# Un Modelo de Reglas de Asociación Espacio-Temporales

Mariano S. Kohan S., Juan M. Ale

Facultad de Ingeniería, Universidad de Buenos Aires
marianokohan@gmail.com, ale@acm.org

**Abstract.** El descubrimiento de reglas de asociación es una tarea del área de la minería de datos que permite obtener patrones que muestran conjuntos de elementos que co-ocurren de manera frecuente en un conjunto de transacciones. Con el avance en las tecnologías de recolección de datos han aparecido diversos tipos de modelos para el descubrimiento de reglas de asociación espaciales y reglas de asociación temporales. En este trabajo se propone un modelo de reglas de asociación espacio-temporales, aplicable sobre bases de datos cuyas transacciones contienen componentes espaciales y temporales. Asimismo, se presentan los algoritmos para el descubrimiento de dichas reglas de asociación espacio-temporales, junto con la experimentación realizada con los mismos.

**Palabras claves:** Minería de Datos, Reglas de Asociación, Minería de Datos Espacio-Temporales, Reglas de Asociación Espacio-Temporales

## 1 Introducción

La minería de datos es un conjunto de técnicas que permiten extraer patrones de conocimiento novedosos desde bases de datos de gran volumen. Su aparición se debió principalmente a los avances en tecnologías para la recolección y almacenamiento de datos, frente a la necesidad de analizar esos grandes volúmenes de datos almacenados. Entre las tareas posibles de la minería de datos se destaca la de descubrimiento de reglas de asociación [1]. La misma permite obtener patrones que muestran conjuntos de elementos que co-ocurren de manera frecuente en un conjunto de transacciones. Existen varios algoritmos desarrollados para el descubrimiento de las reglas de asociación, entre los cuales se pueden destacar [2] y [3]. La mayoría de los mismos fueron desarrollados para trabajar con tipos de datos simples: booleanos, categorías de valores, numéricos.

Con el tiempo las tecnologías para recolección de datos se han mejorado, permitiendo recolectar datos de diversas fuentes y de diferente naturaleza. Esto dio lugar a la aparición de áreas dentro de la minería de datos dedicadas a obtener patrones de conocimiento espacial [4], temporal y espacio-temporal [5].

De esta manera, para la tarea de reglas de asociación han aparecido también modelos para el descubrimiento de reglas de asociación espaciales ([6] y [7]) y reglas de asociación temporales ([8] y [9]). Por otro lado, existe el modelo de reglas de asocia-

ción inter-transacciones [10], que permite obtener reglas de asociación en donde los elementos de las mismas se relacionen con los elementos de otras transacciones, mediante atributos que pueden ser de tipo espacial o temporal.

En este trabajo se propone un modelo de reglas de asociación espacio-temporales. El mismo se aplica sobre bases de datos con componentes espaciales y temporales y permite obtener patrones de reglas de asociación que relacionan a los elementos de las transacciones, indicando además las características espaciales y temporales que ocurren entre los elementos de esa relación, de una manera conveniente para cada tipo de conocimiento involucrado.

El resto del trabajo se encuentra organizado de la siguiente manera: la siguiente sección comenta otros trabajos anteriores sobre el tema y su relación con el actual, indicando similitudes y diferencias. La sección 3 plantea los conceptos del modelo de reglas de asociación espacio-temporales desarrollado. En la sección 4 se explican las características principales de los algoritmos que permiten obtener las reglas del modelo. En la sección 5 se describe la implementación de los algoritmos y los detalles de la experimentación realizada con la misma. Finalmente, en la sección 6 se plantea la conclusión y trabajo futuro.

