CIENCIA DE DATOS COMO HERRAMIENTA DE SOPORTE EN LA GESTIÓN PÚBLICA DE CALIDAD DEL AGUA

J. J. Bolano, M. G. Rey, U. Ramirez, J. G. A. Pautsch, E. Zamudio, H. D. Kuna

Instituto de Investigación Desarrollo e Innovación en Informática (IIDII)
Facultad de Ciencias Exactas, Químicas y Naturales, Universidad Nacional de Misiones

hdkuna@gmail.com

RESUMEN

El agua es un recurso vital y como tal requiere una adecuada gestión de su calidad para la subsistencia de la población en todas sus dimensiones. En la provincia de Misiones, existen varias organizaciones que intervienen de alguna manera en la gestión de la calidad del agua, aunque se encuentran limitados respecto de la generación de información adecuada para la gestión, pese a esfuersos individuales de algunas de estas organizaciones.

La gestión de calidad del agua es un proceso que requiere datos. Estos datos son generados a un alto costo, ya que cada estudio implica un traslado al lugar donde se realizan las muestras, gastos en recursos humanos, y gastos en el análisis de las propiedades del agua para determinar su calidad, entre otros.

En esta línea de investigación se aborda la aplicación de procesos de ciencia de datos que podrían contribuir en la generación de información dirigida a la gestión de calidad del agua.

A través de un análisis descriptivo se determinaron parámetros asociados a la calidad del agua, destacándose la temperatura, turbiedez, conductividad, pH, P-ortoFosfatos, solidos disueltos totales, oxígeno disuelto, sólidos suspendidos totales y coliformes fecales.

Asismismo, se evidencia la necesidad de ampliar el desarrollo de conjuntos de datos que integren variables relevantes para la gestión de la calidad del agua.

CONTEXTO

Esta línea de investigación se desarrolla en el ámbito Programa de Investigación en Computación (PICom), perteneciente al Instituto de Investigación Desarrollo e Innovación en Informática de la Facultad de Ciencias Exactas Químicas y Naturales, Universidad Nacional de Misiones (IIDII, FCEQyN, UNaM). El PICom desarrolla líneas de investigación relacionadas con la explotación de información y la robótica.
El grupo de trabajo aborda principalmente las áreas relacionadas con el tratamiento de datos y el descubrimiento de información a partir de éstos mediante técnicas de ciencia de datos, en distintos campos de aplicación. Este contexto ha permitido contribuir con estrategias para la detección de datos anómalos, expansión de consultas, algoritmos de ranking, generación automática de perfiles y desambiguación de entidades, recomendación y selección de grupos de expertos.

1. INTRODUCCION

Una adecuada gestión de calidad del agua es determinante para la subsistencia de la población en todas sus dimensiones. Esto es evidente, en general debido a la dependencia del recurso para la vida, y en particular, debido a los usos del recurso en diversas aplicaciones como el consumo humano, el desarrollo agrícola, industrial, y recreativo.

 RedUNCI - UNdeC. ISBN: 978-987-24611-3-3

En la Argentina, existen varios organismos que intervienen de alguna manera en la gestión de la calidad del agua, algunos inclusive con presencia en la Provincia de Misiones. Estos organismos incluyen al Instituto Nacional del Agua (INA), la Comisión Mixta del Río Paraná (COMIP), la Entidad Binacional Yacyretá (EBY), la Universidad Nacional de Misiones (UNaM), y varias otras dependencias del estado y organizaciones privadas, entre otros. En particular, la EBY ha desarrollado un Sistema de Información Geográfica (GIS) llamado GNOMA, con el objetivo de mantener datos de calidad del agua proveniente de diversas organizaciones. Por el momento, GNOMA carece de herramientas que permitan la generación de conocimiento a partir de sus datos contenidos para su uso en la gestión de la organización. Algunos ejemplos de posibles aplicaciones de generación de conocimiento, incluyen la posibilidad de clasificar zonas en forma automática que permitan indicar el uso potencial del agua en dicha zona.

Las características geográficas de provincia de Misiones, en relación a su carácter mesopotámico definido por sus ríos, arroyos, reservas, así como a sus límites geográficos nacionales e internacionales, establecen un especial interés por la gestión de la calidad del agua.

Mantener una adecuada gestión de calidad del agua implica entre otras importantes actividades, disponer de información relevante, confiable y oportuna. Sin embargo, la diversidad de organizaciones y la falta de información consolidada y disponible en relación a la calidad del agua dificultan la tarea de la gestión.

