AJEA - Actas de Jornadas y Eventos Académicos de UTN
VI Jornadas de Intercambio y Difusión de los Resultados de Investigaciones de los Doctorandos en Ingeniería
DOI: https://doi.org/10.33414/ajea.1068.2022

Hacia una Estrategia de Recomendación por Similitud Semántica en Repositorios con Grandes Volúmenes de Datos de Medición y Evaluación

Towards a Semantic Similarity Recommendation Strategy for Big Data Repositories

Presentación: 4 y 5 de Octubre de 2022

Doctorando:
María Laura Sánchez Reynoso
Data Science Research Group. La Pampa - Argentina
mlsanchezreynoso@gmail.com

Director:
Mario José Diván

Resumen

Un proceso de medición debiera ser repetible, extensible a nuevos requerimientos y sus medidas debieran poder ser comparables. Esto es así, para poder posibilitar que el mismo proceso de medición pueda ser implementado por diferentes actores, de manera que sus resultados sean comparables a lo largo del tiempo. Definido el proceso de medición, las medidas necesitan ser interpretadas a fin de poder tomar decisiones y a partir de ello recomendar cursos de acción. Puede suceder que dependiendo de la entidad que se encuentre bajo monitoreo, y una situación en particular, no siempre se dispone con antecedentes o conocimientos previos. Ello provocaría que, a pesar de haber identificado la situación, no sería posible brindar recomendaciones o cursos de acción configurado al escenario específico. Adicionalmente el Internet de las Cosas (en inglés, Internet of Things - IoT) ha permitido llevar a cabo el desarrollo de estrategias de recolección de datos, soportando procesos de toma de decisiones en tiempo real. El objetivo consiste en desarrollar una estrategia de recomendación a partir de mediciones, basado en el análisis de similitud semántica que permita brindar experiencias y-o recomendaciones por analogía con otra entidad.

Palabras clave: Similitud Semántica, Recomendación, Medición.

Abstract

A measurement process should be repeatable, extensible to new requirements, and comparable measurements. This is so to enable the same measurement process to be implemented by different actors so that their results are comparable over time. Once the measurement process has been defined, the measurements need to be interpreted

in order to be able to make decisions and, based on this, recommend courses of action. It may happen that depending on the entity that is under monitoring, and a particular situation, there is not always background information or prior knowledge available. This would mean that, despite having identified the situation, it would not be possible to provide recommendations or courses of action configured for the specific scenario. Additionally, the Internet of Things (in English, Internet of Things - IoT) has allowed the development of data collection strategies, supporting decision-making processes in real-time. The objective is to develop a recommended strategy based on measurements, based on semantic similarity analysis that allows providing experiences and/or recommendations by analogy with another entity.

Keywords: Semantic Similarity, Recommendation, Measurement.

Introducción

La medición y evaluación (M&E) es un proceso a través del cual es posible conocer el comportamiento asociado a una entidad bajo monitoreo, detectando en forma temprana desvíos ó anomalías respecto de su conducta esperada (G. Rossi and D. Schwabe, 2008). El empleo de marcos formales de M&E, permite acomodar los términos, conceptos y las relaciones entre ellos de manera de promover la consistencia, la repetitividad y la extensibilidad del proceso de M&E (M. J. Divan and M. de los A. Martin, 2017), como así también su automatización (Sánchez-Reynoso, 2014).

Se ha definido PAbMM (Processing Architecture based on Measurement Metadata), una estrategia de procesamiento en línea implementada sobre Apache Storm especializada en implementar proyectos de medición y evaluación. La idea en PAbMM (Diván, M., & Sánchez Reynoso, M. , 2017) se centra en reducir el espacio de búsqueda en memoria de las experiencias y/o conocimiento previo de aquellas entidades vinculadas. De este modo, cuando no exista conocimiento previo o experiencia previa, PAbMM podrá recomendar determinados cursos de acción de acuerdo a la estructura y comportamiento de la entidades bajo monitoreo que sean similares. Estas dos últimas variantes, suponen que para localizar entidades similares (junto con su experiencia ó conocimiento asociado), tanto la definición de su estructura como la de su comportamiento asociado son homogéneos semánticamente. El punto es que los proyectos de M&E bajo monitoreo, pueden ser variados, al igual que sus entidades, incluso en la definición de sus atributos se podría utilizar sinónimos. En este aspecto, es poco probable que en dos proyectos que monitorean entidades, tengan dos atributos que se encuentren definidos en español exactamente en los mismos términos y sin tener alguna variante. Lo que podría ser probable, es que en dos proyectos dados en los cuales son monitoreadas entidades, sean definidos dos atributos de entidades diferentes utilizando conceptos eventualmente equivalentes o sinónimos.

