Proyecto de laboratorio

Calidad de datos de una librería

Fecha de Entrega: 29/10, hora 23:59

Realidad planteada

El trabajo estará enfocado en la calidad de datos (CD) sobre libros y conceptos relacionados con éstos.

Consideremos una librería que tiene la necesidad de gestionar la calidad de sus datos. Esta librería es el resultado de la unificación de dos librerías que eran totalmente independientes hasta que la más grande compró a la más pequeña, transformándose en una única librería. Cada una de las librerías originales tenía sus bases de datos y uno de los desafíos que enfrenta la nueva librería es la unificación de todos los datos.

Los encargados de la librería saben que los datos tienen muchos problemas de calidad y que éstos se verán potenciados por la integración de los *datasets* provenientes de ambas librerías, pero no saben cómo enfrentar estos problemas. Nuestro trabajo es evaluar la calidad de los datos de la base de datos integrada y dejar un conjunto de especificaciones que permitan la mejora de la calidad de los datos actuales y futuros.

Los datasets de las librerías (que serán publicados en el sitio EVA del curso) son archivos en formato csv que cumplen con las siguientes estructuras:

Librería 1:

L1-Books (title, description, authors, publisher, publishedDate, categories)

Librería 2:

L2-Books (ISBN, Book-Title, Book-Author)

L2-Ratings (User-ID, ISBN, Book-Rating)

Para la nueva librería se construirá una tabla integrada, con la siguiente estructura:

NL-Books (ISBN, title, description, authors, publisher, publishedDate, categories, avg-rating)

De ahora en más, nos referiremos a cada una de las librerías que serán integradas como L1 y L2, mientras que la nueva librería será nombrada como NL.

Se sabe que en la NL trabajarán 3 usuarios: un administrador, un publicista digital y un analista de datos. El administrador se encargará de la gestión de los datos de la librería, el publicista realizará tareas de recomendación y promoción de libros en el sitio Web, y el analista se encargará del análisis de los datos buscando estudiar comportamientos, preferencias y relaciones entre los clientes.

El objetivo final es que los datos de NL, cumplan con las siguientes propiedades:

- Cada libro deberá tener asociado un título y al menos un autor. Por otro lado, en esta librería pretenden tener al menos 500 libros y el 20% de ellos debe ser parte de la lista de los 100 mejores libros, la cual se encuentra publicada en el sitio Web *Goodreads* [1].
- Desde ya, se sabe que el usuario administrador realizará, con mucha frecuencia, ciertas consultas a la base de datos. Por ejemplo, le interesará conocer los libros cuya publicación sea del año actual, los libros de la editorial Wiley, o el top 3 de los libros con mayor score, según el rating de los lectores.
- Para que las tareas del publicista puedan ser realizadas correctamente es necesario que la base de datos de la NL sea actualizada todos los viernes. Además, este usuario realiza sus tareas esperando que el 60% de los libros tengan al menos un score mayor o igual a 5.
- Por otro lado, para la efectividad de las tareas del analista de datos se especifica que al menos el 95% de los libros debe tener ISBN, sus títulos deben estar correctamente escritos y para los nombres de los autores debe aparecer al menos un nombre y un apellido.
- Finalmente, se destaca que los tiempos de respuesta del sitio Web de la NL no puede superar los 3 segundos.

Tareas a realizar

Este proyecto se enfoca en la ejecución de la Fase 1 – *DQ Planning* y la Fase 2 – *DQ Assessment*, de la metodología de calidad *CaDQM*. Esto implica la ejecución de las siguientes etapas:

Fase 1:

- ST1 - Elicitation
- ST2 – Data Analysis
- ST3 – User Requirements Analysis

Fase 2:

- ST4 – *DQ Model Definition*
- ST5 – *DQ Measurement*
- ST6 – *DQ Assessment*

De la **Fase 1**, algunas tareas ya fueron realizadas. Por lo tanto, contarán con resultados de dichas tareas, que deberán utilizar en su trabajo y serán parte de sus resultados finales. De esta fase, la etapa que deben realizar por completo es **ST2**, mientras que para ST1 y ST3 contarán con partes de las tareas ya realizadas. Para ST1 y ST3 la única referencia será la letra del proyecto, ya que no tenemos información extra de la realidad ni contamos con usuarios reales.

De la **Fase 2**, la etapa principal para este proyecto es **ST4**, mientras que las siguientes etapas serán opcionales (según el alcance al que llegue cada trabajo).

Se pide:

1. **Realizar las actividades de la Fase 1**: Estas actividades incluirán la ejecución de un *Data Profiling*, utilizando alguna de las estrategias vistas en clase. Se deberá obtener como resultado de la Fase 1, el *Modelo de Contexto* (siguiendo los componentes y su clasificación por usuario, vistos en clase) y la lista de *Problemas de Calidad*. Estos resultados incluirán a los resultados intermedios que se entregan en el Anexo de este documento.
2. **Realizar las actividades de la Fase 2, etapa ST4:** Esto implica: i) asignar prioridades a cada uno de los problemas de CD reportados, ii) especificar un Modelo de CD que considere el Modelo de Contexto y el reporte de problemas de CD priorizado. Tener en cuenta que el Modelo de CD debe cumplir los siguientes requisitos:
    a. Contener, al menos, 3 dimensiones de CD
    b. Para cada dimensión, al menos 2 factores de CD
    c. Para cada factor, al menos 1 métrica de CD
    d. Para cada métrica, al menos 1 método de CD con su respectivo método aplicado

En forma Opcional:

3. **Seleccionar uno de los factores de calidad propuestos y ejecutar ST5:** Obtener el valor de CD para al menos 1 método aplicado.

4. **Para los resultados obtenidos en ST5, ejecutar ST6**: Por lo tanto, para el o los valores de CD obtenidos en la etapa ST5, se pide una evaluación del valor de CD obtenido, teniendo en cuenta el Modelo de Contexto.

Estructura del Informe

El documento debe contener:

- Carátula con número de grupo, nombre, apellido y CI de cada uno de los integrantes
- Introducción
- Descripción de las actividades desarrolladas en cada una de las Etapas ejecutadas, y los resultados obtenidos al final de cada Fase. En particular, en la etapa ST4, el Modelo de CD se debe especificar con el formato presentado en clase de teórico.
- Conclusiones sobre la metodología aplicada, los resultados obtenidos, y la experiencia personal en la realización de este trabajo.

REFERENCIAS

[1] https://www.goodreads.com/list/show/2681.Time_Magazine_s_All_Time_100_Novels

ANEXO

Resultados intermedios de la Fase 1

Problemas de Calidad de Datos

Libros duplicados

Libros valorados que no forman parte de los datasets

Libros sin autores

Libros sin ISBN

Libros sin editores

Ratings con escalas diferentes

Campos con valores nulos

Fechas con distintos formatos

Componentes del contexto

Dominio de Aplicación (AD): Librería

Usuarios:

U1, Administrador

U2, Publicista digital

U3, Analista de datos

Tareas:

U1: Administración y gestión de la librería

U2: Recomendación de libros y promoción de la librería

U3: Análisis de datos

Filtrado de datos:

DF1: Libros cuya publicación sea del año actual

DF2: top 3 de los libros con mayor score

DF3: libros editados por Wiley

Requerimientos del sistema:

SR: los tiempos de respuesta del sitio Web no deben superar los 3 segundos