Hacia un Modelo Integrado de Procesamiento de Flujos de Datos

Mario Diván[1 2], **Luis Olsina**[2], **Hernán Molina**[2]
[1] Facultad de Ciencias Económicas y Jurídicas,
[1 2] GIDIS_Web, Facultad de Ingeniería
Universidad Nacional de La Pampa
[mjdivan, olsinal, hmolina]@ing.unlpam.edu.ar / mjdivan@eco.unlpam.edu.ar

Resumen. *El presente paper presenta un modelo integrado de procesamiento de flujos de datos con el fin de mejorar la toma de decisiones basada en contextos mediante la incorporación de metadatos basados en una ontología de medición. En particular se discute la recolección-adaptación de datos dentro del modelo integrado de procesamiento, y se aborda la problemática de la definición de un esquema para el intercambio continuo de mediciones basadas en un marco conceptual de medición y evaluación, como así también el protocolo asociado a la transmisión de las mismas. Dicho esquema y protocolo, permiten el intercambio de metadatos vinculados a mediciones y sus contextos asociados, con el objeto de permitir un análisis consistente de los mismos que contribuya a una mejora en la toma de decisión susceptible al contexto.*

Palabras clave: *Flujos de Datos, CINCAMI/MIS, CINCAMI/MIP, Procesamiento de Flujos de Datos, Clasificación, XML, Métricas, Contexto.*

1. Introducción

En la actualidad, un nuevo tipo de aplicaciones intensivas en el procesamiento de datos, tales como aquellas aplicadas al procesamiento de señales y sensores en general donde es menester disponer de una respuesta inmediata, requieren un tratamiento diferenciado del enfoque tradicional de análisis de datos basado en persistencia [1]. Este nuevo tipo de aplicaciones necesita procesar un conjunto de datos en forma continua, a medida que se generan; de manera que el mismo sistema sea capaz de poder tomar decisiones o ajustar modelos de soporte a la toma de decisiones. Dichos modelos, generados al momento en que el dato arriba y es procesado, se aplican sobre los datos actuales a los efectos de obtener la decisión, dado que carecen de tiempo y/o recursos para un procesamiento secundario en un entorno de persistencia [2].

Dentro del tipo de aplicaciones mencionadas en el párrafo anterior, se encuentran aquellas que procesan datos de mediciones y basan todo o parte de su comportamiento en función de los mismos. En este contexto, el marco conceptual de medición y evaluación C-INCAMI (*Context* - *Information Need*, *Concept model*, *Attribute*, *Metric* and *Indicator*) establece una ontología y estructura que incluye los conceptos y relaciones necesarias, a los efectos de dar forma a cualquier proyecto de medición y evaluación [3,4] que se desee llevar a cabo. A partir de este marco, se ha desarrollado una herramienta web denominada INCAMI_Tool [3], la cual permite la gestión integral de metadatos asociados a uno o más proyectos de medición, almacenando a tal fin los mismos en un repositorio XML basado en el modelo ontológico C-INCAMI.

Dentro del ámbito del marco conceptual C-INCAMI y el tipo de aplicaciones mencionadas, el presente artículo pretende plantear un modelo integrado de procesamiento de flujos de datos que, mediante la incorporación de metadatos sustentados en la ontología de medición definida en C-INCAMI, permita una corrección y análisis consistente de los datos a los efectos de incrementar la precisión en el proceso de toma de decisión basado en el contexto de medición. En este último sentido y planteado el modelo integrado de procesamiento, el artículo se focalizará en establecer la estructura (bajo la forma de un esquema XML) y el protocolo de comunicación para el intercambio de mediciones en el marco de los procesos de recolección y análisis del modelo integrado de procesamiento.

El presente artículo se organiza en seis secciones. La sección dos aborda el modelo integrado de procesamiento de flujos de datos y dentro de sus sub-secciones, se expondrán las funcionalidades asociadas a cada elemento del modelo. La sección tres planteará el esquema para el intercambio de mediciones basadas en C-INCAMI (CINCAMI/MIS). La sección cuatro planteará el protocolo para el intercambio de las mediciones basadas en C-INCAMI (CINCAMI/MIP). La sección cinco expondrá los trabajos relacionados y por último, la sección seis esboza algunas conclusiones y trabajos a futuro.

