ISBN: 978-987-4998-33-0
DOI:https://doi.org/10.33414/AJEA.4.414.2019



# Estrategia de Recomendación por Similitud en Repositorios con Grandes Volúmenes de Datos de Medición y Evaluación

## A Recommendation Strategy based on Similarity in Big Measurement Repositories

Presentación: 17/10/2019

Doctorando:

**María Laura Sánchez Reynoso**

Data Science Research Group
Facultad de Ciencias Económicas y Jurídicas
Universidad Nacional de La Pampa
mlsanchezreynoso@eco.unlpam.edu.ar

Director/es:

**Mario José Diván**

Data Science Research Group
Facultad de Ciencias Económicas y Jurídicas
Universidad Nacional de La Pampa
mlsanchezreynoso@eco.unlpam.edu.ar

## Resumen

La Arquitectura de Procesamiento de Flujos de Datos es una estrategia de procesamiento basada en Apache Storm focalizada en proyectos de medición y evaluación. Los proyectos de medición y evaluación son definidos utilizando un marco forma de medición y evaluación que permite establecer previamente la entidad bajo monitoreo junto con sus conceptos asociados. Este acercamiento permite guiar el procesamiento y toma de decisiones basada en datos de múltiples proyectos concurrentes sustentado en la semántica de las etiquetas definidas en el alcance de cada proyecto. La estrategia incorpora un comportamiento activo, lo que implica que ante una situación tipificada es posible proveer recomendaciones y-o instruir cursos de acciones basado en la experiencia previa. Sin embargo, es posible que existan situaciones para las cuales una entidad bajo monitoreo no cuente con experiencia o conocimiento previo (por ejemplo, por que constituye una nueva situación), lo que imposibilitaría brindar sugerencias. Para abordar esta situación, la idea es detectar entidades bajo monitoreo similares estableciendo un puntaje de similitud que permita brindar experiencias y-o recomendaciones por analogía con otra entidad en caso de ausencia de estas en la entidad objeto de análisis. Aquí se presenta un avance parcial de esta línea de investigación. Así, se ha logrado establecer coeficientes de similitud estructurales y comportamentales, analizar el efecto de técnicas de descarte selectivo en el procesamiento de flujos, fomentar la interoperabilidad de proyectos de medición y evaluación y su efecto en la confiabilidad del sistema de medición y evaluación.

Palabras claves: Procesamiento de Datos en Tiempo Real, Mediciones, Entidades, Similitud, Grandes Volúmenes de Datos

## Abstract

The data stream processing architecture is an Apache Storm-based processing strategy focused on the Measurement and Evaluation (M&E) projects. Projects are defined using an M&E framework which allows previously establishing the entity under monitoring jointly with their associated concepts. This approach is able to guide the processing and decision making based on data coming from multiple concurrent projects supported by the semantics of each tag defined into each project scope. The strategy incorporates active behaviour, which implies it is possible to provide recommendations or instruct courses of actions given a situation based on previous experiences. However, it is possible there exist situations for which there are no specific knowledge or previous experiences (e.g. in front of new situations), being no possible to provide suggestions. Thus, the idea is to detect entities under monitoring by similarity, establishing a score that allows providing experiences and knowledge by analogy. In this way, when there is no experience available for a given entity, other experiences and knowledge could be retrieved from third entities by analogy, being able to provide approximated suggestions. Here, advances of the research line are introduced in terms of the establishment of structural and behavioural coefficients, the effect of load shedding techniques in the data stream processing, the perspectives of measurement and evaluation projects based in a project-common definition language jointly with its effect on the reliability of the measurement and evaluation system.

Keywords: Real-Time Data Processing, Measurement, Similarity, Big Data.

## 1. Introducción

La Arquitectura de Procesamiento basada en Metadatos de Mediciones (en inglés, Processing Architecture base don Measurement Metadata -PAbMM-) [1] es un motor de procesamiento de flujos de mediciones basada en topologías de Apache Storm[1]. La arquitectura se encuentra especializada en el monitoreo de entidades bajo análisis en proyectos de medición y evaluación.

