Lección 3.3 Clasificación en base a la estructura :

Como ya comentábamos en la lección anterior, una de las formas en que se materializa la heterogeneidad de las fuentes es en el…

- **Nivel de estructura de los datos**: La forma en que se organizan los datos para facilitar su procesamiento usando un computador
    - Mayor nivel de estructura → Procesamiento más sencillo
    - Menor nivel de estructura → Procesamiento más complejo
- Por ejemplo, no es lo mismo el procesamiento y análisis de un texto escrito en lenguaje humano que uno al que ya se han aplicado un conjunto de reglas definidas para la estructuración de la información que contiene
    - ej. modelo relacional, lenguaje de marcas XML,...
- Cuando menor es el nivel de estructura más difícil es su procesamiento (integración y verificación de la calidad) y, por tanto, más complicada es la extracción de conocimiento.
- En base a su nivel de estructura las fuentes de datos pueden clasificarse en:
    - Estructuradas
    - Semi estructuradas
    - No estructuradas

1. Fuentes estructuradas

- Nivel más alto de estructura → Procesamiento eficiente y eficaz
- Los datos se almacenan con una estructura bien definida y que aplica unas normas muy estrictas
- El ejemplo más claro son las **bases de datos relacionales**
    - La información se almacena en tablas y se definen relaciones entre dichas tablas
    - Las tablas se componen de filas (tuplas) y columnas (campos o atributos)
    - Toda la información se almacena de acuerdo al esquema relacional definido
- Por lo general, los almacenes de datos (Data Warehouses) usan esta tecnología como almacenamiento subyacente
    - Aunque se aplique el modelo multidimensional (que vimos en el tema 2) en lugar del modelo relacional
- Suelen contener metadatos: información sobre los propios datos que ayuda en su interpretación (ej. descripciones, unidades de medida usadas....)

2. Fuentes semi estructuradas

El procesamiento de información estructurada es el más sencillo y ,desde hace tiempo, se consigue realizar de forma eficiente para grandes volúmenes de datos. Sin embargo, en Big Data, la mayoría de las fuentes externas de las que disponemos son de las consideradas semi o nada estructuradas.

- Los datos se almacenan conforme a conjunto de reglas menos estrictas y más flexibles
    - El nivel de estructura puede variar según su aplicación y, por tanto, también la dificultad de procesamiento
    - A medio camino entre datos estructurados y nada estructurados.
- Algunos de los formatos semi estructurados más usados:
    - XML, JSON, CSV, Excel…
- En algunos los datos se organizan conforme a un esquema o modelo de datos bien definido
    - XML (DTD y XML Schema)
- Suelen contener metadatos
- Algunas de las fuentes que usualmente se distribuyen en estos formatos:
    - Open Data, redes sociales, datos de sensores, logs de servidores web...

(ejemplo de JSON)

3. Fuentes no estructuradas

Por último, tenemos los datos no estructurados

- Menor nivel de estructura: No tienen una estructura definida de forma explícita
    - Ejemplos: texto (en lenguaje natural), vídeo, audio, imágenes...
    - Sí pueden tener algún tipo de estructura implícita: ej. División en párrafos de un texto, escenas de una película, estribillo de una canción…
- Para un computador puede llegar a ser muy difícil de interpretar
    - Es frecuente que, lo que es más fácil de interpretar para un humano sea lo más difícil de interpretar para un computador

- El **80% de las fuentes** disponibles en Big Data son **no estructuradas**.
- Existen algunas técnicas que permiten aprovechar este tipo de fuentes…
    - Procesamiento del lenguaje natural (PLN) → Estructuración
    - Minería de datos → Descubrimiento automático de conocimiento implícito en los datos
- En una aplicación Big Data es frecuente trabajar con fuentes de datos de los 3 tipos
- Como veremos en el módulo 4, la tecnología de almacenamiento y procesamiento de Big Data centra sus esfuerzos en dar soporte efectivo y eficiente a las nuevas fuentes semi y no estructuradas, y facilitar la integración de estas con otras fuentes altamente estructuradas.
- Puede ser necesario añadir estructura a las fuentes que no la tienen antes de aplicar algún proceso de análisis para el descubrimiento de conocimiento
    - Mediante procesos ETL (los veremos en la lección 5)

Enlaces de interés:

- http://www.computerweekly.com/blogs/cwdn/2010/10/ibm-80-percent-of-data-is-unstructured-so-what-do-we-do.html