2. Modelo Integrado de Procesamiento de Flujos de Datos

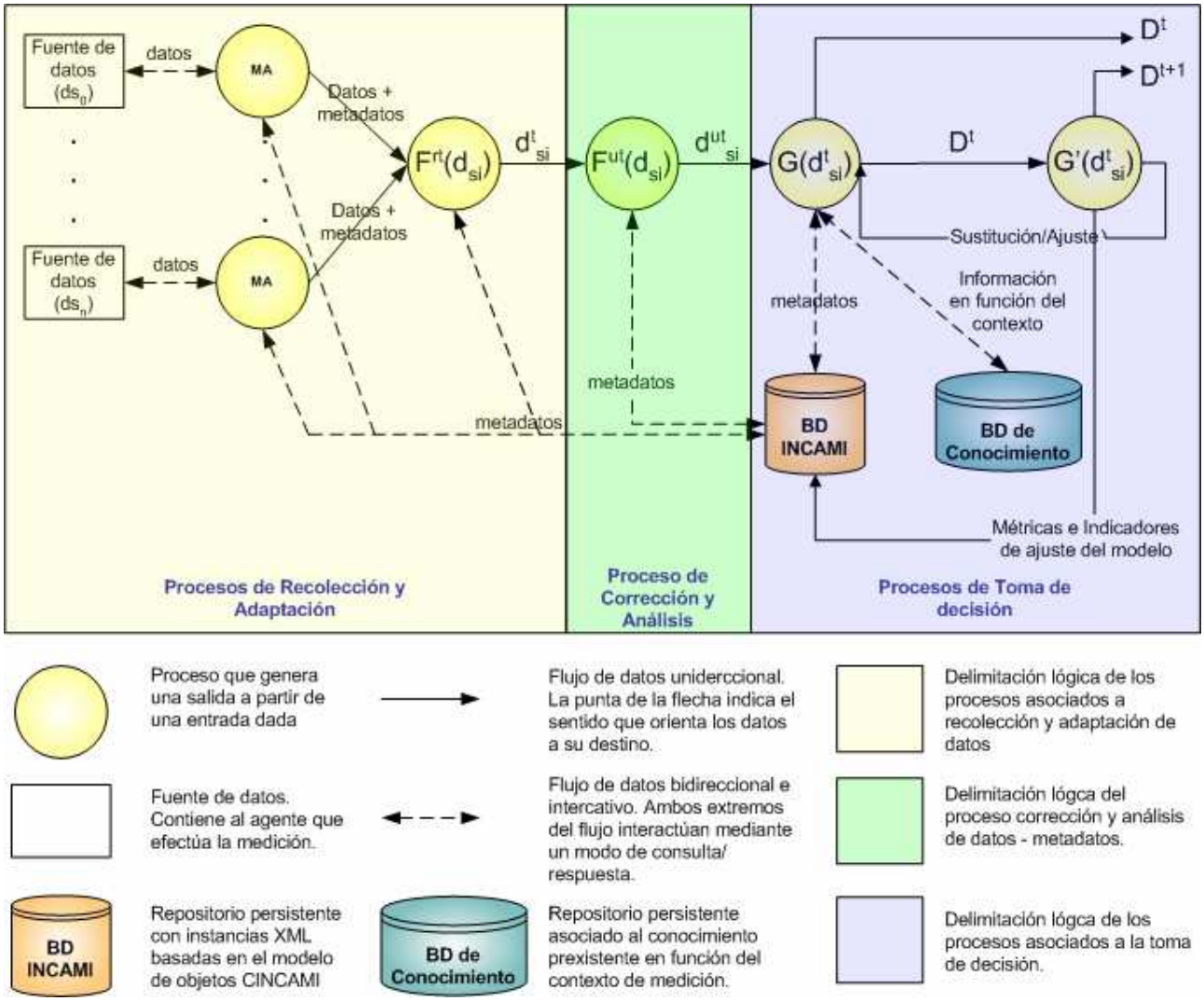
2.1 Motivación

La idea de procesar los datos continuamente ante su arribo, "al vuelo", implica una clara diferenciación con respecto a los sistemas de gestión de bases de datos (SGBD) [5] según el enfoque tradicional. A los efectos de diferenciar el contexto de datos tradicional y el denominado como flujos de datos (data streams), se podría tener en cuenta, por ejemplo, el problema de la *detección de fraudes* en las empresas de telefonía [6]. Dichas empresas desearían seguramente poder detectar los potenciales fraudes al momento en que se están por producir, o bien localizarlos mientras ocurren, para de este modo poder minimizar las posibles pérdidas asociadas. Esto último motiva a la incorporación de técnicas provenientes del campo de la minería de datos sobre el área de flujos de datos –mining data streams / stream mining- a los efectos de medir y evaluar el comportamiento típico sobre los flujos de datos asociados a las comunicaciones y poder construir modelos de decisión ajustables "al vuelo" capaces de interpretar y detectar desvíos en las métricas e indicadores definidos. Precisamente, la motivación en usar el marco de medición y evaluación C-INCAMI, es que permite especificar los datos y metadatos de las métricas e indicadores en cuestión, además de las propiedades de contexto.

Por otro lado, si se toma un entorno basado en modelos tradicionales de persistencia de datos, aún haciendo empleo de técnicas de minería de datos, se estarían analizando datos históricos y en caso de que se pudiese localizar un fraude, esto sólo representaría un hecho pasado y en donde el costo asociado al mismo se habrá tornado en una pérdida. Así la idea de adaptar técnicas de clasificación [7] a flujos de datos semi estructurados enmarcados dentro de las especificaciones del marco, se asocia a generar y ajustar los modelos de modo que eviten o minimicen dichas pérdidas, sin necesidad de analizar datos históricos sino por el contrario, analizar datos en el mismo momento en que se generan o arriban, pudiendo adaptar los modelos y tomar decisiones más robustas en base al contexto vinculado con los datos.

2.2. Modelo Integrado de Procesamiento de Flujos de Datos

Conceptualmente, la idea en términos de procesamiento (ver figura 1) es la siguiente. A partir del ingreso de los datos y metadatos a una función $F^{rt}(d_{si})$ –que tendrá por objeto la reunión de los flujos en función de las métricas asociadas- se transmiten a una función $F^{ut}(d_{si})$ –que tendrá por objeto suavizar las mediciones vinculadas a cada métrica-. Suavizados los datos, se aplicará el modelo actual de clasificación $G(d^t_{si})$ en base al marco de métricas e indicadores definidos junto con el conocimiento previo almacenado, produciendo una decisión en tiempo t, a la cual se denomina D^t. Luego, el modelo actual de clasificación se *ajusta y/o sustituye* en base a los nuevos datos y situación contextual para producir la decisión en tiempo t+1 con el nuevo modelo, permitiendo la comparación de decisiones entre D^t y D^{t+1}. Ambas decisiones permitirán pro-activamente disparar alarmas, notificaciones, ajustes y/o lo que el usuario de C-INCAMI haya definido en términos de indicadores, a los efectos de detectar las desviaciones on-line.



En donde:

MA	Measurement Adapter (Adaptador de Mediciones)
d_{si}	Flujo de datos (data set) 'i'
$F^{rt}(d_{si})$	Función de reunión de d_{si} (Join Function)
$F^{ut}(d_{si})$	Función de transformación o suavizado del d_{si} unificado
d^t_{si}	Flujo de datos 'i' unificado
d^{ut}_{si}	Flujo de datos 'i' unificado y suavizado
$G(d^t_{si})$	Aplicación del modelo de clasificación actual (por ej. árbol)
D^t	Decisión en tiempo 't' obtenida mediante el modelo actual
D^{t+1}	Decisión en tiempo 't+1' obtenida mediante el modelo ajustado
$G'(d^t_{si})$	Sustitución/Ajuste del modelo actual a partir de la generación del nuevo modelo.
KPI	Key Process Indicador (Indicador de proceso clave)
Knowledge DB	Base de datos de conocimiento

Figura 1: *Esquema del modelo integrado de procesamiento de flujos de datos*

2.3 Fuentes de Datos y el Adaptador de Mediciones

Se entiende por *fuentes de datos* –data sources- a cualquier origen plausible de generar nuevos datos de un modo continuo e ilimitado. Se entiende por *nuevos datos* a aquellos que no existen previamente en la memoria de procesamiento en el momento exacto en que arriban a la función de transformación $F^{ut}(d_{si})$ (ver sección 2.5); pero en caso de que existiesen, su arribo aporta algún elemento adicional que permite incorporar un nuevo punto de vista al análisis automático de

datos. Las fuentes de datos para el contexto de procesamiento en continuo se asumen como *ilimitadas*, que a su vez pueden aprovisionar los datos de a *ráfagas*.

Las datos se procesan según el concepto de *ventanas* –windows- [8] surgido a partir de la idea de data streams. El procesamiento por ventanas puede ser en dos modos: el primero implica una *ventana en función temporal* que procesará tantos registros como arriben en un período "p"; mientras que el segundo implica una *ventana en función de hitos*; esto es, aquella que procesará tantos datos como existieran dentro de las ventanas definidas por dichos hitos. De este supuesto se desprende como corolario que *los flujos de datos son parcelizables*.

