Práctico 6
Ecosistemas de Bases de Datos

Filminas

- Bases de datos orientadas a objetos -- Bases de Datos 2022
- Datos no estructurados -- Bases de Datos 2022
- Recuperación de Información -- Bases de Datos 2022
- Minería de datos -- Bases de Datos 2022

Lecturas

Apartado 2.7 Silberschatz
Capítulo 8 Silberschatz
Apartados 9.1, 9.2 y 9.3 Silberschatz
https://es.wikipedia.org/wiki/L%C3%B3gica_de_descripci%C3%B3n
https://en.wikipedia.org/wiki/Unstructured_data
https://es.wikipedia.org/wiki/B%C3%BAsqueda_y_recuperaci%C3%B3n_de_informaci%C3%B3n
https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico
Blockchain https://es.wikipedia.org/wiki/Cadena_de_bloques
Data Warehouse https://es.wikipedia.org/wiki/Almac%C3%A9n_de_datos
Data Lake https://en.wikipedia.org/wiki/Data_lake
Data Cube https://en.wikipedia.org/wiki/Data_cube
OLAP Cube https://en.wikipedia.org/wiki/OLAP_cube
Seguridad de la Información https://es.wikipedia.org/wiki/Seguridad_de_la_informaci%C3%B3n

Ejercicios

Bases de Datos orientadas a objetos

1. Explique por qué la persistencia es un aspecto importante a desarrollar para el uso de lenguajes de programación orientados a objetos en bases de datos, en particular, qué limitaciones presentan estos lenguajes con respecto a la persistencia.

2. Mencione por lo menos dos conceptos que incorpore un lenguaje de descripción de bases de datos para representar objetos, por contraste con un lenguaje de descripción de bases de datos no orientado a objetos.

3. La primera tabla tiene valores que no son propios de 1FN, la segunda tabla está normalizada. Explique cómo la forma de expresión de la primera tabla contribuye a un modelado más intuitivo, usando el concepto de tipo complejo.

título	*lista-autores*	*editorial (nombre, sucursal)*	*lista-palabras-clave*
Compiladores	{Gómez, Santos}	(McGraw-Hill, Nueva York)	{traducción, análisis}
Redes	{Santos, Escudero}	(Oxford, Londres)	{Internet, Web}

título	*autor*	*nombre-editorial*	*sucursal-editorial*	*palabra-clave*
Compiladores	Gómez	McGraw-Hill	Nueva York	traducción
Compiladores	Santos	McGraw-Hill	Nueva York	traducción
Compiladores	Gómez	McGraw-Hill	Nueva York	análisis
Compiladores	Santos	McGraw-Hill	Nueva York	análisis
Redes	Santos	Oxford	Londres	Internet
Redes	Escudero	Oxford	Londres	Internet
Redes	Santos	Oxford	Londres	Web
Redes	Escudero	Oxford	Londres	Web

4. El siguiente fragmento de código SQL:1999 tiene algunas propiedades de orientación a objetos, mencione por lo menos dos.

```
create type Editorial as
      (nombre varchar(20),
      sucursal varchar(20))
create type Libro as
      (título varchar(20),
      array-autores varchar(20) array [10],
      fecha-pub date,
      editorial Editorial,
      lista-palabras-clave setof(varchar(20)))
create table libros of type Libro
```

5. ¿Qué propiedad de la orientación a objetos se está expresando en el siguiente fragmento de código?