## 2 Trabajos Relacionados

Uno de los primeros trabajos que trata sobre el descubrimiento de reglas de asociación espacio-temporales es [11]. El mismo maneja datos censales los cuales se procesan para obtener transacciones correspondientes a polígonos espaciales definidos en los datos. Temporalmente las variables registradas se toman como la diferencia entre los valores de los años 1970 y 1990, permitiendo registrar aumentos o disminuciones en el tiempo. Sobre estos datos se aplican algoritmos convencionales para el descubrimiento de reglas de asociación. En [12] se plantea un método para el descubrimiento de reglas de asociación espacio-temporales donde también se procesan los datos para obtener transacciones y luego se aplican los algoritmos convencionales para el descubrimiento de reglas de asociación. En este caso el método de conversión, denominado *pivoting*, es más genérico y según los atributos seleccionados en el mismo, permite obtener reglas del tipo espacial, temporal, o espacio-temporal.

Por último, [13] plantea un modelo para el descubrimiento de reglas de asociación espacio-temporales en bases de datos de objetos móviles. Las reglas de asociación de este modelo tratan de representar los movimientos de los objetos entre regiones entre dos intervalos de tiempo.

De los trabajos nombrados, solamente [13] plantea un modelo y un algoritmo particular para el descubrimiento de las reglas de asociación, aunque las bases de datos sobre las que se descubren las mismas no son las convencionales.

En el presente trabajo se propone un modelo para el descubrimiento de reglas de asociación del tipo espacio-temporales como en [13], aunque aplicable sobre bases de datos de transacciones que contienen atributos del tipo espacial y temporal, más similares a las tratadas en los trabajos [11] y [12].

## 3 Modelo

En esta sección se explican los conceptos principales del modelo de reglas de asociación espacio-temporales.

El modelo trabaja con bases de datos que contengan elementos espaciales y temporales, planteando un esquema general para la estructura de las mismas. El elemento temporal se representa mediante un timestamp señalando el tiempo en que se registra la transacción; el espacial es una posición geográfica (posición en el plano, coordenadas en grados u otro similar) asociada al item en la transacción, indicando donde se registro el mismo. Representando ambos elementos en conjunto queda:

*Sea I={ $i_1$, $i_2$, $i_3$, ... $i_k$, ... $i_n$ } un conjunto de items posibles, se define a la* ***base de datos D*** *como un conjunto de duplas, denominadas transacciones (**Tr**), de la forma*

$$Tr = \left\langle t_i, \left\{ \left\langle i_k, \left(x_j, y_j\right) \right\rangle, \ldots \right\} \right\rangle$$

*donde $t_i$ será el timestamp en el que se registra la transacción y ($x_j$, $y_j$) cada una de las posiciones geográficas asociadas a cada item $i_k$, indicando la ubicación espacial donde se registraron cada uno de los mismos dentro del timestamp $t_i$.*

Un itemset es un conjunto de items ([1]) asociado a una componente espacial y otra temporal. La componente espacial se basa en el modelo de reglas de asociación inter-transacciones [10] tomando los 2 elementos de las posiciones geográficas como atributos dimensionales. En este modelo se permite relacionar un item con los demás en otras transacciones a través de los componentes de las posiciones geográficas. De esta manera en cada itemset extendido ([10]) los items se relacionan con respecto a un item base (correspondiente a la posición geográfica de referencia del itemset). En el modelo actual, se aplica esta idea salvo que se utiliza un conjunto de predicados espaciales definidos entre dos posiciones geográficas, identificado como *PredS*, para señalar las relaciones espaciales entre los items del conjunto. Esto permite relacionar items ubicados en cualquier posición geográfica de la base de datos y no solamente los cercanos, limite dado por la ventana de desplazamiento ([10]). La componente temporal se basa en el modelo de reglas de asociación temporales generalizadas de [8]. De esta manera, está compuesta del *lifespan* o tiempo de vida del itemset, identificado como *l*. El mismo representa el período en el cual el itemset se verifica en la base de datos ([8]).