El *Adaptador de Mediciones* (*Measurement Adapter*) es una pieza de software configurable capaz de interactuar con la fuente de datos y ser configurado respecto de la o las métricas directas a medir, las condiciones contextuales a considerar, el acceso al repositorio de instancias XML basado en el modelo CINCAMI y discernir si la fuente de datos [9,10] ha brindado un valor determinista o bien indicado un valor o valores predictivos junto con sus probabilidades asociadas [11]. El Adaptador de Mediciones tiene por objeto actuar de interface entre la fuente de datos y la función de reunión (ver sección 2.4), a los efectos de incorporar a los datos obtenidos desde la fuente de datos, los metadatos vinculados a la métrica directa en medición más información temporal y contextual asociada. El cómo estructurará los datos y metadatos a transmitir y de qué modo lo hará son el objeto de discusión de las secciones tres (CINCAMI/MIS) y cuatro respectivamente (CINCAMI/MIP). De este modo, los datos no arriban conceptualmente en forma unitaria a la función de reunión, sino que vendrán acompañados adicionalmente, de metadatos que permitirán discernir entre su significado [2]. Al igual que ocurre en el modelo relacional de datos, es posible que no se disponga de valor para un atributo determinado en una medición dada. Además de la ausencia de valor para un atributo en la medición, debe considerarse como restricción que es factible que las líneas de comunicación de datos se dañen y se interrumpa el flujo de datos y metadatos total o parcialmente a la función de reunión.

2.4 Función de Reunión (Join Function)

La función $F^{rt}(d_{si})$ tiene asociada la responsabilidad de informar el estado actual de procesamiento a los adaptadores de mediciones (on-line u off-line) y reunir los flujos de datos, unificarlos ordenándolos dentro del flujo de acuerdo a la métrica a la cual se asocia la medición en primer lugar y la fecha-hora de medición en segundo lugar. La función cuenta con acceso al repositorio XML basado en el modelo de objetos CINCAMI para consultar información adicional sobre los metadatos informados desde el adaptador de mediciones. Esta etapa es claramente una actividad en donde el procesamiento paralelo tendría notables ventajas en materia de rendimiento [12], motivo por el cual tal procesamiento se efectuará de dicho modo al momento de su implementación.

2.5 Función de Suavizado

La función $F^{ut}(d_{si})$ tiene asociada dos responsabilidades. En primer lugar, debe dar coherencia al orden de procesamiento y en segundo lugar, debe suavizar [13,14] cada uno de los flujos de datos acorde al procesamiento por ventanas indicado anteriormente.

La coherencia en el orden de procesamiento estará en función de los metadatos que arriben en forma conjunta con los datos, los cuales permiten interpretar automáticamente y basado en CINCAMI, el significado de los datos y su relación con las restantes entidades del proyecto de medición. Por otro lado, la suavización de los datos tiene por objeto identificar y resolver inconvenientes propios del dato original tal como el ruido, outliers y ausencia de valor.

2.6 Función de Decisión

La función $G(d^{t}_{si})$ de decisión, tomará el orden de procesamiento y los datos suavizados para aplicar el modelo de clasificación actual [15] y así obtener la decisión en el tiempo "t". Esta etapa

comunica los datos en su estado original, el orden de procesamiento y la decisión en tiempo "t" a la etapa asociada a la generación del nuevo modelo de clasificación. La etapa de generación del nuevo modelo de clasificación permitirá retroalimentar el comportamiento de la función de decisión $G(d^t_{si})$, brindándole información sobre el nuevo modelo de clasificación generado junto con la decisión "t+1".

2.7 Función de Sustitución/Ajuste del Modelo de Clasificación

La función $G'(d^t_{si})$ recibirá la decisión en tiempo "t", el orden de procesamiento y los datos en su estado original [16] para la aplicación de diferentes técnicas, a los efectos de ajustar y/o sustituir el modelo de clasificación actual.

El contexto de los datos como su naturaleza, influyen empíricamente en el modelo de clasificación a adoptar y/o ajustar. Cuando se habla de su influencia empírica, se hace referencia a la utilización de la base de datos de conocimiento con los modelos históricos asociados a cada uno de sus contextos, como así también a los datos de las mediciones, dentro de la base de datos INCAMI para incorporarse dentro del proceso de ajuste/sustitución.

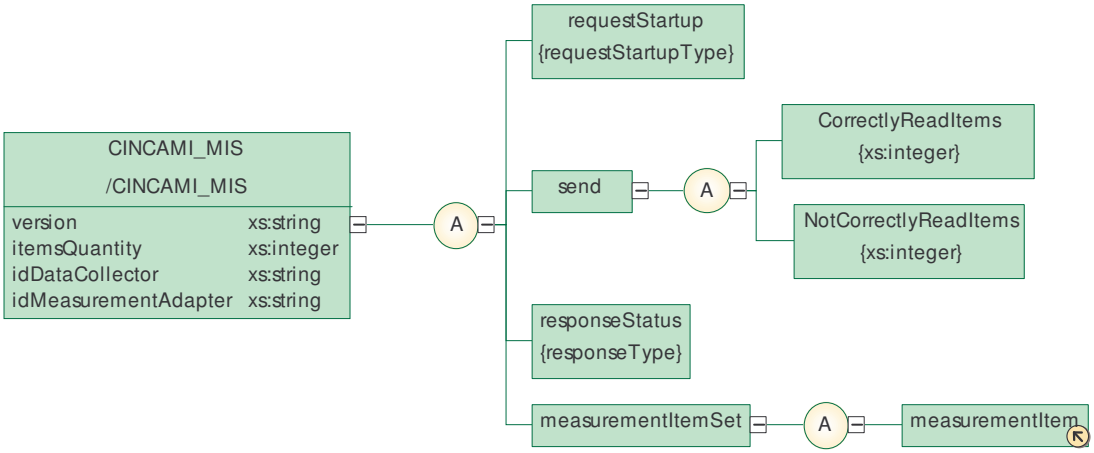
Existen numerosas técnicas para obtener nuevos modelos de clasificación que van desde los tradicionales árboles hasta complejas redes neuronales [17]. Cabe resaltar que no es simple indicar qué técnica es mejor que otra, debido a que las mismas son sensiblemente dependientes del contexto de aplicación, así como de la naturaleza de los datos. Esto último representa un desafío no menor a los efectos de automatizar el proceso de generación del modelo de clasificación y de la decisión de *cuál de los modelos obtenidos aplicar* con vistas a sustituir/ajustar el modelo de la etapa "t" [18]. Para esta investigación, se ha decidido acotar la función de clasificación sólo a técnicas y métodos asociados a árboles a los efectos de no extender el límite de estudio.

