Procesamiento de Flujo de Datos Enriquecidos con Metadatos de Mediciones: Un Análisis Estadístico

Mario Diván[1,2], Luis Olsina[2], Silvia Gordillo[3]

[1] Facultad de Ciencias Económicas y Jurídicas, UNLPam, Santa Rosa, La Pampa, Argentina
[2] GIDIS_Web, Facultad de Ingeniería, UNLPam, General Pico, La Pampa, Argentina.
[3] LIFIA, Facultad de Informática, UNLP, La Plata, Buenos Aires, Argentina.
{mjdivan,olsinal}@ing.unlpam.edu.ar
gordillo@lifia.info.unlp.edu.ar

Resumen. Este trabajo discute el proceso de análisis estadístico que se efectúa sobre las mediciones generadas en fuentes de datos heterogéneas y enviadas a través de flujo de datos (data streams), las cuales arriban para su procesamiento junto con los metadatos asociados a la definición formal de un proyecto de medición y evaluación. Esto permite guiar el análisis en forma consistente en busca de problemáticas típicas asociadas a los datos. Se prueba el prototipo generado para el proceso de análisis estadístico en un ambiente controlado, a los efectos de contrastar empíricamente los tiempos insumidos por el mismo y detectar las principales causas de variabilidad del sistema.

Palabras Clave: Medición, Flujo de Datos, C-INCAMI, Análisis Estadístico.

1. Introducción

En la actualidad existen aplicaciones que procesan datos a medida que se generan en forma continua, con el fin de responder a consultas y/o adecuar su comportamiento en función del propio arribo de los datos [1]. En tales aplicaciones, el arribo de un nuevo dato representa la llegada de un valor asociado a un comportamiento sintáctico, careciendo a menudo de sustento semántico y formal.

Desde el punto de vista de sustento semántico y formal para la medición y evaluación (M&E), el marco conceptual C-INCAMI (*Context-Information Need*, *Concept model*, *Attribute*, *Metric* and *Indicator*) establece una ontología que incluye los conceptos y relaciones necesarios para especificar los datos y metadatos de cualquier proyecto de M&E [2]. Por otra parte, en el *Enfoque Integrado de Procesamiento de Flujos de Datos* (EIPFD) [3] se ha planteado la necesidad de integrar los flujos de datos heterogéneos con metadatos basados en el marco C-INCAMI, permitiendo de este modo un análisis consistente de las mediciones, considerando su contexto de procedencia y su significado dentro de un proyecto de M&E. A partir de este enfoque de procesamiento de flujos de datos y considerando el tipo de aplicaciones mencionadas, el presente artículo realiza un análisis estadístico "al vuelo" sobre el flujo de datos de un modo consistente, sustentado en información

contextual y metadatos embebidos dentro de los propios flujos, lo que permitiría incrementar la robustez del análisis. Como contribuciones específicas se plantea, *(i) relacionado con la métrica:* la detección de desviaciones con respecto a la definición formal, identificación de outliers y de ausencia de valor; *(ii) relacionado con el grupo de mediciones:* la detección instantánea de correlaciones, identificación de los factores de variabilidad del sistema y la detección de tendencia sobre el propio flujo de datos, considerando la situación contextual de la fuente generadora de las mediciones.

El presente artículo se organiza en cinco secciones. La sección 2 resume los objetivos y motivaciones. La sección 3 plantea el modo en que los metadatos inciden en el análisis estadístico de los diferentes flujos de datos y expone la forma en que se lleva adelante dicho análisis. La sección 4 realiza la simulación del prototipo y el análisis de sus resultados. Por último, se presentan las consideraciones finales.

2. Procesamiento de Flujos de Mediciones. Objetivo y Motivación

En el EIPFD [3] se plantearon las componentes de una arquitectura especializada en la gestión de flujos de mediciones. En este sentido, la idea central del EIPFD es: automatizar los procesos de recolección permitiendo la incorporación de fuentes heterogéneas, analizar y detectar anomalías sobre los datos ante el propio arribo, y tomar decisiones on-line en base a la definición formal de un proyecto de M&E. En cuanto al análisis y detección de anomalías, este artículo propone un enfoque on-line con técnicas estadísticas como análisis descriptivo, correlación y componentes principales, las que se abordarán, conceptualmente, desde la óptica del proceso y empíricamente desde la simulación de carga de trabajo (secciones 3 y 4).

