Ficha de Resumen de Artículo Científico: A Data Quality Assessment Model and Its Application to Cybersecurity Data Sources

Ficha elaborada por: Ines Salamanca Estévez

Título del artículo: A Data Quality Assessment Model and Its Application to Cybersecurity Data Sources

Autor/es del artículo: Noemí DeCastro-García, Enrique Pinto

Año: 2020

Enlace: https://link.springer.com/chapter/10.1007/978-3-030-57805-3_25

Resumen

Este artículo presenta un modelo multidimensional para evaluar la calidad de los datos, crucial para sistemas de Big Data y fuentes de ciberseguridad. Propone un conjunto de dimensiones de calidad, una metodología de evaluación y una fórmula matemática para generar una puntuación de calidad. Se desarrolló una herramienta de software en Python que automatiza este proceso, aplicándose exitosamente a un conjunto real de datos de eventos de ciberseguridad para clasificar y mejorar la fiabilidad de las fuentes de información.

Ideas principales

- **Necesidad de Evaluación de Calidad**: Ante la proliferación de sistemas Big Data y la dependencia en múltiples fuentes (especialmente en ciberseguridad), es imprescindible evaluar la calidad de los datos para asegurar la toma de decisiones adecuadas.
- **Modelo Multidimensional y Metodología**: El estudio define un modelo que evalúa la calidad de los datos a través de diversas dimensiones (ej., cantidad, completitud, veracidad, consistencia, frecuencia, relevancia y precio), utilizando métodos de medición tanto automáticos como manuales.
- **Puntuación y Clasificación de Fuentes**: Se aplica una fórmula matemática para calcular una puntuación global de calidad para cada fuente de datos, permitiendo obtener una clasificación (ranking) que compara su fiabilidad y utilidad.
- **Desarrollo de Herramienta Software**: Se creó una aplicación en Python (disponible en GitHub) que automatiza la evaluación de la calidad de los datos según el modelo propuesto, facilitando el análisis rápido y fiable de grandes volúmenes de información.
- **Aplicación Práctica y Conclusión**: El modelo y la herramienta se validaron con éxito en un dataset real de 27 fuentes de eventos de ciberseguridad. Los resultados permiten identificar fuentes de datos deficientes o mejorables, orientando decisiones estratégicas y la mejora continua de la calidad de los datos.

Material usado

Datos:

- **Dataset real de eventos de ciberseguridad**: Compuesto por 25,446,964 registros (filas) con 113 características (columnas), provenientes de 27 fuentes diferentes y cubriendo aproximadamente 24 horas. Este dataset fue proporcionado por INCIBE bajo acuerdo de confidencialidad.

Software:

- Sistema operativo: Windows 10.
- Lenguaje de programación: Python (distribución Anaconda 2.7 y 3.7).
- Aplicación desarrollada: RIASC Python application (código fuente alojado en GitHub bajo licencia GNU GPLv3).

Hardware:

- PC con procesador Intel i5 y 8 GB de RAM.

Herr. matemáticas:

- **Función de Clasificación por Colores (Ecuación 1):** $\text{ColourSj(Di)} = \begin{cases} \text{Green} & \text{si VSj(Di)} > \text{bi} \\ \text{Yellow} & \text{si VSj(Di)} \in \text{Ii} \\ \text{Red} & \text{si VSjDi)} < \text{ai} \end{cases}$ para determinar el estado de cada dimensión de calidad respecto a umbrales.
- **Fórmula para la Puntuación de Calidad Global (Ecuación 2):** $\text{VSj} = \text{w+} \cdot \frac{\#\text{GreenSj(Di)}}{\#\text{(Di)}} + \text{w0} \cdot \frac{\#\text{YellowSj(Di)}}{\#\text{(Di)}} + \text{w-} \cdot \frac{\#\text{RedSj(Di)}}{\#\text{(Di)}}$.
  Esta fórmula combina las evaluaciones de las dimensiones individuales para obtener una puntuación total por fuente.
- **Normalización de Datos:** División por el número total de datos para obtener valores en el rango $[0, 1]$ para ciertas dimensiones (cantidad, completitud, nivel de información, veracidad desconocida y veracidad).

Palabras clave

Ej.: aprendizaje automático, redes neuronales, diagnóstico, imágenes médicas

Referencias relevantes

1 ISO/IEC 25000 (series de estándares de calidad de productos y sistemas de software).

2,4 ISO/IEC 25012 - Data Quality Model (Estándar de dimensiones de calidad de datos).

5 Referencia a un estudio que propone seis dimensiones primarias de calidad de datos (completitud, unicidad, puntualidad, validez, precisión y consistencia).

Observaciones / Comentarios

Estudio de la calidad de los datos del MICs.