Una vez obtenido el nuevo modelo de clasificación, éste se comunica a la función $G(d^t_{si})$ junto con la decisión en tiempo "t+1", para actualizar el modelo y así ser empleado en el procesamiento de las ulteriores ventanas. Se entiende que la decisión en tiempo "t+1" se obtiene mediante la aplicación del nuevo modelo de clasificación. Finalmente, se compara la semántica de las decisiones producidas en tiempo "t" versus "t+1" y aquella que mejor se adecue globalmente a las condiciones actuales del contexto, será la adoptada. Esto último presenta interrogantes no tan simples de responder, como por ejemplo: ¿Cómo saber cuál decisión es la que mejor se adecua al contexto actual? ¿Qué parámetros regirán el comportamiento del contexto? Posibles soluciones a estos desafíos serán ampliados en trabajos futuros.

3. Esquema de Intercambio de Mediciones Basado en C-INCAMI (CINCAMI/MIS)

El esquema de intercambio de mediciones (Measurement Interchange Schema - MIS) basado en C-INCAMI (CINCAMI/MIS), tiene por objeto definir la estructura y tipología de los datos a intercambiar entre el adaptador de mediciones y la función de reunión (ver sección 2.4) dentro de los procesos de recolección y adaptación del modelo integrado de procesamiento de flujos (ver figura 2). Los conceptos empleados dentro de la estructura del mensaje CINCAMI/MIS se encuentran claramente definidos en términos ontológicos dentro de C-INCAMI [3], lo que permite una identificación clara del significado asociado a un valor dado en cada una de las componentes de la medición.

Tanto el adaptador de mediciones como la función de reunión, tienen acceso al repositorio XML basado en el modelo C-INCAMI [3, 4]. A partir de éste último, pueden obtener datos y metadatos del proyecto de medición mediante CASTOR [19], un framework de vinculación de datos para JAVA.



Donde:

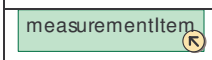
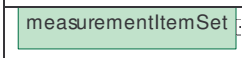
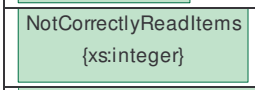
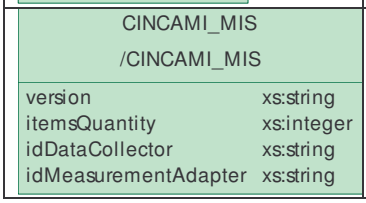
	Implica que la etiqueta superior está compuesta por todas las etiquetas unidas por líneas a partir del signo "-", *sin importar el orden de aparición de las mismas ya que son perfectamente conmutativas*.
	Implica que la etiqueta superior está compuesta por todas las etiquetas unidas *secuencialmente* por líneas a partir del signo "-". El orden está indicado de arriba hacia abajo a partir del símbolo "-" y no es conmutativo, esto implica que para el procesamiento las etiquetas deben arribar en el orden indicado.
	Las líneas continuas establecen una dependencia jerárquica. La etiqueta de la izquierda se compone de la etiqueta o conjunto de etiquetas unidas por líneas continuas a su derecha a través del símbolo "-".
	El contenido de la etiqueta es explicado en forma separada del actual esquema.
	Etiqueta compuesta por sub-etiquetas. No es una etiqueta de último nivel u hoja.
	Etiqueta compuesta por sub-etiquetas, en donde se especifican atributos junto con el tipo de datos asociado.
	Etiqueta de último nivel u hoja. Encierra entre llaves el tipo de datos que se le asocia.
	Etiqueta principal o raíz del esquema (CINCAMI_MIS). Luego del nombre de la etiqueta y separado por la línea horizontal, se especifican sus atributos junto con su tipo de datos asociado. El orden de especificación de los atributos no tiene importancia y es conmutativo.

Figura 2: *Vista resumida del esquema CINCAMI/MIS*

La etiqueta principal del mensaje, denominada CINCAMI_MIS, además de delimitar completamente el mensaje, posee atributos cuya función es la siguiente:

- **version**: Representa la versión del esquema CINCAMI/MIS que utilizará el adaptador de mediciones para el intercambio de las mismas. En la actualidad sólo existe una versión, por lo que por defecto es "1.0".
- **itemsQuantity**: Cantidad de ítems de mediciones a intercambiar o transmitir. Este atributo puede ser 0 (cero) en aquellos mensajes CINCAMI/MIS asociados a tareas de coordinación entre el adaptador de mediciones y la función de reunión. Esto último podrá ser apreciado en la sección cuatro referida a CINCAMI/MIP.
- **idDataCollector**: Identificador del agente que ejecuta el método de medición y obtiene la misma.
- **idMeasurementAdapter**: Identificador del adaptador de mediciones que recolecta, adapta y transmite las mediciones ante la función de reunión.

A los efectos de simplificar la terminología, diremos que los mensajes que el adaptador de mediciones envíe a la función de reunión se los conocerá como "*mensajes de solicitud*", mientras que aquellos mensajes transmitidos desde la función de reunión al adaptador de mediciones se los referenciará como "*mensajes de respuesta*". De este modo, se verá que algunas de las componentes del mensaje CINCAMI/MIS se asociarán solo a los mensajes de solicitud, mientras que otras sólo se vincularán a los mensajes de respuesta. La etiqueta CINCAMI_MIS de este modo se integra por:

- **requestStartup**: Es una etiqueta opcional y sólo está presente en mensajes CINCAMI/MIS de solicitud. Esta puede asumir sólo dos valores (*startup* o *reverse*) los cuales serán analizados en forma conjunta con CINCAMI/MIP en la siguiente sección. Su función básica es coordinar el inicio de la transmisión de mediciones desde el adaptador de mediciones.
- **send**: Es una etiqueta opcional y sólo está presente en mensajes CINCAMI/MIS de respuesta. Su función es indicar a la función de reunión cuántos ítems han sido correctamente leídos (C*orrectlyReadItems*) y cuales no han podido ser leídos completamente o han presentado inconvenientes en alguno de sus componentes (*NotCorrectlyReadItems*).
- **responseStatus**: Es una única etiqueta opcional y sólo está presente en mensajes CINCAMI/MIS de respuesta. Su función es indicar el estado actual de la función de reunión, el cual puede ser online u offline.
- **measurementItemSet**: Es una etiqueta opcional y sólo está presente en mensajes CINCAMI/MIS de solicitud. Esta etiqueta tiene por objeto agrupar al conjunto de mediciones tomadas por el adaptador de mediciones a los efectos de la transmisión. Es importante aclarar que las mediciones (*measurementItem*) dentro del conjunto no respetan necesariamente un orden de generación.