3. Aspectos del Diseño y del Proceso del Análisis Estadístico

Las mediciones vienen asociadas con metadatos que contienen su definición formal, como así también las propiedades contextuales relacionadas al ámbito del valor medido. Estos metadatos permiten organizar las mediciones en dos niveles. El primer nivel agrupa las mediciones por grupo de seguimiento mientras que dentro de cada grupo de seguimiento, en el segundo nivel, las mediciones se estructuran por cada métrica, la cual mide un determinado atributo para una entidad dada. Así, la contribución de los metadatos a la organización, comparación y análisis de las mediciones es sustancial, por lo que mantiene agrupadas las unidades lógicas de medición a nivel de métrica sin perder relación con el grupo de seguimiento al que pertenece. A su vez, mantiene la relación de cada medición con su situación contextual posibilitando la comparación entre grupos de seguimiento. De este modo, los metadatos permiten guiar las técnicas de *load shedding* [4], dado que ante una situación de potencial desborde, es factible priorizar automáticamente las métricas críticas y aplicar el descarte selectivamente.

En el EIPFD, el *analizador estadístico*, es una pieza de software responsable de

llevar adelante el proceso de análisis estadístico (*Analysis & Smoothing Function – ASF*). *ASF* recorre cada uno de los grupos de seguimiento y aplica on-line dos técnicas multivariadas: *Análisis de Componentes Principales* y *Análisis de Correlación* [5]. El análisis de componentes principales se emplea para reducir la dimensionalidad de los problemas y el análisis de correlación, verificará si eventualmente existen correlaciones del tipo lineal entre las métricas que conforman el grupo de seguimiento. El grupo de seguimiento puede ser entendido, por ejemplo (ver escenario de uso en [3]) como un paciente trasplantado ambulatorio en particular, en donde cada métrica hace referencia a un atributo que se desea monitorear (por ej. la frecuencia cardíaca (FC), la que es medida e informada continuamente). Así, el análisis de componentes principales buscará identificar qué métricas (variables) incorporan mayor variabilidad al paciente (grupo de seguimiento, sistema) y el análisis de correlación intentará identificar potenciales relaciones entre las métricas monitoreadas sobre el paciente, a los efectos de detectar situaciones de arrastre, por ej., la temperatura ambiental (propiedad contextual) con respecto a la FC (métrica).

Una vez que ASF culmina el análisis de un grupo de seguimiento y previo a avanzar sobre otro grupo, analiza descriptivamente cada métrica que lo compone a los efectos de detectar anomalías, ruido, outliers y/o tendencias basándose en la definición formal de la métrica o propiedad contextual, generando en paralelo sinopsis [6]. De detectarse alguna situación, ésta se informa al *tomador de decisiones* para que en base a lo definido en el proyecto de M&E actúe proactivamente.

4. Análisis de Resultados de la Simulación

El prototipo asociado al EIPFD implementa la funcionalidad que va desde la integración de las fuentes de datos heterogéneas y *grupos de seguimiento*, hasta el ASF, incluyendo la definición formal de los proyectos medición y el repositorio de metadatos C-INCAMI. Además implementa C-INCAMI/MIS para el intercambio de las mediciones y el buffer multinivel basado en metadatos. La implementación del prototipo se ha desarrollado en JAVA, empleando R [7] como motor de cálculo estadístico y el *CRAN* (*Comprehensive R Archive Network*) RServe para permitir el acceso TCP/IP desde la aplicación a R, primando la comunicación directa.