Los proyectos de medición y evaluación monitoreados en forma automática por PAbMM son definidos en términos del marco C-INCAMI (Context – Information Need, Concept Model, Attribute, Metric and Indicator)[2], [3], siguiendo la estrategia GOCAME (Goal-Oriented Context-Aware Measurement and Evaluation)[4]. Si bien hay diferentes tipos de marcos de medición y evaluación [3, 2], C-INCAMI cuenta con el valor añadido de tener una ontología subyacente, la cual permite definir los términos, conceptos y relaciones entre ellos a los efectos de definir y automatizar un proyecto de medición y evaluación.

Una propuesta de extensión del marco C-INCAMI fue presentada en [5], en donde entre otros aspectos se incorporó la posibilidad de modelar medidas no deterministas, datos complementarios, aspectos de trazabilidad relacionados a las fuentes de datos, restricciones, entre otros aspectos.

C-INCAMI/MIS es un esquema de intercambio de mediciones basado en el marco C-INCAMI, el cual en su segunda versión incorpora las extensiones del marco C-INCAMI [5], permitiendo integrar medidas provenientes desde fuentes de datos heterogéneas, a la vez que mantiene la consistencia y trazabilidad respecto del origen. Dentro del flujo de medidas pueden informarse tanto la medida como sus metadatos descriptivos asociados. Incluso, existe una librería denominada cincamimis [6], que permite generar flujos C-INCAMI/MIS mediante los formatos de datos XML (describir acrónimo) o JSON (describir acrónimo), haciendo empleo de compresión vía GZIP. Más aún, la librería CINCAMI/PD fue liberada en 2018 y permite el intercambio de definiciones de proyectos de M&E bajo formatos XML o JSON.

PAbMM guía su estrategia de procesamiento a partir de los metadatos. Por ejemplo, si un valor está asociado con la métrica del atributo temperatura ambiental, mediante la definición del proyecto se sabe la escala, unidad y valores típicos asociados. De este modo, un valor de 55ºC podría implicar a) una fuente de calor cercana -incendio, etc.-, b) un posible error de calibración, o c) un error en el sensor. En todos los casos, PAbMM disparará una alarma al tomador de decisiones para que se informe junto con las recomendaciones del caso de acuerdo con la entidad bajo monitoreo.

[1] https://storm.apache.org

Ahora bien, en el momento que una o más recomendaciones deben ser adjuntadas a la decisión a la cual eventualmente arribe el tomador de decisiones de PAbMM, se supone que las mismas son pertinentes a la decisión, o al menos, tienen cierto grado de incidencia en la misma. Cuando la entidad bajo monitoreo cuenta en la memoria organizacional con experiencia específica previa y-o conocimiento, la labor se simplifica por cuanto tal experiencia o conocimiento es específica a esa entidad. Ahora bien, cuando tal entidad no tiene experiencia alguna respecto a una situación dada, o bien, no posee el suficiente conocimiento previo, el objetivo es desarrollar una estrategia que recupere las recomendaciones desde entidades que sean similares en término semántico.

No obstante, PAbMM es una topología de Apache Storm por lo que procesa flujos de medidas en tiempo real, mientras que la memoria organizacional es una base de datos columnar, lo que implican modelos de procesamiento de datos totalmente diferentes. Por ello, la estrategia planteada de búsqueda de recomendaciones en entidades similares debe ser factible de implementación en memoria para responder al tomador de decisiones de PAbMM en tiempo real, dado que los tiempos de búsqueda en una base de datos columnar no serían factibles.

Actualmente, PAbMM incorpora una estrategia de recomendación de entidades similares desde el punto de vista estructural y comportamental. El punto de vista estructural pre-calcula la similitud de entidades a partir de los atributos comunes que describen la entidad bajo análisis, lo cual permite arribar a una puntuación estructural. Mientras que el punto de vista comportamental analiza el estado actual y evolución de la serie de datos asociada con cada atributo descriptivo de la entidad, a los efectos de determinar si aun siendo similares estructuralmente, el comportamiento de las medidas para cada atributo es similar o no mediante análisis de correlación. Claro que ambos coeficientes no consideran el abordaje semántico en la definición de los atributos característicos de la entidad y-o su comportamiento, aspecto que pretende ser abordado mediante el presente plan.