La etiqueta *measurementItem* (ver figura 3) es el eje central de CINCAMI/MIS, ya que dentro del mismo se transmiten los datos de la medición y el contexto donde se efectuó la misma. Esta etiqueta se integra por las siguientes sub-etiquetas:

- **idEntity**: Identificador de la entidad objeto de medición.
- **idMetric**: Identificador de la métrica a la cual corresponde la medición.
- **Measurement**: Datos asociados a la medición en sí misma.

Una medición puede ser actual/real o estimada. En caso de que sea actual, el valor indicado será determinado con un margen de precisión dado e indicado en el proyecto de medición bajo C-INCAMI. En caso de que sea estimado, esto podría darse porque un valor no es determinísticamente obtenible, el valor tendrá asociado una etiqueta vinculándolo a su frecuencia empírica (rate) y un atributo de agrupamiento denominado *idMeasure*.

El tipo de medición es indicado a través de la etiqueta *measureTypeField,* el cual es obligatorio en caso de que exista medición. La etiqueta *value* es obligatoria y su semántica varía en función del tipo de medición. Si el tipo de medición es actual, value representará un valor determinístico, mientras que si el tipo de medición es estimated, value representará una estimación cuya frecuencia empírica se indica obligatoriamente mediante la etiqueta *rate.* Puede ocurrir que el adaptador de mediciones ante la estimación de un valor para un contexto de medición y métrica dada, posea como resultado de la medición a un conjunto de valores estimados con sus probabilidades respectivas en lugar de tan solo un valor y su probabilidad asociada [11], en este caso y a los efectos de señalar a la función de reunión que se trata de ésta situación, se informan estos datos como etiquetas *measurement* independientes vinculados a través de un mismo valor en *idMeasure* donde éste último es generado por el adaptador de mediciones en función de la fuente de datos asociada.

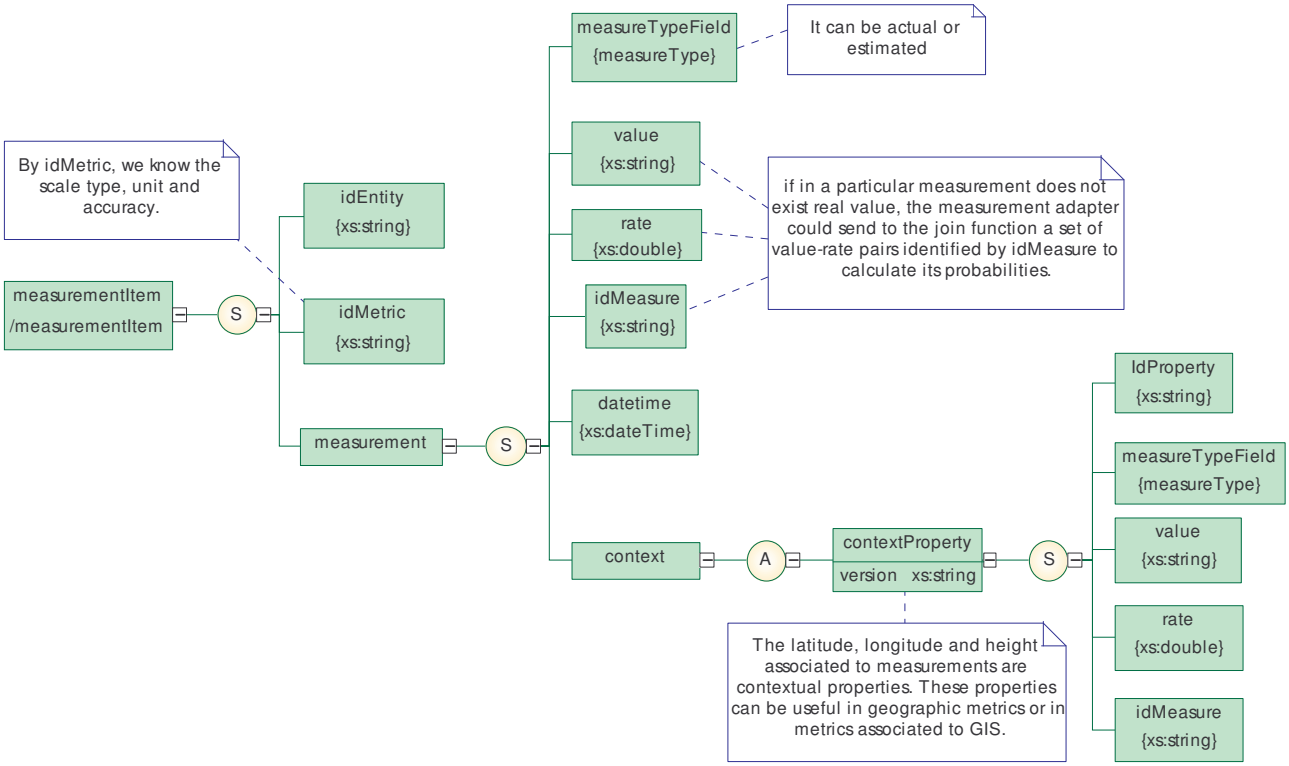


Figura 3: *Detalle de la etiqueta measurementItem*

El campo *datetime* dentro de la etiqueta *measurement*, representa la fecha y hora en que se toma la medición, mientras que la etiqueta compuesta *context* agrupará cada una de las propiedades de contexto en el que se toma la medición.

Cada propiedad de contexto [4] es representada por la etiqueta *contextProperty*, la cual posee un identificador de propiedad de contexto (*idProperty*), el tipo de valor que contiene la propiedad (*measureTypeField*: al igual que las mediciones podrá ser actual o estimated) y el valor asociado a la medición (*value)*. Este último elemento, puede ser actual o estimado en función de lo indicado en *measureTypeField. En* caso de ser estimado, se le asocia la frecuencia empírica dada por el elemento *rate* y en caso de que se presente una medición estimada con múltiples valores-probabilidad, podrá asociar las mismas mediante el elemento idMeasure a nivel de la propiedad del contexto.

4. Protocolo para el Intercambio de Mediciones (CINCAMI/MIP)

El protocolo para el intercambio de mediciones (*Measurement Protocol Interchange* - MIP) basado en C-INCAMI (CINCAMI/MIP), tiene por objeto definir los pasos necesarios para llevar adelante la transmisión de las mediciones entre el adaptador de mediciones (*Measurement Adapter* - MA) y la función de reunión, como así también las operaciones necesarias para coordinar a ambos en caso de que hubiese existido alguna interferencia durante la transmisión (ver figura 4).