La simulación se desarrolló generando los datos de las mediciones en forma pseudo-aleatoria, considerando dos parámetros: cantidad de métricas (cada métrica en la simulación se corresponde con una variable) y cantidad de mediciones por variable. La simulación varió en forma discreta el parámetro de la cantidad de variables en el flujo de datos de 3 a 99 y el parámetro del volumen de mediciones por variable de 100 a 1000. El prototipo, R y Rserve se ejecutaron sobre una PC con procesador AMD Athlon x2 64bits con 3GB de RAM y SO Windows Vista Home Premium.

Del proceso de simulación, se han obtenido 1390 mediciones sobre los tiempos totales de procesamiento en base a la evolución de la cantidad de variables y mediciones, lo que permite estadísticamente confluir a resultados comprobables que permitieron validar el prototipo en un contexto controlado. La gráfica de la Figura 1b muestra claramente cómo la evolución de la cantidad de variables afecta notablemente el tiempo de procesamiento total del flujo de datos, incrementándolo con respecto a lo

expuesto por la gráfica (a); aquí se observa que el incremento en el tiempo de procesamiento que se produce debido al aumento de las mediciones es ínfimo en comparación al referido por las variables. Este último aspecto indica que los mecanismos de load shedding realmente consiguen su objetivo de evitar desbordes y no alterar el tiempo de procesamiento del flujo frente a la variación del volumen del mismo, mientras que la incorporación de variables influye dado que además del volumen del dato que se incorpora por la variable extra, se introduce la interacción con las variables preexistentes que es la causa y diferencia principal en términos de tiempo de procesamiento, con respecto al incremento producido por las mediciones.

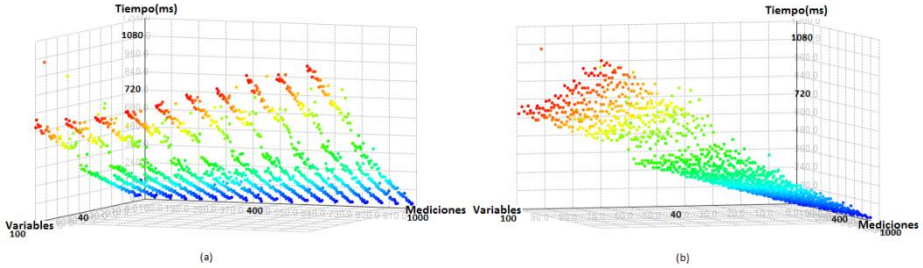


Figura 1 *Comparativa de la evolución del tiempo de procesamiento total (ms) frente a la evolución de la cantidad de variables y mediciones*

Ambas partes de la gráfica de superficie (a) y (b) representan cada punto con un color el cual se asocia a la cantidad de variables. A partir de la simulación, se definieron las siguientes variables que han sido objeto de medición considerando al flujo, como la entidad bajo análisis en cada una de las 1390 mediciones indicadas:

- **Startup**: *Tiempo en ms necesario para inicializar las funciones de análisis*
- **AnDesc**: *Tiempo en ms necesario para efectuar el análisis descriptivo sobre el flujo completo*
- **Cor**: *Tiempo en ms necesario para efectuar el análisis de correlación por grupo de seguimiento dentro del flujo completo*
- **Pca**: *Tiempo en ms necesario para efectuar el análisis de componentes principales por grupo de seguimiento dentro del flujo completo*
- **Total**: *Tiempo total en ms necesario para efectuar todos los análisis sobre el flujo completo*

Los parámetros de la simulación, a los efectos del análisis estadístico de los resultados, se representan como *qVar* para indicar la cantidad de variables del flujo y *meds* para indicar la cantidad de mediciones por variable en el flujo. En adelante y a los efectos de simplificar la lectura del análisis estadístico, los parámetros qvar y meds se referenciarán directamente como variables, al igual que se lo hace con *startup*, *andesc*, *cor*, *pca* y *total*.