*Sea X un* ***k-itemset*** *de la forma* $X = \{i_1(POS_1), i_2(POS_2), \ldots i_k(POS_k)\}$, $P = \{pred_1, pred_2, \ldots pred_p\}$ *un conjunto de predicados espaciales posibles definidos entre dos variables e I un conjunto de items posibles definidos sobre una base de datos D, con la estructura señalada*

***PredS*** *para X vale:* $\Pr edS_X = \{pred_{j2}(POS_1, POS_2), pred_{j3}(POS_1, POS_3), \ldots pred_{jk}(POS_1, POS_k)\}$

*El* ***lifespan*** *de X es:* $l_x = [t_1, t_2]$, *donde*

$t_1 = mín\{t / Tr \in D \wedge verifica(Tr, X)\}$ $\quad t_2 = máx\{t / Tr \in D \wedge verifica(Tr, X)\}$

En la definición de *PredS* anterior, cada $POS_i$ representa una variable de posición, que permite identificar para cada predicado espacial a que par de posiciones geográficas de cada item del itemset se aplica el mismo. Dentro de la definición de *lifespan* se

comento el concepto de un itemset verificado en una transacción de la base de datos. Para este caso, además de hallarse los items del itemset contenidos dentro de la transacción ([1]), se deben verificar entre las posiciones geográficas de los items de la misma los predicados espaciales definidos dentro de la componente PredS del itemset.

*Sea X un k-itemset y una transacción Tr perteneciente a D, se indica que* ***Tr verifica X*** *(señalándolo con* ***verifica(Tr,X)****) si:*

$$\left(\exists i_k \in Tr : i_k = i_1 \wedge \forall i_l \in X, l > 1 : \exists i_{k'} \in Tr, \left(i_{k'} = i_l \wedge pred_{jl}((x_k, y_k),(x_{k'}, y_{k'}))\right)\right)$$

De esta manera se puede identificar a $\boldsymbol{V_D(X, [t,t'])}$ como el conjunto de transacciones dentro del intervalo de tiempo [t,t'] que verifican X ([8]):

$$V_D(X,[t,t']) = \{Tr \in D / t \in [t,t'] \wedge verifica(Tr, X)\}$$

En este caso se aplica también la definición de *verifica* del modelo que considera la componente espacial del itemset.

Considerando lo anterior, las métricas del modelo se definen de igual manera a [8], aunque considerando que cuando se utilice dentro de las mismas el concepto de $V_D(X, [t,t'])$ se utilizará el definido en este modelo, ya que el mismo incluye el concepto de verificación de un itemset dentro de una transacción que considera la componente espacial del itemset. Junto con estas métricas se definen los umbrales de valor mínimo, **σ** y **τ,** para identificar a los itemsets frecuentes.

Pueden existir itemsets que no sean frecuentes dentro de su lifespan sino en subintervalos contenidos dentro de estos últimos. Para estos casos se utiliza el concepto de *lifespan frecuente* de un itemset, definido como el conjunto de subintervalos dentro de su lifespan donde el itemset es frecuente ([8]). Del mismo no se consideran todos los subintervalos donde se superen los umbrales de soporte, sino aquellos que superen el umbral del soporte y soporte temporal y sean maximales respecto a este último. Para estos itemsets el lifespan frecuente pasa a ser su componente temporal, representada por *fl.*

Una regla de asociación espacio-temporal va a estar compuesta por dos itemsets, mostrando los componentes espacial y temporal de los mismos como el conocimiento espacio-temporal aportado por la regla.

*Sea I un conjunto de items posibles definidos sobre una base de datos D, se define a una* ***regla de asociación espacio-temporal*** *como la expresión:*

$$X \rightarrow Y : [\Pr edS_{X \cup Y}; fl_{X \cup Y}] : [\{s_1, s_2, \ldots, s_m\}; \{c_1, c_2, \ldots, c_m\}]$$

*donde*

$$X, Y: \text{itemsets de la forma } X = \{i_1(POS_1), i_2(POS_2), \ldots i_k(POS_k)\},$$

$$Y = \{i_{k+1}(POS_{k+1}), i_{k+2}(POS_{k+2}), \ldots i_l(POS_l)\}$$

$$\Pr edS_{X \cup Y} = \{pred_{j2}(POS_1, POS_2), pred_{j3}(POS_1, POS_3), \ldots predj_k(POS_1, POS_k), \ldots, pred_{j(k+1)}(POS_1, POS_{k+1}), \ldots pred_{jl}(POS_1, POS_l)\}$$

$$fl_{X \cup Y} = \{[t,t'] \subseteq l_{X \cup Y} / |[t,t']| \geq \tau \wedge \sup(X \cup Y, [t,t']) \geq \sigma \wedge (\neg \exists [t_2, t_2'] : ([t,t'] \subset [t_2, t_2'] \wedge \sup(X \cup Y, [t_2, t_2']) \geq \sigma))\}$$