Cuando cada MA desea transmitir por primera vez, envía un mensaje en CINCAMI/MIS en donde indica en la etiqueta *requestStartup (nivel 1)* el valor "**startup**". La función de reunión responde a este mensaje indicando en el elemento *responseStatus (nivel 1)* si se encuentra "**online**" u "**offline**" a los efectos de receptar mediciones. Si durante el envío del mensaje de startup ocurriese un timeout o algún tipo de excepción, el MA reenvía nuevamente el mensaje indicando el valor "startup" hasta tanto la función de reunión retorne alguno de los dos valores indicados. Solo en caso de que el MA reciba en responseStatus el valor "**online**" podrá transmitir las mediciones.

Cada vez que el MA envía un mensaje de solicitud con mediciones (**send**), recibirá en responseStatus el estado actual de la función de reunión.

De este último modo, si la función de reunión indica como respuesta a un mensaje **send** el valor **"offline"** en responseStatus o bien ocurriese un timeout/excepción, el MA deberá coordinar el intercambio mediante el envío de un mensaje de solicitud en donde requestStatus tendrá el valor "**startup**" hasta tanto la función de reunión responda efectivamente con el valor **"online"** en responseStatus.

El comportamiento del MA, en términos de envío/recepción de mediciones a la función de reunión, dependerá del estado actual en que éste se encuentre. Básicamente se podrá encontrar en tres estados (ver figura 5):

1. **startup**: El MA se encuentra inicializado y listo para transmitir
2. **sending:** El MA se encuentra transmitiendo mediciones a la función de reunión

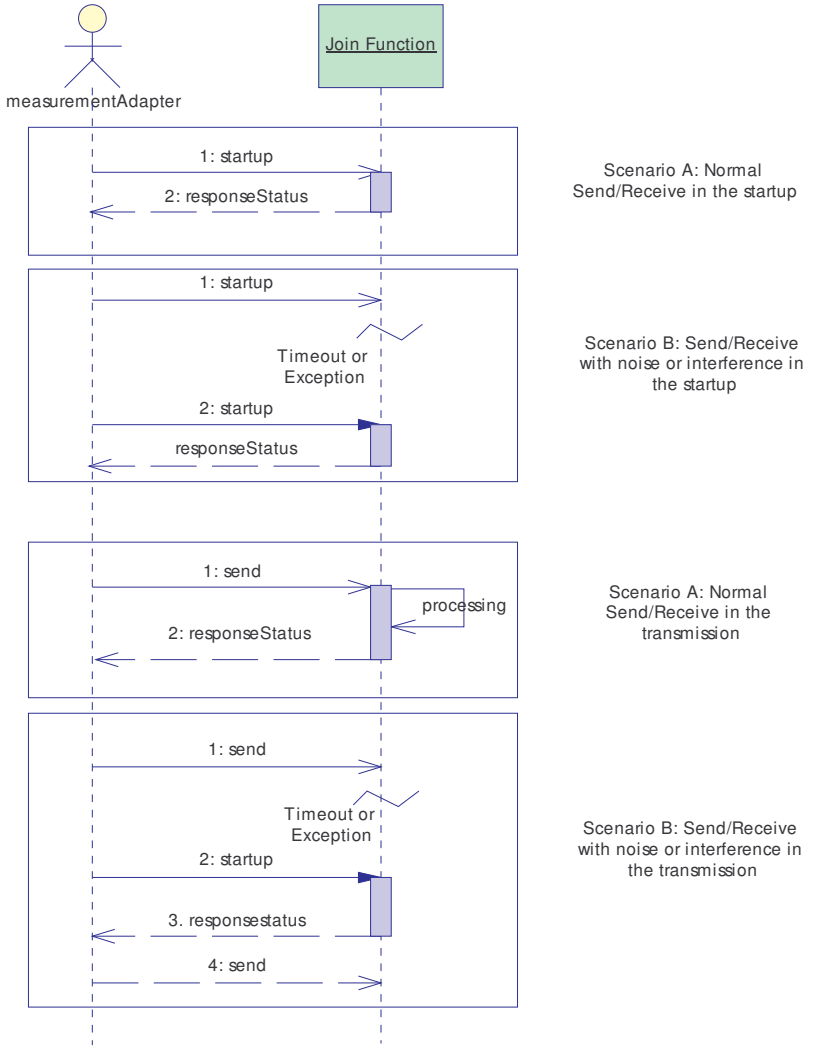


Figura 4: *Escenarios posibles para las fases de startup y transmisión*

3. **waitingStartup**: El MA ha señalado su deseo de inicializar la transmisión de mediciones a la función de reunión y ésta aún no ha respondido o bien, ha indicado offline.. En forma análoga para la función de reunión, su respuesta a los diferentes MA estará en función de su estado actual (ver figura 6)