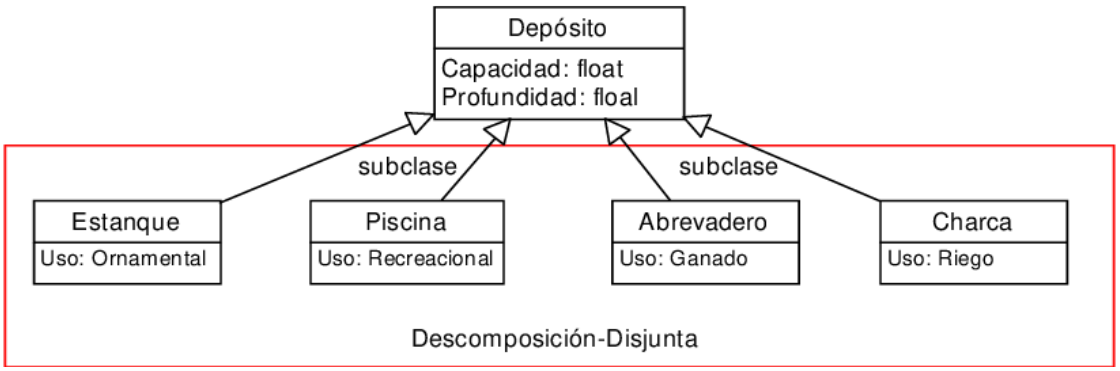
```
create table estudiantes of Estudiante
      under persona
create table profesores of Profesor
      under persona
```

6. ¿Qué propiedad de la orientación a objetos se está expresando en el siguiente fragmento de código?

```
create table ayudantes
of Ayudante
      under estudiantes, profesores
```

7. Las bases de datos orientadas a objetos pueden ser menos eficientes que las relacionales. Para mejorar la eficiencia de una base de datos orientada a objetos, en

algunos casos se opta por traducir algunas partes a una base de datos relacional, para hacer las consultas más eficientes. Proponga una traducción a un modelo relacional de la siguiente porción de una ontología geográfica, y haga la correspondiente traducción a tablas.



8.

Datos no estructurados

1. Dé un ejemplo de cómo una colección de datos no estructurados (por ejemplo, documentos, imágenes, audios) se puede albergar en una base de datos relacional. Explique qué información no queda representada de forma satisfactoria en ese almacenamiento porque el lenguaje de consulta no puede satisfacer alguna necesidad de información habitual sobre esos datos no estructurados.

2. Explique qué herramientas aporta una base de datos basada en documentos a diferencia de una base de datos relacional, para albergar datos con estructuras variables.

3. Identifique por lo menos tres tipos de datos no estructurados diferentes que se producen en su entorno.
4.

Recuperación de información

1. Explique cómo se pueden representar documentos para recuperación de información.

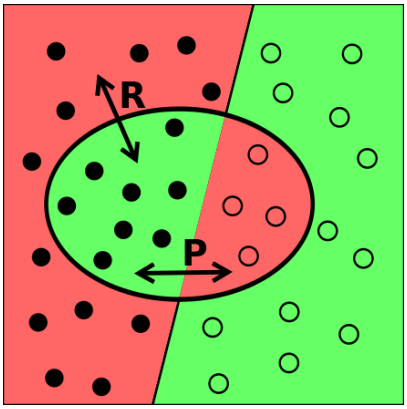
2. Compare una representación de un documento mediante las palabras más frecuentes con una representación de un documento mediante todas las secuencias de dos palabras que se encuentran en el documento, en términos de costes y adecuación descriptiva.

3. Explique la diferencia entre el modelo de recuperación de información booleano y el modelo de recuperación de información vectorial.

4. En el siguiente texto se describe el modelo booleano extendido. Describa el modelo booleano, el vectorial, explique las limitaciones de cada uno y explique cómo el modelo booleano extendido trata de superarlas.

   The goal of the Extended Boolean model is to overcome the drawbacks of the Boolean model that has been used in information retrieval. The Boolean model doesn't consider term weights in queries, and the result set of a Boolean query is often either too small or too big. The idea of the extended model is to make use of partial matching and term weights as in the vector space model. It combines the characteristics of the Vector Space Model with the properties of Boolean algebra and ranks the similarity between queries and documents. This way a document may be somewhat relevant if it matches some of the queried terms and will be returned as a result, whereas in the Standard Boolean model it wasn't.