Para poder presentar al usuario las reglas con suficiente representatividad e implicancia, se definen las medidas de interés de las mismas de manera similar a [8]: soporte y confianza, las que se calculan para cada intervalo frecuente. Estos valores del soporte y la confianza calculados son los que aparecen dentro del segundo corchete de

la regla (cada $s_n$ corresponde al soporte y cada $c_n$ a la confianza). Junto con estas medidas de interés se definen los umbrales mínimos, $\sigma$ y $\theta$, para considerar cada regla.

Considerando lo señalado, se puede expresar la interpretación de las reglas:

*"Dentro de los intervalos de tiempo definidos por $fl_{x\cup y}$, se registran en conjunto dentro $s_1$, $s_2$, ..., $s_m$ % de las transacciones los elementos de $X\cup Y$ con las relaciones espaciales representadas según $PredS_{x\cup y}$, definidas entre las posiciones geográficas donde se registraron los items. Además cada vez que se registran los elementos del conjunto de items X, en $c_1$, $c_2$, ...,$c_m$ % de las transacciones dentro de los intervalos de tiempo definidos por $fl_{x\cup y}$, tienden a registrarse los valores del conjunto de items Y con las relaciones espaciales representadas según $PredS_{x\cup y}$, definidas entre las posiciones geográficas donde se registraron los items".*

## 4 Algoritmos

El descubrimiento de las reglas de asociación espacio-temporales propuestas por el modelo se puede dividir en dos etapas de manera similar a la tarea de [1]. A continuación se describen las características más importantes de los algoritmos utilizados en las mismas.

### 4.1 Descubrimiento de Itemsets Frecuentes

Como base para el desarrollo del algoritmo para el descubrimiento de los itemsets espacio-temporales frecuentes se consideró el algoritmo Apriori ([2]), utilizado para el descubrimiento de itemsets frecuentes. El mismo se extendió incorporando las modificaciones correspondientes para el descubrimiento de las componentes espaciales y temporales. Para estas últimas se tuvieron en cuenta las modificaciones propuestas en [8], extendidas para el caso del modelo actual. El algoritmo propuesto se denomina STApriori, y se caracteriza principalmente por los elementos descriptos a continuación.

**Generación de Itemsets Candidatos.** Las extensiones se ubican en los diferentes pasos de la etapa:

- Paso de Junta: se agrega en la condición de junta verificar también la igualdad de los predicados de PredS correspondientes a cada itemset frecuente que formará al candidato, hasta el penúltimo elemento de cada componente. A diferencia de lo propuesto en [8], para este modelo el lifespan del nuevo candidato no es posible de ser determinado en base a los itemsets que lo forman. De esta manera, se genera de manera dinámica durante la contabilización.
- Paso de Poda: se debe considerar que los "subconjuntos" de un candidato se encuentran conformados por los items más los predicados espaciales que conforman PredS. Además, considerando que el lifespan se genera de manera dinámica, se puede señalar para cada nuevo candidato $i_c$, formado en base a los itemsets $i_j$ e $i_k$,

$l_{i_c} \subseteq l_{nec} = l_{i_j} \cap l_{i_k}$. Esto se puede utilizar como poda adicional de los nuevos candidatos verificando que $|l_{nec}| \geq \tau$, similar a como se plantea en [8].

**Verificación en Transacciones y Contabilización de Itemsets.** Para determinar si un itemset se verifica dentro de una transacción se aplica la definición dada por el modelo. Si el itemset se verifica en la transacción, se actualizan los campos contabilizadores, correspondientes a los básicos junto con los necesarios para la determinación del lifespan dinámicamente. Además, se mantiene un histograma de transacciones en cada intervalo $\Delta t$ ([8]).

