ISBN: 978-950-42-0200-4
DOI: https://doi.org/10.33414/ajea.5.751.2020

Recomendación por similitud semántica en repositorios con grandes volúmenes de datos de medición

Semantic Similarity-based Recommendation in Big Data repositories

Presentación: 6 y 7 de Octubre de 2020

Doctorando:
María Laura Sánchez Reynoso
Data Science Research Group, Facultad de Ciencias Económica y Jurídicas, Universidad Nacional de La Pampa - Argentina
mlsanchezreynoso@eco.unlpam.edu.ar

Director/a:
Mario José Diván

Resumen

Los proyectos de medición y evaluación son definidos utilizando un marco formal de medición y evaluación, el cual permita considerar la entidad bajo monitoreo, la cual forma parte de nuestro análisis. Puede suceder, que dicha entidad, no cuente con experiencia o conocimiento previos, provocando esto que no se pueda recomendar sugerencias en situaciones determinadas. En este sentido, y de manera de poder solucionar dicha situación, es que se plantea la idea de recomendar de acuerdo con la similitud semántica que presenten las entidades bajo monitoreo, considerando las mediciones surgidas del proyecto de medición y evaluación definidos previamente. El objetivo es justamente poder 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, indicando los resultados obtenidos en lo que respecta a similitud semántica, coeficientes de similitud, estrategias de recomendación y actualización de diferentes componentes de la arquitectura de procesamiento de flujo de datos.

Palabras clave: Similitud Semántica, Repositorio, Datos, Medición

Abstract

The measurement and evaluation projects are defined using a formal measurement and evaluation framework, which allows us to consider the entity under monitoring, which is part of our analysis. It may be that the entity under monitoring has no previous experience or knowledge, which means that no suggestions can be recommended in certain situations. In this sense, and to solve this situation, the idea is to recommend according to the semantic similarity presented by the entities under monitoring, considering the measurements arising from the project of measurement and evaluation previously defined. The objective is precisely to be able to detect similar entities under monitoring by establishing a similarity score that allows for the provision of experiences and/or recommendations by analogy with another entity in the absence of these in the entity under analysis. Here we present a partial advance of this line of research, indicating the results obtained in terms of semantic similarity, similarity coefficients, recommendation strategies, and updating of different components of the data flow processing architecture.

Keywords: Semantic Similarity, Repositories, Big Data, Measurement

Introducción

El contexto de búsqueda en un repositorio de grandes volúmenes de datos, basado en un almacenamiento persistente y teniendo en cuenta los tiempos de acceso, es decir lectura y escritura son completamente diferentes del acceso en memoria. En los motores de procesamiento de flujo de datos, el dato será leído y procesado hasta una vez, las decisiones se toman en tiempo real, los tiempos de acceso están asociados con la memoria y todo ello deberá ocurrir dentro de los límites establecidos por los recursos con los que se dispone en ese instante (por ejemplo, memoria, procesador, etc.).

Ahora bien, cuando una entidad no posee eventualmente experiencia previa, la idea en PAbMM (Diván & Sánchez Reynoso, Behavioural Similarity Analysis for Supporting the Recommendation in PAbMM, 2017) consiste en reducir el espacio de búsqueda en memoria respecto de las experiencias previas y conocimiento de las entidades vinculadas. Así, aun cuando no exista experiencia previa específica, PAbMM, podrá recomendar curso de acción vinculados con entidades cuya estructura y comportamiento son similares a la monitoreada y recomendar por similitud. El análisis de similitud estructural se calcula a partir de los metadatos asociados con la definición del proyecto de M&E (Schwaber & Sutherland, 2017) basado en C-INCAMI. Sin embargo, puede suceder que dos entidades estructuralmente similares, se comporten en forma diferente. En tal caso, dicha situación es abordada mediante el coeficiente de similitud comportamental, el cual es obtenido mediante análisis de correlación multivariado a partir de las métricas que cuantifican el comportamiento de la entidad (Diván & Sánchez Reynoso, Behavioural Similarity Analysis for Supporting the Recommendation in PAbMM, 2017; Diván M. , Data-Driven Decision Making, 2017).

Teniendo en cuenta lo antes mencionado, en lo que se refiere a coeficientes (estructurales y comportamentales) y a las entidades con y sin experiencia previa, cabe plantear el siguiente interrogante ¿Cómo se puede recomendar cursos de acción por similitud semántica en repositorios de grandes volúmenes de datos? De este modo, el objetivo del presente trabajo de investigación consiste en recomendar cursos de acción en memoria, a partir de repositorios que contienen grandes volúmenes de datos de medición y evaluación basados en el marco formal C-INCAMI, a fin de poder detectar entidades semánticamente similares, y de este modo reutilizar su conocimiento y experiencia en el proceso de toma de decisiones en tiempo real cuando sea requerido.

Desarrollo

Dentro del marco del grupo de investigación CIEDAyAp (Grupo de investigación en Ciencias de Datos y sus Aplicaciones), se puedo establecer en forma general las pautas correspondientes a la estrategia de similitud semántica (Sánchez Reynoso, M & Diván, M, 2019) donde el objetivo es poder brindar recomendaciones desde entidades que sean similares semánticamente.