Salvo excepciones, los esfuerzos individuales de las organizaciones en relación a la recolección y mantenimiento de datos de calidad del agua, difícilmente son traducidas en información útil para la gestión. Más aún, la gran cantidad de datos relacionados con la gestión de calidad del agua mantenidos por estas organizaciones, junto con la diversidad de sus estrategias de almacenamiento, dificultan la tarea de producir información a partir de dichos datos.

En la actualidad se evidencia una tendencia respecto del uso de la ciencia de datos aplicada a la gestión en diversos ámbitos. En particular, el crecimiento de datos disponibles y los costos en el poder de cómputo, permitieron un gran desarrollo de estrategias de ciencia de datos. Estas estrategias de ciencia de datos, permiten la generación de conocimiento a partir del descubrimiento de patrones en grandes conjuntos de datos. Asimismo, estas estrategias permiten el desarrollo de modelos de predictivos con precisiones que superan al humano en varias tareas específicas, como el reconocimiento de patrones o la predicción de variables cuantitativas.

En resumen, una adecuada gestión de calidad del agua requiere de estrategias que permitan generar conocimiento a partir de datos caracterizados por un gran volumen y heterogeneidad.

Como consecuencia, resulta necesario desarrollar estrategias que permitan generar conocimiento a partir de datos asociados a la gestión de calidad del agua. Sin embargo, hasta el momento, organizaciones relevantes en la provincia de Misiones aún no disponen de recursos avanzados para la aplicación de procesos relacionados con ciencia de datos que le permitan un aprovechamiento de los datos asociados a la calidad del agua, los cuales son particularmente costosos de generar.

2. LINEAS DE INVESTIGACION Y DESARROLLO

En la actualidad se evidencia una tendencia respecto del uso de la ciencia de datos

aplicada a la gestión en diversos ámbitos. Estas aplicaciones incluyen áreas de importancia como la agricultura [1].

El crecimiento de datos disponibles y los costos en el poder de cómputo, permitieron un gran desarrollo de estrategias de ciencia de datos. Estas estrategias de ciencia de datos, permiten la generación de conocimiento a partir del descubrimiento de patrones en grandes conjuntos de datos. Asimismo, estas estrategias permiten el desarrollo de modelos de predictivos con precisiones que superan al humano en varias tareas específicas, como el reconocimiento de patrones o la predicción de variables cuantitativas.

Recientemente se conoció la creación de un emprendimiento a nivel global denominado Aquagenuity [2], el cual busca proveer servicios de información relacionados con la calidad del agua. Por otra parte, en el ámbito del aprendizaje automático, la compañía SomData ha desarrollado una inteligencia artificial para la detección de microorganismos en el agua [8]. Otras organizaciones como el Group on Earth Observations (GEO) desarrollan trabajos en relación al monitoreo, gestión y toma de decisiones relacionados con la calidad del agua [3].

La disponibilidad de datos es esencial para proyectos de ciencia de datos. En este sentido, existen esfuerzos internacionales para la creación, mantenimiento y disposición pública de datos relacionados con calidad del agua como es el caso del paquete de datos Global Open Data for Agriculture and Nutrition [4].

Actualmente es frecuente el uso de grandes cantidades de datos (Big Data) en varios dominios de aplicación. Esto también ocurre en el ámbito de la gestión del agua, como en el análisis de datos para la gestión de aguas urbanas [5].

Esta línea de trabajo aborda el desarrollo de procesos de ciencia de datos que permitan la identificación de patrones y generación de información dirigida a la toma de decisiones en relación a la gestión de la calidad del agua.

Esta línea de trabajo incluye los siguientes objetivos:

-Identificar las necesidades específicas en relación a la gestión pública de la calidad del agua en el ámbito de la provincia de Misiones.

-Determinar las fuentes y las características de los datos asociados a la gestión pública de la calidad del agua.

-Desarrollar mecanismos de procesamiento de datos disponibles que permitan su uso en modelos de ciencia de datos dirigidos a la gestión de la calidad del agua.

-Desarrollar estrategias de ciencia de datos, incluyendo el entrenamiento de modelos de aprendizaje automático, para dar soporte a los requerimientos de información de la gestión de la calidad del agua.

-Evaluar el impacto de la implementación de los procesos desarrollados de ciencia de datos en organizaciones intervinientes en la gestión de calidad del agua.

3. RESULTADOS OBTENIDOS/ESPERADOS

Hasta el momento se trabajó con el procesamiento de un conjunto de datos de calidad del agua del Río Paraná censados en el intervalo de tiempo para los años 1980 a 1990. A través de un análisis descriptivo se determinaron parámetros asociados a la calidad del agua, destacándose la temperatura, turbiedez, conductividad, pH, P-ortoFosfatos, solidos disueltos totales, oxígeno disuelto,

sólidos suspendidos totales y coliformes fecales.