La matriz de correlación de Pearson expuesta en la Figura 2(a), en primer lugar confirmaría una relación lineal entre la cantidad de variables (*qvar*) y el tiempo total de procesamiento del flujo (*total*) dado la presencia de un coeficiente de 0.95. En segundo lugar, se puede concluir que el tiempo total de procesamiento guardaría una fuerte relación lineal con respecto al tiempo del análisis descriptivo con un coeficiente 0.99, siguiéndole en ese orden el tiempo de *pca* con 0.9 y *cor* con 0.89.

Las matrices resultantes del análisis de componentes principales expuestas en la Figura 2 b) y c), revelan cuáles de las variables aportan mayor variabilidad al sistema. En este sentido y a modo de ejemplo, el primer autovalor (fila 1, Figura 2 b)) explica el 66% de la variabilidad del sistema y si se observa su composición en la matriz de autovectores (col. e1, Figura 2 c)), las variables que más contribuyen en términos absolutos son *AnDesc, Cor, Pca* y *qVar*.

Coeficientes de correlación

Correlación de Pearson: Coeficientes\probabilidades

	qVar	meds	Startup	AnDesc	Cor	Pca	Total
qVar	1.00	1.00	0.15	0.00	0.00	0.00	0.00
meds	8.9E-05	1.00	0.11	5.8E-06	0.00	0.00	0.00
Startup	-0.04	-0.04	1.00	0.28	0.30	0.68	0.41
AnDesc	0.97	0.12	-0.03	1.00	0.00	0.00	0.00
Cor	0.79	0.40	-0.03	0.85	1.00	0.00	0.00
Pca	0.77	0.45	-0.01	0.85	0.89	1.00	0.00
Total	0.95	0.21	-0.02	0.99	0.89	0.90	1.00

(a)

Autovalores

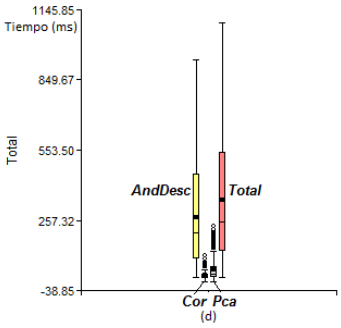
Lambda	Valor	Proporción	Prop Acum
1	4.62	0.66	0.66
2	1.11	0.16	0.82
3	0.99	0.14	0.96
4	0.14	0.02	0.98
5	0.10	0.01	1.00
6	0.02	3.3E-03	1.00
7	3.3E-05	4.7E-06	1.00

(b)

Autovectores

Variables	e1	e2	e3
qVar	0.43	-0.31	-0.08
meds	0.14	0.87	0.18
Startup	-0.02	-0.22	0.98
AnDesc	0.45	-0.19	-0.04
Cor	0.44	0.14	0.04
Pca	0.44	0.18	0.07
Total	0.46	-0.11	-0.01

(c)



(d)

Figura 2. (a) *Matriz de Correlación de Pearson,* **(b)** *Matriz de Autovalores y* **(c)** *Matriz de Autovectores asociadas al Análisis de Componentes Principales (PCA) y* **(d)** *Boxplot de las variables AnDesc, Cor, PCA y Total*

De este modo, si se deseara reemplazar las siete variables enunciadas por las nuevas tres variables (e1 a e3), se estaría explicando el 96% de la variabilidad del sistema, donde las principales variables en término de aporte están asociadas con AnDesc, Cor, Pca y qVar. El sistema sólo es afectado en un 16% por la evolución de las mediciones y en un 14% por el tiempo de inicialización, lo que implica un punto importante de resaltar dado que la única variable externa al prototipo que no puede ser controlada por el mismo, el volumen de arribo de las mediciones, sólo representa un 16% y en ningún caso representó una situación de desborde en la cola de servicios.

De las cuatro variables que más variabilidad aportan al sistema, tres de ellas componen parte del tiempo total de procesamiento por lo que mediante el box plot de la **Error! Reference source not found.**, puede corroborarse que la variable más influyente a la magnitud del tiempo total de procesamiento es *AnDesc*. Adicionalmente, debe destacarse que el mayor tiempo obtenido para procesar 99 variables con 1000 mediciones (99000 mediciones en total por flujo) fue de 1092 ms, es decir 1,092 segundos, sobre un hardware básico y totalmente accesible en el mercado, lo que permite establecer un umbral de aplicabilidad del prototipo.