De este modo, el objetivo principal del presente trabajo consiste en desarrollar una estrategia de recomendación en memoria, a partir de repositorios de medición y evaluación basados en el marco formal C-INCAMI, a los efectos de localizar entidades bajo monitoreo semánticamente similares, y poder así, reutilizar su conocimiento y experiencia en el proceso de toma de decisiones en tiempo real cuando sea requerido.

El trabajo se organiza en cinco secciones. Sección 2 define el contexto de la investigación junto con las limitaciones definidas sobre el mismo. Sección 3 describe sintéticamente la metodología de trabajo. Sección 4 describe resultados parciales publicados junto con aquellos que han sido aceptados y presentados estando a la fecha en prensa. Sección 5 discute algunas conclusiones y trabajos futuros.

## 2. Contexto y límites

El objetivo y problemática introducida se limita a proyectos de medición y evaluación. Se establece el marco formal de medición y evaluación C-INCAMI para el análisis de similitud en términos de la entidad bajo monitoreo o análisis. Este aspecto no es trivial, dado que el concepto de similitud refiere directamente a la entidad bajo análisis. La estrategia empleada para definir los proyectos de medición y evaluación es GOCAME, la cual establece los pasos necesarios para definir un proyecto. A partir de allí, las entidades bajo análisis se caracterizan mediante atributos, como así también las propiedades del contexto asociado. Tales atributos y propiedades de contexto son descritos en forma narrativa en español.

## 3. Metodología

A partir de la estrategia de similitud que se defina en términos del marco de medición y evaluación en consonancia con GOCAME, se trabajará en forma iterativa con un ciclo de vida en espiral para construir diferentes componentes adicionales a PAbMM.

El esquema en espiral permitirá efectuar el esquema de pruebas en forma unitaria y de integración, a la vez que permite ciclos cortos con productos compatibles con la metodología Scrum. Las iteraciones del ciclo

contrastarán los productos respecto a lo esperado en términos del proceso y su lógica, contrastando su funcionamiento.

Al finalizar el ciclo de desarrollo en espiral, una librería Java implementando la lógica de los procesos de puntuación semántica/comportamental/estructural, extensiones y-o diferentes artefactos serán plausibles de generación de publicaciones parciales.

## 4. Resultados

De acuerdo con lo indicado precedentemente en términos de la metodología, a la luz del contexto, objetivo y límites establecidos, se cuenta con los siguientes resultados parciales debidamente publicados:

- "*An Architecture for the Real-Time Data Stream Monitoring in IoT*" [7]: brinda una perspectiva integral del procesamiento de flujos de datos de medición guiados por metadatos, capaz de brindar recomendaciones a situaciontes tipificadas. https://doi.org/10.1007/978-981-13-8759-3_3

- "*The Real-Time Measurement and Evaluation as System Reliability Driver*"[8]: Analiza la confiabilidad del sistema de medición y evaluación basado en el procesamiento de flujos de datos. https://doi.org/10.1201/9781351117661-11

- "*A library for articulating the measurement streams with columnar data*" [9]: Define una librería capaz de articular flujos de mediciones con bases de datos columnares para almacenamiento monolítico de datos. https://www.sciencepubco.com/index.php/ijet/article/download/23373/11680

- "*Fostering the Interoperability of the Measurement and Evaluation Project Definitions in PAbMM*"[10]: Establece el esquema de intercambio de definiciones proyectos de medición basados en C-INCAMI con el objetivo de fortalecer la interoperabilidad de sistemas de medición heterogéneos. https://doi.org/10.1109/ICRITO.2018.8748766

- "*A Load-Shedding Technique based on the Measurement Project Definition*"[11]: Define una estrategia en la cual el descarte selectivo de medidas puede ser llevado adelante mediante el empleo de los metadatos asociados con la definición del proyecto de medición y evaluación. https://doi.org/10.1007/978-981-13-9406-5_122