Un análisis de la distribución de frecuencia de los valores avistados con los expertos en la calidad del agua determinó que no siguen una distribución normal estándar y no existe un marco de referencia sobre el cual poder contrastar los datos actualmente, dado que la naturaleza del Río Paraná es única en su contexto geográfico e hidrológico.

La ausencia de un marco de referencia para poder analizar el comportamiento de los datos incentivó a tomar la iniciativa en la reconstrucción de los datos históricos del Río Paraná a través de la articulación de los datos disponibles que disponen las organizaciones que se encargan de medir la calidad del agua del río.

Además de organizar los datos disponibles que se encuentran dispersados en varias entidades, otro desafío que enfrenta este proyecto es la propia naturaleza del Río Paraná, que posee un gran nivel de auto-depuración debido a su tamaño y fluidez. Un mismo parámetro de calidad del agua puede variar excesivamente en su medición si se realiza en un instante posterior y a una escasa distancia de la medición original.

Para definir el proceso de ciencia de datos a aplicar, [6] realiza un mapeo sobre la literatura relacionada con la predicción de la calidad del agua haciendo uso de técnicas de inteligencia computacional (IC), donde se detecta que la mayor proporción de trabajos estudian métodos híbridos que involucran más de una tecnología de IC para la construcción de un modelo predictivo de calidad de agua. En el mismo estudio se detecta la tendencia a la investigación de técnicas de Redes Neuronales Artificiales (RNA) para la construcción de modelos predictivos.

4. FORMACIÓN DE RECURSOS HUMANOS

El grupo de trabajo se compone de un equipo interdisciplinario, el cual desde la especialidad de las áreas de trabajo de sus integrantes contribuyen a la generación de conocimiento en el área de ciencia de datos por un lado, y en la gestión de calidad del agua por otro. Asimismo, integrantes del equipo de trabajo han colaborado en forma conjunta en objetivos comunes relacionados con el área de calidad del agua.

En particular a la gestión de calidad del agua, el grupo ha desarrollado actividades de colaboración con la Comisión Mixta del Río Paraná (COMIP), organismo del estado nacional, para procesos de digitalización de datos de calidad del agua a partir de informes históricos. A partir de dichas actividades, se desarrollaron procesos de control de calidad de datos dirigidos a la identificación desvíos, datos faltantes y anómalos.

Estas actividades implicaron el desarrollo de procesos específicos para garantizar la compatibilidad y calidad de los datos para su exportación a un Sistema de Información Geográfica gestionado por la Entidad Binacional Yacyretá (EBY).

Las líneas de investigación presentadas cuentan con doce integrantes relacionados con las carreras de Ciencias de la Computación de la UNaM. El grupo de investigación desarrolla dos tesis de grado articulando sus trabajos con una beca de Estímulo a las Vocaciones Científicas del Consejo InterUniversitario Nacional (CIN) y una beca UNaM; dos tesis de maestría en curso y una finalizada. Asimismo, las líneas de investigación y sus integrantes se vinculan con grupos de la Universidad de Castilla-La Mancha, España y la Universidad de Sonora, México.

5. BIBLIOGRAFÍA

[1] J. J. Dabrowski, A. Rahman, A. George, S. Arnold, and J. McCulloch, "State Space Models for Forecasting Water Quality Variables," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*, 2018, pp. 177–185.

[2] D. Lazovskam, "Aquagenuity, plataforma de calidad del agua – ExpokNews," 2019. [Online]. Available: https://www.expoknews.com/aquagenuity-plataforma-de-calidad-del-agua/. [Accessed: 08-Aug-2019].

[3] G. El Serafy, A. Taylor, and S. Greb, "Water Quality Aquatic Services: Use of Data Science in the Aquawatch GEO Initiative," in *20th EGU General Assembly, EGU2018, Proceedings from the conference held 4-13 April, 2018 in Vienna, Austria, p.19859*, 2018, vol. 20, p. 19859.

[4] Global Open Data for Agriculture and Nutrition, "Introducing the Agricultural Open Data Package: BETA Version | GODAN," 2016. [Online]. Available: https://www.godan.info/news/introducing-agricultural-open-data-package-beta-version. [Accessed: 08-Aug-2019].

[5] R. Harsh, G. Acharya, and S. Chaudhary, "Scope of Big Data Analytics in Bikaner Urban Water Management," *Int. J. Comput. Intell. IoT*, vol. 2, no. 3, 2018.

[6] I. D. Lopez, A. Figueroa, and J. C. Corrales, "Un mapeo sistemático sobre predicción de calidad del agua mediante técnicas de inteligencia computacional," *Rev. Ing. la Univ. Medellín*, vol. 15, no. 28, pp. 35–52, 2016.