**Verificación de Itemsets Frecuentes según Parámetros de Umbrales.** Se aplica el esquema de ([8]): se deben verificar los dos umbrales definidos, los cuales se aplican primero sobre el lifespan según las definiciones del modelo. En caso de que el itemset no verifique el umbral de soporte y si el de soporte temporal, se aplica el algoritmo a-posteriori ([8]) para hallar los subintervalos maximales frecuentes contenidos dentro del lifespan.

### 4.2 Generación de Reglas de Asociación Espacio-Temporales

Para el armado de las reglas se consideró el algoritmo Faster Algorithm propuesto en [2], al que se le aplicaron modificaciones propias del modelo. En el mismo, las reglas armadas contendrán al item principal (aquel respecto del cual se relacionan los demás) ubicado en el antecedente de las mismas, para evitar posibles confusiones en interpretar las reglas por parte del usuario. Además, en la obtención de los consecuentes, se debe tener en cuenta que los mismos se componen de una componente espacial PredS conteniendo predicados, los cuales serán relativos respecto al item principal de las reglas. Por último, para el cálculo de la confianza de las reglas es necesario determinar el soporte de la premisa en el intervalo frecuente del itemset que forma la regla, extendiéndose para este caso las propuestas de [8].

## 5 Implementación y Experimentación

Se ha desarrollado una implementación del algoritmo sobre el sistema Weka [14], el cual contiene un conjunto variado de algoritmos para ser utilizados en las tareas de minería de datos. Para el cálculo de las relaciones espaciales se utiliza GeoTools [15]. La misma es una librería desarrollada en java para la manipulación de datos geoespaciales, la cual se utiliza en la implementación de sistemas relacionados con la tecnología de los GIS.

Para la discretización de los atributos continuos existentes en los datasets se ha desarrollado un filtro, que utiliza los algoritmos de clustering existentes en Weka para discretizar atributos seleccionados en los datos de entrada.

### 5.1 Experimentación

La solución implementada ha sido aplicada sobre un dataset utilizado en el área del transporte de gas, que contiene mediciones del gas natural transportado a través de determinados gasoductos. La componente temporal de este dataset es la fecha en que se registraron las variables medidas, y la espacial la posición geográfica de la estación de medición, en coordenadas de longitud y latitud.

Se ejecutó un caso de ejemplo con este dataset para analizar los resultados obtenidos con el mismo. Se armó un dataset tomando los valores medidos por 8 estaciones de medición a lo largo de todo un año (365 instantes de tiempo). Los valores de las variables medidas fueron discretizados utilizando el algoritmo de clustering EM ("Expectation Maximisation"), provisto por Weka. En la tabla 1 se muestra la descripción de las variables utilizadas junto con los conjuntos de discretización obtenidos (EM identifica a los mismos mediante la media y desvío estándar de la distribución normal que modeliza a cada uno).

Para los predicados espaciales se utilizaron relaciones de tipo distancia-orientación. Los componentes de las mismas fueron discretizados con los siguientes intervalos:

- distancia ("DIST"): MC=[ 0; 40), C=[ 40; 100), MDC=[ 100; 220), MED=[ 220; 450), MDL=[ 450; 900), L=[ 900; 1800) y ML=[ 1800; $\infty$) (en km)
- orientación ("OR"): SO=[ -157.5; -112.5), S=[ -112.5; -67.5), SE=[ -67.5; -22.5), E=[ -22.5; 22.5), NE=[ 22.5; 67.5), N=[ 67.5; 112.5), NO=[ 112.5; 157.5) y O=[ 157.5; 180] $\cup$ [-180; -157.5) (en grados)