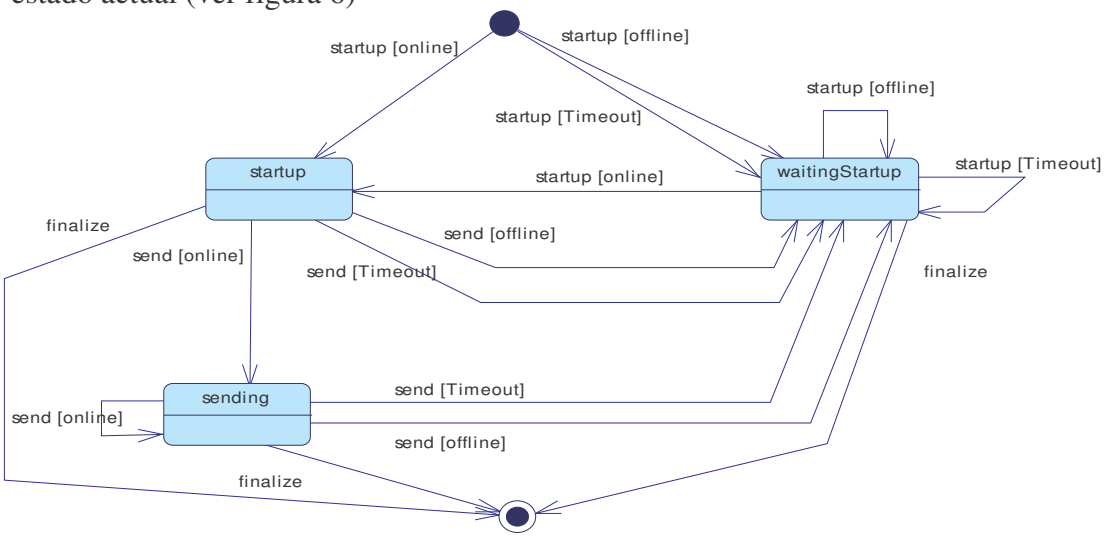
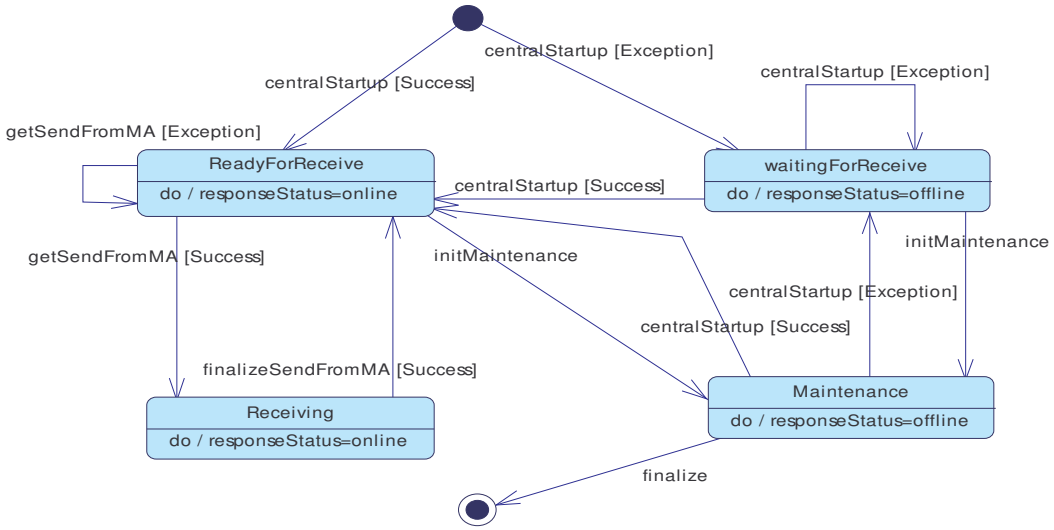


Figura 5: *Diagrama de estados para el adaptador de mediciones asociado al envío/respuesta de flujos CINCAMI/MIS*

En forma análoga al comportamiento del MA, el comportamiento de la función de reunión dependerá de su estado actual los cuales pueden ser (ver figura 6):

1. **WaitingForReceive**: La función de reunión ha sufrido algún inconveniente al intentar inicializarse y se mantiene a la espera de una correcta inicialización.
2. **ReadyForReceive**: La función de reunión ha sido inicializada correctamente y se encuentra lista para recibir mediciones desde los MA
3. **Receiving**: la función de reunión se encuentra recibiendo mediciones desde los MA
4. **Maintenance**: la función de reunión se encuentra abocada a tareas de mantenimiento propias tales como re-configuración, ajuste de parámetros de procesamiento, entre otras.



En donde:

centralStartup	Su responsabilidad es inicializar los recursos necesarios para que la función de reunión sea capaz de procesar en línea las mediciones informadas desde los MA
getSendFromMA	Su responsabilidad es tomar un flujo de mediciones desde el MA y comenzar a procesarlo. Puede recibir tanto flujos de mediciones como de coordinación
finalizeSendFromMA	Su responsabilidad es notificar al MA cuando se ha culminado el procesamiento del flujo de datos e indicar el estado actual de la función de reunión
initMaintenance	Su responsabilidad es indicar el inicio del proceso de mantenimiento, llevar adelante el proceso en sí y notificar cuando éste culmine a través de centralStartup.
finalize	Su responsabilidad es finalizar la escucha desde los MA y liberar los recursos reservados

Figura 6: *Diagrama de estados para la función de reunión asociado a la recepción/respuesta de flujos CINCAMI/MIS*

5. Trabajos Relacionados

En cuanto al procesamiento de flujos de datos, existen trabajos relacionados tales como los llevados adelante por el proyecto STREAM de la Universidad de Standford [20,21,22] y el proyecto Aurora desarrollado en colaboración entre el MIT y las Universidades de Brown y Brandeis [23,24,25]. Básicamente, los proyectos han abordado desde diferentes ópticas, los tópicos y problemáticas de las operaciones tradicionales de bases de datos dentro del contexto de procesamiento de data streams. En este sentido, numerosos enfoques efectuados por ambos proyectos son útiles en las funciones de reunión de datos dentro del proceso de corrección y análisis del modelo integrado de procesamiento de flujos.

En el campo de stream mining y en particular, con respecto a funciones de clasificación y agrupamiento sobre data streams, existen trabajos de Yu y otros de Watson Research Center - IBM directamente vinculados, tales como [26, 27, 28], entre otros. Estas investigaciones contribuyen y nutren los procesos de toma de decisión del modelo integrado de procesamiento.

Singh et al [29] han efectuado enfoques interesantes desde el punto de vista del empleo de minería de datos adaptable mediante factores contextuales. Esto último se encuentra directamente vinculado con los aspectos de generación del modelo de clasificación susceptible al contexto dentro de los procesos de toma de decisión del modelo integrado de procesamiento planteado.

Si bien existen enfoques puntuales en las áreas específicas como los mencionados, nuestra propuesta incorpora a la medición un conjunto de metadatos sustentados en una ontología claramente definida en C-INCAMI. Estos metadatos permiten incorporar información precisa sobre las definiciones de métricas, tipos de escala, unidades, métodos, entre otros conceptos, que agregan valor al dato, permitiendo a la función de reunión discernir claramente su significado, ayudando notoriamente a que el proceso de corrección y análisis sea efectuado de un modo consistente. El hecho de incorporar consistencia en el proceso de corrección y análisis, repercutirá necesariamente y positivamente en la función de decisión $G(d^t_{si})$ y en la función de sustitución/ajuste $G'(d^t_{si})$, lo que producirá modelos más precisos y adecuados a las condiciones contextuales de la medición a la hora de la toma de decisión.

6. Conclusiones y Trabajo Futuro

El presente trabajo ha planteado la visión del modelo de procesamiento integrado de flujos de datos y se ha focalizado en la sección referida a los procesos de recolección y adaptación de mediciones. Este modelo de procesamiento, sustentado en una ontología de medición claramente definida en C-INCAMI, permite incorporar significado a las mediciones y su contexto, con el objeto de lograr un proceso de corrección y análisis de datos consistente. Dicha consistencia, permitirá que los procesos de toma de decisión se encuentren mejor ajustados al contexto de medición y su decisión por ende, dispondrá de mayor precisión en términos contextuales.