5. Consideraciones Finales

El artículo ha discutido cómo la presencia de los metadatos basados en un marco formal de M&E e incorporados en forma conjunta con las mediciones, permiten una organización de los data streams que incorpora consistencia en el análisis estadístico, de modo que identifica las componentes formales de los datos y su contexto asociado.

Se ha probado a partir del análisis estadístico de los resultados de la simulación, que el prototipo que implementa el EIPFD es más susceptible al incremento de la cantidad de variables que al incremento de la cantidad de mediciones por variable en términos de tiempo de procesamiento. Mediante el Análisis de Componentes Principales se ha comprobado que *andesc* es quien define la mayor proporción del tiempo final de procesamiento del flujo de datos. Considerando un entorno de prueba implementado mediante un hardware accesible en el mercado, se pudo establecer como patrón de comparación a los efectos de poder evaluar consistentemente los ámbitos de aplicación, que para procesar 99000 mediciones (99 variables y 1000 mediciones/variable) el tiempo máximo arrojado ha sido 1092ms. Como corolario del análisis estadístico, ha podido comprobarse la efectividad de los mecanismos de load shedding dentro del buffer multinivel, dado que la evolución en la cantidad de mediciones no ha comprometido el funcionamiento del prototipo ni tampoco afectado en gran proporción al tiempo final de procesamiento del flujo de mediciones.

Existen trabajos que se enfocan en el procesamiento sintáctico del flujo tales como Aurora & Borealis [8] y STREAM [9]. Nuestro prototipo realiza la gestión de metadatos en forma conjunta con las mediciones basado en un marco formal de M&E, lo que permite guiar la organización de las mediciones, posibilitando análisis consistentes y comparables desde el punto de vista estadístico, e incorporando una actuación proactiva ante la detección de potenciales situaciones de anomalía o bien, a partir de la decisión del clasificador on-line en base al flujo.

Como trabajo a futuro y a los efectos de culminar con la funcionalidad del prototipo, se pretende implementar los clasificadores on-line del EIPFD y sus mecanismos de retroalimentación, ajustando el DM para poder incorporar dentro de su proceso las decisiones surgidas de los clasificadores.

Reconocimientos. Esta investigación está soportada por los proyectos PICT 2188 de la Agencia de Ciencia y Tecnología y 09/F052 por la UNLPam, Argentina.

Referencias

1. Gehrke J., Balakrishan H., Namit J. "Towards a Streaming SQL Standard" in *VLDB*, Auckland, New Zealand, 2008.
2. Olsina L., Papa F., Molina H. "How to Measure and Evaluate Web Applications in a Consistent Way," in *Ch. 13 in Web Engineering* Springer Book HCIS, 2008, pp. 385–420.
3. Diván, M., Olsina, L. "Enfoque Integrado para el Procesamiento de Flujos de Datos: Un Escenario de Uso," in *CIBSE*, pp. 374-387, 2009.
4. Rundensteiner W., Mani M., Wei M. "Utility-driven Load Shedding for XML Stream Processing," in *International World Wide Web*, Beijing, China, pp. 855-864, 2008.
5. Johnson, D. *Métodos Multivariados Aplicados al Análisis de datos*. México: Thomson Editores, 2000.
6. Jiang Q., Chakravarthy S. *Stream Data Processing: A Quality of Service Perspective*. Springer, 2009.
7. R Software Foundation, *R Software*. Vienna, Austria: The R Foundation for Statistical Computing, 2010.
8. Ahmad Y., Balazinska M., Cetintemel U., Cherniack M.. Hwang J., Lindner W., Maskey A., Rasin A., Ryvkina E., Tatbul N., Xing Y., Zdonik S., Abadi D."The Design of the Borealis Stream Processing Engine," in Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, pp. 277-289, 2005.
9. The Stream Group, "STREAM: The Stanford Stream Data Manager," 2003.