**Tabla 1.** Variables y conjuntos de discretización para el caso de ejemplo ejecutado

| Variable | Descripción | Conjuntos de discretización (etiqueta = [media; desvío estándar]) |
|---|---|---|
| VOL | caudal de gas que midió la estación | VL=[ 7.91; 5.81], L=[ 24.81; 10.49], ML=[ 335.9; 190.59], MH=[ 1091.95; 349.71], H=[ 2004.96; 186.35], VH=[ 2540.13; 177.61] |
| HEAT_VAL | poder calorífico del gas | VL=[ 0; 4.21], L=[ 37.58; 0.05], ML=[ 37.64; 0.14], MH=[ 38.57; 0.05], H=[ 38.81; 0.52], VH=[ 38.96; 0.03] |
| PRESS | presión estática | L=[ 4.73; 4.2], M=[ 20.5; 0.59], H=[ 56.21; 4.78] |
| TEMP | temperatura del gas medido | L=[ 10.03; 4.88], M=[ 16.06; 1.53], H=[ 24.07; 3.62] |

Los valores de los parámetros utilizados en el algoritmo para el descubrimiento de las reglas de asociación son: $\sigma = 0.9$, $\tau = 120$ días, $\Delta t = 60$ días y $\theta = 0.95$.

Los siguientes son algunos ejemplos de reglas obtenidas con este caso de ejemplo que se destacan por su utilidad de uso para el usuario, en base al conocimiento aportado por el mismo respecto al dominio de aplicación actual de los datos.

```
PRESS_H(POS1) → TEMP_H(POS2)
[ { DIST_MED-OR_E(POS1,POS2) }; { [29/08/2003; 28/12/2003] } ]
[ { 0.96 }; { 0.96 } ]
```

La regla mostrada relaciona valores "altos" de presión con valores "altos" de temperatura. Estos últimos pueden ser una consideración de seguridad a tener en cuenta por el usuario. Teniendo en cuenta que el gas se transfiere elevándolo a presiones altas, se puede prestar atención a las características de la regla en los casos que sea necesario tener en cuenta la seguridad del gasoducto en base a la transferencia del gas.

```
VOL_VL(POS1), HEAT_VAL_L(POS2) → PRESS_L(POS3)
[ { DIST_MDC-OR_S(POS1,POS2), DIST_MDL-OR_N(POS1,POS3) };
{ [30/06/2003; 31/12/2003] } ]
[ { 0.9135 }; { 1 } ]
```

Esta regla presenta como antecedente valores "muy bajo" de volumen y "bajo" de poder calorífico. Además, los items de la misma se repiten en otras reglas con predicados similares (solo varia la orientación). Considerando que el producto de las dos variables del antecedente mide la energía transferida, estas reglas identificadas relacionan un valor bajo de energía con disminuciones en la transferencia del gas.

```
VOL_VL(POS1), PRESS_H(POS2), PRESS_M(POS3) → PRESS_L(POS4)
[ { DIST_MED-OR_SO(POS1,POS2), DIST_MED-OR_N(POS1,POS3),
DIST_MDL-OR_N(POS1,POS4) }; { [01/01/2003; 31/12/2003] } ]
[ { 1 }; { 1 } ]
```

En esta regla se observa, arrancando desde un valor "muy bajo" de volumen, la disminución de los valores registrados de presión, a medida que la relación espacial se orienta al norte y aumenta su distancia. De esta manera, se puede considerar como un patrón de relaciones espaciales que señala la disminución de los valores de presión registrados (que representan la transferencia del gas).

Debido a la cantidad de reglas obtenidas, es posible encontrar información útil en otras reglas; en base a un mejor conocimiento de uso, por parte del usuario, de los resultados provistos por la solución implementada, así como mediante el uso de aplicaciones de postprocesamiento de los resultados obtenidos.

**Pruebas de Performance.** Junto con el caso de ejemplo se ejecutaron otras pruebas para analizar la performance del algoritmo, respecto a los tiempos de ejecución en función del tamaño del dataset utilizado. Las pruebas se ejecutaron sobre un máquina con un procesador AMD Athlon™ 64 3000+ (1.81GHz), con una memoria RAM de 2GB. El sistema operativo fue MS Windows XP SP2. La versión de java fue la 5. La memoria de la jvm fue configurada a un máximo de 1,5 GB.