En lo que respecta a la estrategia de priorización de contenido, se pudo establecer en forma parcial los lineamientos de definición de dicha estrategia que permitirá, llevar a cabo la priorización del conocimiento/experiencia previa de la entidad bajo análisis. (Diván M. &., 2018) y se aplicaron técnicas de visualización de datos, (Sánchez Reynoso M. &., 2020) de manera de poder contar con una serie de guías que permitan mostrar los datos de las mediciones obtenidas. A la hora de plantear los mecanismos de organización, se definieron previamente los coeficientes de similitud semántica y sintáctica, de modo de poder diseñar dichos mecanismos y permitir de este modo llevar a cabo las mejores recomendaciones. (Diván M. &., A library for articulating the measurement streams with columnar data, 2018). Por otro lado, en lo que respecta a la implementación de la estrategia de recomendación, se trabajó en forma en forma parcial en base al análisis de los mecanismos que se encuentran definidos también parcialmente. (Diván M. &., Extending the Data Stream Processing Strategy to Scenario Analysis, 2019). La idea es trabajar en forma iterativa para ir incorporando diferentes componentes a PAbMM, que permitan modificar el mismo en forma gradual. (Diván M. &., 2019)

Resultados

Teniendo en cuenta las actividades de investigación llevadas a cabo dentro del grupo de investigación CIEDAAyAp (Ciencias de Datos y sus Aplicaciones), se obtuvieron muy buenos resultados, en lo que respecta al avance de cada una de las actividades asociadas con el Doctorado.

En seste sentido es importante mencionar que se realizaron cursos de posgrado relevantes a la temática del plan de trabajo del Doctorado, obteniendo muy buenas calificaciones en cada una de las evaluaciones individuales de los mismos. Por otro lado, se publicaron una serie de numerosos artículos internacionales, a medida que se avanza con el desarrollo de coeficientes semánticos

y sintácticos, permite ir mostrando en forma gradual el avance de la investigación asociados con conceptos tales como similitud semántica, medición, evaluación, entre otros.

Se llevó a cabo también la participación en congresos, conferencias (Sánchez Reynoso M. &., A Systematic Literature Mapping on the Similar emantically Entities in Measurement Projects, 2019) (Diván M. &., 2018), (Sánchez Reynoso M. &., 2020) (Diván M. &., 2019) (Sánchez Reynoso M. &., 2019), seminarios y jornadas, a fines con la temática, lo que permitió exponer cada uno de los artículos publicados (Sánchez Reynoso M. &., Assesment of semantic similarity in entities under monitoring: A systematic literature mapping, 2020), (Diván M. &., Strategies based on IoT for supporting the decisison making in Agriculture: A Systematic Literature Mapping, 2020) y obtener feedback de otros investigadores con respecto a los avances y nuevas actualizaciones asociados a los temas expuestos. Esto ayudó a trabajar en forma colaborativa con investigadores y grupos de trabajo de otras Universidades.

La participación en forma activa dentro del Grupo de Investigación denominado CIEDAAyAp, facilitó la interacción en forma continua con otros investigadores y con el director de tesis, para poder avanzar en la escritura de publicaciones en revistas indexadas (Sánchez Reynoso M. &., Improving the Rela-Time Searching in the Organizational Memory, 2019) (Diván M. &., A library for articulating the measurement streams with columnar data, 2018) y capítulos de libros, (Diván M. &., Managing the Data Meaning in the Data Stream Processing: A Systematic Literature Mapping, 2020) (Diván M. &., The Real-Time Measurement and Evaluation as System Reliability Driver, 2018) (Diván M. &., An Architecture for the Real-Time Data Stream Monitoring in IoT, 2019) como así también ir delineando la estrategia para comenzar con la escritura de tesis.

El año 2019 fue bastante productivo y satisfactorio, en cuanto los objetivos propuestos, publicaciones realizadas (Diván M. &., Extending the Data Stream Proecssing Strategy to Scenario Analysis, 2019) (Diván M. &., 2020) y asistencia a congresos, dado que los mismos fueron alcanzados con holgura, permitiendo ello poder avanzar de manera tranquila e ir organizando y delineando la estrategia de escritura de tesis del Doctorado.

Conclusiones

De acuerdo con el avance en forma gradual, del trabajo de investigación llevado a cabo, podemos determinar que en lo que respecta a marcos de medición y evaluación, existen diversidad en los mismos. Sin embargo, es importante tener en cuenta a la hora de recomendar acciones en tiempo real, que existen diferentes opciones por medio de las cuales se puede llevar a cabo un determinado proceso de medición y evaluación, provocando ello que las recomendaciones por similitud semántica en repositorios de grandes volúmenes de datos varíen de manera sensible.

En términos de la línea de investigación, dentro del grupo de investigación CIEDAAyAp (Ciencias de Datos y sus Aplicaciones), se generaron resultados en forma parcial tenido en cuenta entidades bajo monitoreo, actualizaciones a la arquitectura de procesamiento de flujos de datos basado en escenarios.