La incorporación de metadatos dentro de las mediciones y su contexto asociado, se ha planteado dentro del esquema CINCAMI/MIS, con el objeto de conseguir un intercambio consistente de las mediciones y su contexto. Este intercambio, se sustenta en el protocolo CINCAMI/MIP a los efectos de coordinar coherentemente el flujo de datos entre el adaptador de mediciones y la función de reunión. Esto último, le permite a la función de reunión discernir entre problemáticas inherentes a la comunicación de aquellas propias de los datos, con la finalidad de garantizar la sincronización entre las partes. Así el proceso de corrección y análisis se abstrae de las problemáticas de comunicación y se aboca a las problemáticas propia de los datos.

Dado que el proceso de corrección y análisis dispone de metadatos basados en una ontología de medición claramente definida, tiene la ventaja de ser capaz de identificar el significado de los datos y su contexto asociado, lo que permite un análisis consistente de los mismos. Así, el proceso de toma de decisión puede adecuarse a los contextos de medición e incrementar la precisión de las correspondientes decisiones.

Como trabajo a futuro, se abordará e implementarán los procesos vinculados a recolección y adaptación dentro del modelo integrado de procesamiento de flujos de datos, como así también lo referido a CINCAMI/MIS junto con el protocolo CINCAMI/MIP. En este último sentido, se pretende que estructuralmente los procesos de recolección y adaptación estén preparados para ser especializados a fuentes de datos de diferente naturaleza.

Referencias

1. Babcock B., Babu S., Datar M., Motwani R., and Widom J. (2002) Models and Issues in Data Stream Systems. In proc. of 21st ACM Symposium of Principles of Database Systems (PODS 2002). Madison, USA.
2. Fan W., Huang Y., Wang H. and Yu P. (2004). Active Mining of Data Streams. In proc. of Int'l Conference on Data Mining (SIAM2004). Florida, USA.

3. Olsina L, Papa F., Molina H. (2007) How to Measure and Evaluate Web Applications in a Consistent Way. Chapter 13 in Springer Book, Human-Computer Interaction Series, titled *Web Engineering: Modelling and Implementing Web Applications*; Rossi, Pastor, Schwabe, & Olsina (Eds.), pp. 385–420.
4. Molina H, & Olsina L. (2007). Towards the Support of Contextual Information to a Measurement and Evaluation Framework. In proc. of 6th Int'l Conference on the Quality Information and Communications Technology (QUATIC07). IEEE CS Press, Lisbon, Portugal. pp. 154–163.
5. Golab L & Oszu T. (2003). Issues in Data Stream Management. ACM SIGMOD Record, Vol. 34: 2, pp.5-14, ISSN 0163-5808.
6. Abidogun O. (2005). Data Mining, Fraud Detection and Mobile Telecommunications: Call Pattern Analysis with Unsupervised Neural Networks. Master thesis. University of the Western Cape.
7. The Y., Zaitun A., and Lee S. (2001). Data Mining Using Classification Techniques in Query Processing Strategies. ACS/IEEE Int'l Conference on Computer Systems and Applications (AICCSA'01).
8. Tao Y., and Papadias D. (2006). Maintaining Sliding Window Skylines on Data Streams, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18: 3, pp. 377-391.
9. Chaudhry N., Shaw K., and Abdelguerfi M. (2005). Stream Data Management. Springer. pp. 1-11.
10. Bose S. and Fegaras L. (2004). Data Stream Management for Historical XML Data . In proc. of Int'l Conference on Management of Data (ACM SIGMOD2004). Paris, France.
11. Abajo Martínez Nicolás (2004). "ANN quality diagnostic models for packaging manufacturing: an industrial data mining case study". In proc. of tenth ACM SIGKDD 2004
12. Foster, Ian (1995). "Designing and Building Parallel Programs: Concepts and Tools for Parallel Software". Addisson-Wesley.
13 . Pérez López C. (2005). Métodos Estadísticos Avanzados con SPSS. Thomson.
14. Johnson D. (2000). Métodos Multivariados Aplicados al Análisis de Datos. Thomson.
15. Aggarwal C., Han J., Wang J. and Yu, P. S. (2004). On Demand Classification of Data Streams, Proc. of Int'l Conf. on Knowledge Discovery and Data Mining (KDD'04), Seattle, WA.
16. Ben-David S., Gehrke J., and Kifer D. (2004). Detecting Change in Data Streams. Proc. of VLDB04.
17. Dong G., Han J., Lakshmanan L.V.S., Pei J., Wang H. and Yu P.S. (2003). Online mining of changes from data streams: Research problems and preliminary results. In Proc. of the Workshop on Management and Processing of Data Streams. In cooperation with the Int'l Conference on Management of Data (ACM-SIGMOD'03), San Diego, CA.
18. Gama J., Medas P., and Rodríguez P. (2005). Learning Decision Trees from Dynamic Data Streams, ACM Symposium on Applied Computing - SAC05.
19. Keith Visco and Assaf Arkin (2008). Castor. http://www.castor.org.
20. U. Srivastava and J. Widom (2004). Flexible Time Management in Data Stream Systems. In proc. of ACM PODS (Principles of Database Systems), 2004.
21. B. Babcock, M. Datar, and R. Motwani (2004). Load Shedding for Aggregation Queries over Data Streams. In Proc. of IEEE ICDE (International Conference on Data Engineering), 2004.
22. B. Babcock, S. Babu, M. Datar, R. Motwani, and D. Thomas (2004). Operator Scheduling in Data Stream Systems. VLDB Journal, Vol. 13:4, pp. 333-353.
23. N. Tatbul, S. Zdonik (2006). Window-aware Load Shedding for Aggregation Queries over Data Streams. In proc. of the 32nd ACM VLDB.
24. E. Ryvkina, A. S. Maskey, M. Cherniack, S. Zdonik (2006). Revision Processing in a Stream Processing Engine: A High-Level Design. In proc. of the 22nd IEEE ICDE.
25. J.-H. Hwang, M. Balazinska, A. Rasin, U. Çetintemel, M. Stonebraker, S. Zdonik (2005). High Availability Algorithms for Distributed Stream Processing. In proc. of the 21st IEEE ICDE.
26. C. Aggarwal, J. Han, J. Wang, and P. S. Yu (2004). On Demand Classification of Data Streams. In Proc. ACM KDD.
27. C. Aggarwal, J. Han, J. Wang, P. S. Yu (2003). A Framework for Clustering Evolving Data Streams. In Proc. ACM VLDB.
28. Yun Chi, Philip S. Yu, Haixun Wang, Richard R. Muntz(2005). Loadstar: A Load Shedding Scheme for Classifying Data Streams. In proc. of SIAM International Conference on Data Mining (SIAM SDM).
29. Singh, S., Vjirkar, P. and Lee, Y. (2003). Context-aware data mining framework for wireless medical application. Lectures Notes in Computer Science, Vol 2736 pp. 381-391.