En esta prueba se analizo el tiempo de ejecución en función de la cantidad de timestamps. Para esto se tomaron conjuntos de transacciones del modelo contiguas y desde el principio del dataset. Se comenzaron con 150 transacciones del modelo, cuya cantidad se fue aumentando hasta llegar a 365. La cantidad de sensores seleccionados fue 6. Debido a que las curvas obtenidas no presentaban una tendencia observable de manera simple, se ejecutaron tomando tres datasets diferentes para poder verificar el comportamiento del algoritmo en la prueba; variándose entre los datasets los sensores y la cantidad de intervalos de discretización a elegir para cada atributo –para este caso se utilizo un filtro de discretización en igual ancho-. Estos datasets se seleccionaron de forma tal que sus transacciones tuvieran todas el mismo tamaño, para evitar que una diferencia en el mismo pudiera influir en las pruebas. Los parámetros dados de entrada al algoritmo son: $\sigma = 0.7$, $\tau = 90$ días, $\Delta t = 20$ días y $\theta = 0.8$. La figura 1 muestra gráficamente los tiempos de ejecución obtenidos en esta prueba.

Se puede observar en el gráfico de resultados una tendencia lineal, aunque con algunos desvíos en la misma. Estos desvíos en las curvas se pueden deber al uso de datos experimentales en las pruebas, y la selección de los mismos; ya que entre los diferentes datasets los desvíos son de diferente tipo. Además, hay que tener en cuenta que si bien se trato de seleccionar un conjunto de items *I* que sea lo mas parecido entre las diferentes ejecuciones de la prueba, siempre aparecieron diferencias respecto al tama-

ño de los conjuntos de discretización, las que habrán influido en los resultados obtenidos.

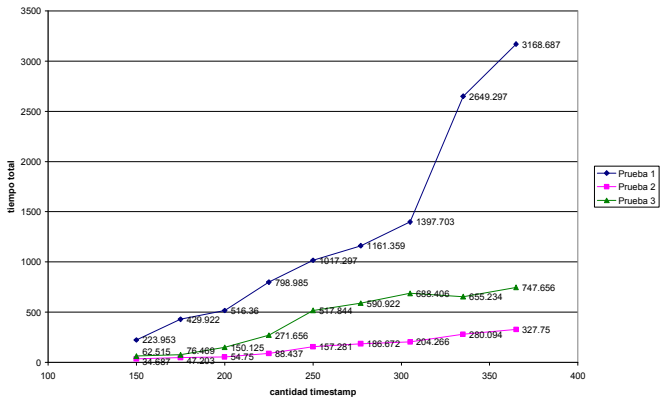


**Fig. 1.** Gráfico de resultados de pruebas de performance

## 6 Conclusiones y Trabajo Futuro

En este trabajo se desarrollo un modelo de reglas de asociación espacio-temporales, que permite obtener las mismas desde bases de datos con transacciones que presentan componentes espaciales y temporales. Este modelo propuesto trata de expresar de una manera conveniente los conocimientos, espacial y temporal, involucrados en las reglas, mediante la integración y extensión de conceptos planteados en distintos modelos seleccionados para el descubrimiento de reglas de asociación temporales, reglas de asociación espaciales y reglas de asociación inter-transacciones. El conocimiento espacial se representa relacionando todos los elementos de la regla entre sí, respecto a un mismo elemento, sin restringirse la relación espacial, y utilizando la notación de predicados espaciales, de uso común en los modelos de reglas de asociación espaciales. El conocimiento temporal se representa mediante intervalos de ocurrencia, los que dependen solamente de las características propias de los datos de entrada.

Se desarrollaron los algoritmos necesarios para el descubrimiento de las reglas de asociación espacio-temporales definidas por el modelo. Los mismos se implementaron en Java, como parte del sistema Weka [14]. Se analizó experimentalmente el comportamiento del modelo utilizando la implementación desarrollada sobre un dataset del área de transporte de gas, pudiéndose encontrar diversos patrones de utilidad de uso. Además se realizaron pruebas de performance con este dataset de uso real, que señalan un comportamiento lineal del algoritmo en el caso más convencional de uso del modelo –el aumento de la cantidad de registros con el avance del tiempo-.