- "*Experiences in the Business Process Modelling at Public Organizations of La Pampa*"[12]: Presenta contextos en los cuales la entidad bajo monitoreo en PAbMM se asocia con procesos de negocios junto con sus implicancias. http://doi.org/10.1109/ICRITO.2017.8342446

- "*Behavioural Similarity Analysis for Supporting the Recommendation in PAbMM*"[13]: Define en PAbMM la alternativa de implementar coeficientes estructurales basados en el marco C-INCAMI, como así también comportamentales basado en el comportamiento de la serie de datos asociada. https://doi.org/10.1109/ICTUS.2017.8285992

- "*Improving the Real-Time Searching in the Organizational Memory*"[14]: Define una alternativa en la estrategia de búsqueda por similitud dentro de la memoria organizacional que contiene experiencias previas y conocimiento de expertos. https://doi.org/10.1016/j.procs.2019.06.043

- "*Extending the Data Stream Processing Strategy to Scenario Analysis*"[15]: Se incorpora la posibilidad de análisis multi-escenarios en el contexto del análisis de medidas vinculados con entidades bajo monitoreo. https://doi.org/10.30534/ijatcse/2019/0181.42019

Los siguientes artículos han sido aceptados y presentados, estando actualmente en prensa:

- Sánchez Reynoso, M & Diván, M (2019) "A Systematic Literature Mapping on the Similar Semantically Entities in Measurement Projects". In Edutainment, The 13th International Conference on E-Learning and Games. Cali, Colombia, August 15-17 of 2019. [In press].

- Diván, M & Sánchez Reynoso, M (2019) "Articulating Heterogeneous Data Streams with the Attribute-Relation File Format". In International Conference on Electrical and Electronic Engineering 2019 (ICON3E2019). Putrajaya, Malaysia, June 24-25 of 2019. [In press]

- Diván, M and Sánchez Reynoso (2019) "Incorporating Scenarios and States Definitions on Real-Time Entity Monitoring in PAbMM". In proc. XLV Latin-American Computer Conference, Panam√°, September 30-October 4 of 2019. [In press].

## 5. Conclusiones

Como aspecto que aporta complejidad al momento de analizar el estado del arte, puede mencionarse la falta de consenso en relación con los marcos de medición y evaluación como así también de los conceptos y relaciones subyacentes. En tal sentido, puede indicarse que existen diferentes alternativas para llevar adelante el proceso de medición, provocando que el concepto de similitud "semántica" pueda variar sensiblemente de acuerdo con el marco escogido.

En términos de la línea de investigación, se han generado resultados parciales considerando entidades bajo monitoreo tangibles e intangibles, extensiones a la arquitectura de procesamiento de flujos de datos basado en escenarios y estados múltiples de la entidad, se han desarrollado extensiones a la estrategia de búsqueda por similitud sobre la memoria organizacional, incorporado técnicas de descarte selectivo guiado por los metadatos de medición, integrado la perspectiva global de la arquitectura de procesamiento, se analizaron los aspectos de confiabilidad de los sistemas de medición y evaluación, se integraron los flujos de datos con modelos de datos columnares, se avanzó con librerías que fomentaba la interoperabilidad en la definición del proyecto de medición y evaluación (aspecto esencial para determinar la similitud entre entidades), y finalmente se llevó adelante un mapeo sistemático de la literatura para analizar las estrategias actuales de abordaje de la similitud semántica entre entidades.

Como trabajo a futuro, se continuará con el análisis de similitud semántica sobre contextos de procesamiento de datos en tiempo real, en donde las sugerencias basada en situaciones tipificadas debieran ser provistas al instante.