Por otro lado, se establecieron lineamientos de definición de la estrategia de priorización de contenido, pautas asociadas a la estrategia de búsqueda por similitud, se aplicaron técnicas de visualización de los datos, se plantearon mecanismos de organización y finalmente un mapeo sistemático de la literatura fue llevado a cabo para continuar con el análisis de las diferentes estrategias, recomendaciones y enfoques, asociados a la similitud semántica en repositorios de grandes volúmenes de datos de medición y evaluación.

Como trabajo a futuro, se continuará con el análisis de similitud semántica, procesos de medición y evaluación, sobre el procesamiento de flujos de datos, permitiendo establecer recomendaciones en tiempo real.

Referencias

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

Diván, M. &. (2018). A library for articulating the measurement streams with columnar data. *International Journal of Engineering and Technology (UAE)*, 7(4.31):234-241. Obtenido de https://www.sciencepubco.com/index.php/ijet/article/download/23373/11680

Diván, M. &. (2018). A Load- Shedding Technique based on the Measurement Project Definition. *5th international Conference on Intelligent Computing, Communications & Devices (ICCD)* (págs. 1027-1033). Xi'an, China: Springer Nature, Singapore. doi:10.1007/978-981-13-9406-5 122

Diván, M. &. (2018). The Real-Time Measurement and Evaluation as System Reliability Driver. En A. &. Anand (Ed.), *System Reliability Management Solutions and Technologies* (págs. 161-188). CRC Press Taylor & Francis Group. doi:10.1201/9781351117661-11

Diván, M. &. (2019). An Architecture for the Real-Time Data Stream Monitoring in IoT. En S. a. S. Tanwar (Ed.), *Multimedia Big Data Computing for IoT Applications: Concepts, Paradigms and Solutions* (págs. 59-100). Springer. doi:10.1007/978-981-13-8759-3_3

Diván, M. &. (2019). Extending the Data Stream Processing Strategy to Scenario Analysis. *International Journalof Advanced Trends in Computer Science and Engineerign (IJATCSE)*, 1-8.

Diván, M. &. (2019). Extending the Data Stream Proecssing Strategy to Scenario Analysis. *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, 8(1.4):1-8. doi:10.30534/ijatcse/2019/0181.42019

Diván, M. &. (2019). Incorporatins Scenarios and States Definitions on Real-Time Entity Monitoring in PAbMM. *XLV Latin American Computer Conference - CLEI.* Panamá. doi:10.1109/CLEI47609.2019.235072.

Diván, M. &. (2020). Dynamic Switching in the Measurements'Collecting from Hetrogeneous Data Sources. *Journal of Physics: Conference Series. Institute of Physics*. doi:10.1088/1742-6596/1529/2/022058

Diván, M. &. (2020). Managing the Data Meaning in the Data Stream Processing: A Systematic Literature Mapping. En J. V. P. Johri (Ed.), *Applications of Machine Learning* (págs. 31-46). Springer Singapore. doi:10.1007/978-981-15-3357-0_3

Diván, M. &. (2020). Strategies based on IoT for supporting the decisison making in Agriculture: A Systematic Literature Mapping. *International Journal of Reasoning-based Intelligent Systems Inderscience*. Obtenido de https://www.inderscience.com/info/ingeneral/forthcoming.php?jcode=ijris

Diván, M. (2017). Data-Driven Decision Making. *1st International Conference on Infocom Technologies and Unmanned Systems (ICTUS).* Dubai.

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.

Sánchez Reynoso, M & Diván, M. (2019). Improving the Real-Time Searching in the Organizational Memory. *8th International Congress of Information and Communication Technology (ICICT)* (págs. 293-304). Procedia Computer Science. Elsevier Ltd.

Sánchez Reynoso, M. &. (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.

Sánchez Reynoso, M. &. (2019). Contributions to the Communication of the Officiaal Advertising's Distribution in Argentina. *4th International Conference on Information Systema and Computer Networks (ISCON).* India. doi:10.1109/ISCON47742.2019.9036298

Sánchez Reynoso, M. &. (2019). Improving the Rela-Time Searching in the Organizational Memory. *Procedia Computer Science, 154*, 293-304. doi:doi.org/10.1016/j.procs.2019.06.043

Sánchez Reynoso, M. &. (2020). Applying Data Visualization Guideline on Forest Fires in Argentina. *10th international Conference - CONFLUENCE Department of Computer Science and Engineering.* Uttar Pradesh, India: Amity University. Obtenido de http://itekcmsonline.com/rps2prod/confluence2020/eproceedings/html/269.xml

Sánchez Reynoso, M. &. (2020). Assesment of semantic similarity in entities under monitoring: A systematic literature mapping. *Revista de Universidad de Antioquía*. doi:10.17533/udea.redin.20200476

Schwaber, K., & Sutherland, J. (2017). *The Scrum Guide. The Definitive Guide to Scrum: The Rules of the Game.* Ken Schwaber and Jeff Sutherland.