Considerando lo antes mencionado, podemos determinar que la principal dificultad, se presenta en la imposibilidad de determinar la similitud semántica a partir de atributos que caracterizan (en español), las entidades bajo monitoreo como así también organizar la experiencia previa ó conocimiento previo, por similitud semántica en memoria, de manera que permita brindar recomendaciones. Esta situación, lleva a plantear el desarrollo de una estrategia de análisis semántico sobre la definición de los atributos que caracterizan la entidad bajo monitoreo, de manera de poder determinar si las mismas son similares o no en concepto y de este modo poder brindar recomendaciones para la toma de decisiones en tiempo real.

Desarrollo

Establecer la estrategia de similitud, que se defina a partir de repositorios de medición y evaluación basados en el marco formal de medición y evaluación C-INCAMI (Context-Information Need, Concept Model, Attribute, Metric,

and Indicator) tiene por fin, localizar entidades semánticamente similares que permitan reutilizar su conocimiento previo y experiencia en el proceso de toma de decisiones.

El marco C-INCAMI (L. Olsina, F. Papa and H. Molina, 2007) (M. Mendonca and V. Basili, 2000), es un marco conceptual el cual define los conceptos, términos y relaciones necesarias para especificar un proceso de M&E. De este modo, se logra a través de su utilización, comprender el concepto de lo que una entidad representa, la caracterización de esta a través de sus atributos, la cuantificación de atributos mediante métricas junto con la interpretación de los mismos, a través de sus indicadores. Dicho marco, toma como guía la estrategia GOCAME (Goal-oriented Context-aware Measurement and Evaluation) (Becker, 2014)la cual es una estrategia multipropósito, sensible al contexto para especificar un proyecto de medición basado en el marco C-INCAMI.

La estrategia de similitud semántica incorpora un comportamiento que permite ante una situación típica, poder proveer recomendaciones y-o recomendar cursos de acción basado en la experiencia previa. A los efectos de establecer los mecanismos de organización, fueron definidos los coeficientes de similitud semántica (Diván, M., 2018) de manera de poder de este modo llevar a cabo las mejores recomendaciones.

Sin embargo, podría suceder que, para la entidad bajo monitoreo, no exista experiencia previa, es decir que se cuente ante la presencia de una nueva situación. En ese caso, no se podría brindar sugerencias o recomendaciones. Para poder brindar solución a ello, se plantea la idea de detectar entidades similares (Sánchez Reynoso, M. & Diván, Mario., 2019) estableciendo un puntaje de similitud que permita brindar recomendaciones por analogía con otra entidad, en el caso de que no exista experiencia previa en la entidad la cual es objeto del análisis.

Teniendo en cuenta lo antes mencionado, se trabajó en la incorporación de diferentes componentes en la arquitectura PAbMM (Diván, M & Sánchez-Reynoso, M, 2019), llevando a cabo un proceso iterativo, de manera de poder modificar la misma en forma gradual.

Resultados

Considerando todas las actividades de investigación que fueron llevadas a cabo dentro del grupo de investigación Ciencia de Datos, se obtuvieron buenos resultados, debidamente publicados como parte del plan de trabajo correspondiente a las actividades del Doctorado.

Los resultados obtenidos, fueron presentados en revistas internacionales, conferencias nacionales e internacionales y en capítulos de libros asociados a la temática de investigación. Es importante destacar que, de los artículos de revistas presentados, tres fueron aceptados en revistas Q1, según clasificación Scimago.

En este sentido, es importante mencionar los avances y resultados publicados desde el año 2017 a la fecha, tal como se visualizan en la Figura 1, de forma cronológica para su referencia.

Figura 1: Descripción gráfica de los diferentes tipos de publicaciones realizadas

Como se puede observar en la figura anterior, hubo una gran participación en las diferentes actividades realizadas dentro del Grupo de Investigación, lo que permitió la interacción continua con otros investigadores al momento de trabajar en forma conjunta con la escritura de publicaciones en revistas indexadas y capítulos de libros.

Desde el inicio del Doctorado a la fecha, se puede observar un avance incremental y productivo en cuanto a los objetivos propuestos, dado que los mismos fueron alcanzados con holgura, permitiendo poder realizar la escritura de tesis de manera tranquila.

