows that the number of generated patterns is polynomial according to the dimension of the patterns and the width of the sliding window. The presented experimental results show the effectiveness and the performance of the proposed approach. This approach guarantees finding all remainder patterns and declares all them interesting without compute their observability intervals.

References

AGRAWAL, R., and SRIKANT, R. 1994. Fast Algorithms for Mining Generalized Association Rules. In *Proc. of the 20th Int'l Conf. on Very Large Databases. Expanded version in IBM Research Report RJ9839*, 478–499.

AGRAWAL, R.; IMIELINSKI, T.; and SWAMI, A. 1993. Mining Association Rules between Sets of Items in Large Databases. In *Proc. of the ACM SIGMOD Conference on Management of Data*, 207–216.

ALLEN, J. F. 1983. Maintaining Knowledge about Temporal Intervals. In *Comm. ACM*, 26(11), 832–843.

ANDRADE, M. A.; CASADIO, R.; and MASOTTI, L. 2001. Tools for Automated Protein Annotation Protein Sequence Analysis in the Genomic Era. In *CLUEB*.

DOUSSON, C., and DUONG, T. V. 1999. Discovering Chronicles with Numerical Time Constraints from Alarm Logs for Monitoring Dynamic Systems. In *Proc. of the 6th Int'l Joint Conf. on Artificial Intelligence*, 620–633.

HÖPPNER, F., and KLAWONN, F. 2001. Finding Informative Rules in Interval Sequences. In *Proc. of the 4th Int'l Symposium, Lecture Notes in Computer Sciences*, 123–132.

HÖPPNER, F. 2001. Discovery of Temporal Patterns-Learning Rules about the Qualitative Behaviour of Time Series. In *Proc. of the 5th European Conf. on Principles and Practice of KDD, Lecture Notes in Artificial Intelligence*, 2168, 192–203.

JENSSEN, T. K.; ÖBERG, L. M. J.; ANDERSSON, M. L.; and KOMOROWSKI, J. 2002. Methods for Large-Scale Mining of Networks of Human Genes. In *the SIAM Conference on Data Mining*.

MANNILA, H.; TOIVONEN, H.; and VERKAMO, A. 1997. Discovery of Frequent Episodes in Event Sequences. In *Data Mining and Knowledge Discovery*, 1(3).

MOIZUMI, K. 1998. *The Mobile Agent Alanning Problem*. Phd thesis, Thayer School of Engineering, Dartmouth College.

OSMANI, A., and LÉVY, F. 2000. Simulating Faults in Telecommunication Network: Reasoning about Incertain Propagation of Events. In *Proc. of the 15th Int'l Conf. on Computers and Their Applications (CATA)*, 15–19.

RABINER, L. R., and JUANG, B. H. 1993. Fundamentals of Speech Recognition. In *Prentice Hall*.