Como trabajo futuro, se analizará la necesidad de mejoras de performance mediante el desarrollo de un nuevo algoritmo para el descubrimiento de itemsets frecuentes. Para el mismo, se aplicarán modificaciones similares a las realizadas sobre Apriori, en

otros algoritmos para el descubrimiento de itemsets frecuentes más eficientes que este, como por ejemplo FP-Growth ([3]).

Por otro lado, se considerará la extensión del modelo para que sea aplicable sobre datasets de objetos móviles.

## Referencias


1. Agrawal, R., et. al.: Mining Association Rules between Sets of Items in Large Databases. In: ACM SIGMOD Conference on Management of Data, pp. 207-216. ACM Press, Washington (1993)
2. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Int. Conf. on Very Large Data Bases, pp. 487-499, expanded version: IBM Research Report RJ9839. Morgan Kaufmann, Santiago (1994)
3. Han, J., et. al.: Mining Frequent Patterns without Candidate Generation. In: The 2000 ACM SIGMOD International Conference on Management of Data, pp. 1-12. ACM Press, Dallas (2000)
4. Adhikary J., et. al.: Knowledge Discovery in Spatial Databases: Progress and Challenges. Technical Report 96-08, SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. Canada (1996)
5. Roddick, J.F., et. al.: YABTSSTDMR - Yet Another Bibliography of Temporal, Spatial and Spatio-Temporal Data Mining Research. In: KDD 2001 Temporal Data Mining Workshop, pp. 167-175. ACM, San Francisco (2001)
6. Koperski, K., Han J.: Discovery of Spatial Association Rules in Geographic Information Databases. In: Egenhofer, M. J., Herring, J. R. (eds.) SSD 95. LNCS, vol. 951, pp. 47-66. Springer, Portland (1995)
7. Lee, Ickjai: Mining Multivariate Associations within GIS Environments. In: Orchard, R., Yang, C., Ali, M. (eds.) IEA/AIE 2004. LNCS, vol. 3029, pp. 1062-1071. Springer, Ottawa (2004)
8. Ale, J., Rossi, G.: The Itemset's Lifespan Approach to Discovering General Temporal Association Rules. In: Second Workshop on Temporal Data Mining in SIGKDD 2002, The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1-10. ACM, Edmonton (2002)
9. Ramaswamy, S., et. al.: On the Discovery of Interesting Patterns in Association Rules. In: 24th Int. Conf. Very Large Data Bases, pp. 368-379. Morgan Kaufmann, Nueva York (1998)
10. Lu, H., et. al.: Beyond Intra-Transaction Association Analysis: Mining Multi-Dimensional Inter-Transaction Association Rules. ACM Transactions on Information Systems, vol. 18, issue 4, pp. 423-454. ACM, Nueva York (2000)
11. Mennis, J., Liu, J. W.: Mining Association Rules in Spatio-Temporal Data. In: 7th International Conference on Geocomputation. Southampton (2003)
12. Gidófalvi, G., Bach Pedersen, T.: Spatio–temporal Rule Mining: Issues and Techniques. In: Min Tjoa, A., Trujillo, J. (eds.) DaWaK 2005. LNCS, vol. 3589, pp. 275-284. Springer, Copenhague (2005)
13. Verhein, F., Chawla, S.: Mining Spatio-Temporal Association Rules, Sources, Sinks, Stationary Regions and Thoroughfares in Object Mobility Databases. In: Lee, M. L., Tan, K. L., Wuwongse, V. (eds.) DASFAA 2006. LNCS, vol. 3882, pp. 187-201. Springer, Singapur (2006)
14. Witten, I., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco (2005)
15. GeoTools. The open source Java GIS toolkit, http://geotools.codehaus.org/. Acceso: 03/05/2008