Conclusiones

Al momento de definir el tema de investigación fueron analizados las diferentes alternativas de marcos de medición y evaluación (M&E), indicando en el estado del arte la existencia de distintas alternativas asociadas con los marcos para definir el proceso de M&E. Ello permite a su vez que exista una variación del concepto de similitud semántica de acuerdo con el marco que se seleccione. En este sentido, fue importante analizar la eficacia de cada uno de los marcos en las distintas áreas de aplicación, a los efectos de poder evaluar las ventajas que presentan los mismos.

Es importante tener en cuenta que cuando se deben realizar recomendaciones y-o sugerencias en tiempo real, considerando la similitud semántica de aquellas entidades que se encuentran bajo monitoreo, los datos podrían variar en forma sensible.

Debemos evaluar si la entidad bajo monitoreo cuenta ó no con experiencia previa, o si dicha experiencia, es suficiente para determinadas situaciones en donde se pueda brindar cursos de acción que ayuden a la toma de decisiones. En este sentido, el objetivo que se persigue al desarrollar una estrategia consiste en recuperar las recomendaciones desde entidades que sean similares en término semántico.

Ahora bien, se puede presentar el caso de que una entidad presenta una situación nueva, de la cual no se tiene antecedentes, o bien si la entidad es nueva y carece de historia, la recomendación en PAbMM es guiada localizando entidades estructural y comportamentalmente similares.

A los efectos del análisis semántico, se definirán los coeficientes de similitud estructural y comportamental dentro de PAbMM. Ambos coeficientes son importantes en lo que respecta a la búsqueda y optimización en memoria

para la arquitectura. Lo que se persigue con la definición de dichos coeficientes es, por un lado, permitir conocer sintácticamente, si dos entidades comparten un cierto conjunto de atributos (coeficiente estructural) y, por otro lado, detectar conductas comunes (coeficiente comportamental) entre las entidades de acuerdo con la distribución de datos en tiempo real de cada atributo común entre las mismas.

De este modo, la propuesta de desarrollar una estrategia que permita brindar recomendaciones por similitud semántica en tiempo real, considerando repositorios de grandes volúmenes de datos de medición y evaluación basado en el marco formal C-INCAMI, permie localizar entidades semánticamente similares, a fin de poder reutilizar su conocimiento y experiencia en el proceso de toma de decisiones cuando sea requerido.

Referencias

G. Rossi and D. Schwabe. (2008). Modeling and Implementing Web Applications with Oohdm. *in Web {Engineering}: {Modelling} and {Implementing} {Web} {Applications}*, 109–155.

M. J. Divan and M. de los A. Martin. (2017). A new storm topology for synopsis management in the processing architecture . *XLIII Latin American Computer Conference (CLEI), 2017*, pp. 1-10.

Sánchez-Reynoso, M. L. (2014). *Estrategia de monitoreo de los procesos previsionales del Instituto de Seguridad Social (La Pampa), soportado por un marco conceptual de medición y evaluación, para análisis preventivo en toma de decisiones* . Universidad Tecnológica Nacional - Facultad Regional Córdoba, Córdoba.

Diván, M., & Sánchez Reynoso, M. . (2017). Behavioural Similarity Analysis for Supporting the Recommendation in PAbMM. *1st International Conference on Infocom Technologies and Unmanned Systems (ICTUS).* Dubai.

L. Olsina, F. Papa and H. Molina. (2007). How to Measure and Evaluate Web Applications in a Consistent Way. (O. P. G. Rossi, Ed.) *Web Engineering: Modelling and Implementing Web Applications*, 385-420.

M. Mendonca and V. Basili. (2000, June). Validation of an approach for improving existing measurement frameworks. (484-499, Ed.) *IEEE Transactions on Software Engineering, 26*(6).

Becker, P. (2014). *Process View of the Quality Measurement and Evaluation Integrated Strategies, PhD Thesis.* . National University of La Plata, La Plata.

Diván, M. (2018). A library for articulating the measurement streams with columnar data. . *International Journal of Engineering and Technology (UAE)*, 234-241.

Sánchez Reynoso, M. & Diván, Mario. (2019). A Systematic Literature Mapping on the Similar emantically Entities in Measurement Projects. . *The 13th International Conference on E-Learning and Games - Edutainment.* Cali, Colombia.

Diván, M & Sánchez-Reynoso, M. (2019). An Architecture for the Real-Time Data Stream Monitoring in IoT. (E. S. Tanwar, Ed.) *Multimedia Big Data Computing for IoT Applications: Concepts, Paradigms and Solutions*, 59-100.