## Referencias


[1] **M. J. Divan, "Processing architecture based on measurement metadata," in 2016 5th International** Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2016, pp. 6–15. https://doi.org/10.1109/ICRITO.2016.7784912

[2] **L. Olsina, F. Papa, and H. Molina, "How to Measure and Evaluate Web Applications in a Consistent Way,"** in Web Engineering: Modelling and Implementing Web Applications, G. Rossi, O. Pastor, D. Schwabe, and L. Olsina, Eds. London: Springer London, 2008, pp. 385–420. https://doi.org/10.1007/978-1-84628-923-1_13

[3] **H. Molina and L. Olsina, "Towards the Support of Contextual Information to a Measurement and Evaluation Framework," in 6th International Conference on the Quality of Information** and Communications Technology (QUATIC 2007), 2007, pp. 154–166. https://doi.org/10.1109/QUATIC.2007.21

[4] **P. Becker, F. Papa, and L. Olsina, "Process conceptual base for enriching a** measurement and evaluation **ontology," in CIBSE 2014: Proceedings of the 17th Ibero-**American Conference Software Engineering, 2014.

[5] **M. Diván and M. de los Ángeles Martín, "Towards a Consistent Measurement Stream Processing from** Heterogeneous Data Sources," **Int. J. Electr. Comput. Eng., vol. 7, no. 6, pp. 3164–**3175, Dec. 2017. https://doi.org/10.11591/ijece.v7i6.pp3164-3175

[6] M. J. Diván and M. de los Ángeles Martín, "A new storm topology for synopsis management in the processing architecture," in 2017 XLIII Latin American Computer Conference (CLEI), 2017, pp. 1–10. https://doi.org/10.1109/CLEI.2017.8226398

[7] M. J. Diván and M. L. Sánchez Reynoso, "An Architecture for the Real-Time Data Stream Monitoring in IoT," in Multimedia Big Data Computing for IoT Applications, S. Tanwar, S. Tyagi, and N. Kumar, Eds. Springer Nature Singapore, 2020, pp. 59–100. https://doi.org/10.1007/978-981-13-8759-3_3

[8] M. J. Diván and M. L. Sánchez Reynoso, "Real-Time Measurement and Evaluation as System Reliability Driver," in System Reliability Management, A. Anand and M. Ram, Eds. Boca Raton : Taylor & Francis, a CRC title, part of the Taylor &: CRC Press, 2018, pp. 161–188. https://doi.org/10.1201/9781351117661-11

[9] M. Diván, M; Sánchez Reynoso, "A Library for Articulating the Measurement Streams with Columnar Data," Int. J. Eng. Technol., vol. 7, no. 4, pp. 234–241, 2018. https://www.sciencepubco.com/index.php/ijet/article/download/23373/11680

[10] M. Diván, M.; Sánchez Reynoso, "Fostering the Interoperability of the Measurement and Evaluation Project Definitions in PAbMM," in 2018 7th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, ICRITO 2018, 2018, pp. 228–234. https://doi.org/10.1109/ICRITO.2018.8748766

[11] M. J. Diván and M. L. Sánchez Reynoso, "A Load-Shedding Technique Based on the Measurement Project Definition," 2020, pp. 1027–1033. https://doi.org/10.1007/978-981-13-9406-5_122

[12] M. Divan, M. L. S. Reynoso, F. Veralli, and M. Ostoich, "Experiences in the business process modelling at public organizations of La Pampa," in 2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2017, vol. 2018-Janua, pp. 326–332. http://doi.org/10.1109/ICRITO.2017.8342446

[13] M. J. Divan and M. L. S. Reynoso, "Behavioural similarity analysis for supporting the recommendation in PAbMM," in 2017 International Conference on Infocom Technologies and Unmanned Systems: Trends and Future Directions, ICTUS 2017, 2018, vol. 2018-Janua, pp. 133–139. https://doi.org/10.1109/ICTUS.2017.8285992

[14] M. L. Sánchez Reynoso and M. Diván, "Improving the Real-Time Searching in the Organizational Memory," Procedia Comput. Sci., vol. 154, pp. 293–304, 2019. https://doi.org/10.1016/j.procs.2019.06.043

[15] M. J. Diván, "Extending the Data Stream Processing Strategy to Scenario Analysis," Int. J. Adv. Trends Comput. Sci. Eng., 2019. Vol. 8 no. 1.4, pp.1-8. https://doi.org/10.30534/ijatcse/2019/0181.42019