5. El siguiente diagrama grafica las medidas de precisión (P) y exhaustividad (R). Describa verbalmente o con una fórmula estas métricas, y explique cómo nos ayudan a entender de forma cualitativa el rendimiento de un sistema de recuperación de información.



6.

Minería de datos

1. Dada la siguiente distinción entre aprendizaje supervisado y aprendizaje no supervisado:

   "*El aprendizaje supervisado se caracteriza por contar con información que especifica qué conjuntos de datos son satisfactorios para el objetivo del aprendizaje. Un ejemplo podría ser un software que reconoce si una imagen dada es o no la imagen de un rostro: para el aprendizaje del programa tendríamos que proporcionarle diferentes imágenes, especificando en el proceso si se trata o no de rostros.*

   En el aprendizaje no supervisado, en cambio, el programa no cuenta con datos que definan qué información es satisfactoria o no. El objetivo principal de estos programas suele ser encontrar patrones que permitan separar y

clasificar los datos en diferentes grupos, en función de sus atributos. Siguiendo el ejemplo anterior un software de aprendizaje no supervisado no sería capaz de decirnos si una imagen dada es un rostro o no pero sí podría, por ejemplo, clasificar las imágenes entre aquellas que contienen rostros humanos, de animales, o las que no contienen. La información obtenida por un algoritmo de aprendizaje no supervisado debe ser posteriormente interpretada por una persona para darle utilidad."

Si queremos identificar estudiantes con alta probabilidad de abandonar el cursado de una materia, ¿ante qué tipo de problema nos encontramos, un problema de aprendizaje supervisado o no supervisado? Justifique su respuesta apelando a los métodos presentados en el texto, e imagine por lo menos tres atributos con los que se caracterizarían los estudiantes en este contexto.

2. Las reglas de asociación

Según la definición original de *Agrawal et al*[3] el problema de minería de reglas de asociación se define como:

- Sea $I = \{i_1, i_2, \ldots, i_n\}$ un conjunto de n atributos binarios llamados **items**.
- Sea $D = \{t_1, t_2, \ldots, t_m\}$ un conjunto de transacciones almacenadas en una **base de datos**.

Cada transacción en D tiene un **ID** (identificador) único y contiene un subconjunto de items de I. Una **regla** se define como una implicación de la forma:

$$X \Rightarrow Y$$

Donde:

$$X, Y \subseteq I \text{ y}$$
$$X \cap Y = \emptyset$$

Los conjuntos de items X y Y se denominan respectivamente "**antecedente**" (o parte izquierda) y "**consecuente**" (o parte derecha) de la regla.

Para seleccionar reglas interesantes del conjunto de todas las reglas posibles que se pueden derivar de un conjunto de datos se pueden utilizar restricciones sobre diversas medidas de "significancia" e "interés". Las restricciones más conocidas son los umbrales mínimos de "soporte" y "confianza".

El 'soporte' de un conjunto de items X en una base de datos D se define como la proporción de transacciones en la base de datos que contiene dicho conjunto de items:

$$\text{sop}(X) = \frac{|X|}{|D|}$$

La 'confianza' de una regla se define como:

$$\text{conf}(X \Rightarrow Y) = \frac{\text{sop}(X \cup Y)}{\text{sop}(X)} = \frac{|X \cup Y|}{|X|}$$

Dada la siguiente base de datos:

ID	Leche	Pan	Mantequilla	Cerveza
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Calcule la confianza de:

$$\{\text{Leche}, \text{Pan}\} \Rightarrow \{\text{Mantequilla}\}$$

3. Explique cómo podría transformar la historia impositiva de los ciudadanos (el registro de los impuestos que han pagado) en un formato apto para aplicar reglas de asociación.

Data Warehousing

1. En el contexto de un ecosistema de información en una institución en la que se ha desplegado un almacén de datos (*data warehouse*), describa la funcionalidad de una componente ETL, glosando el siguiente texto:

> *"Extract, transform, load (ETL) is a three-phase process where data is extracted, transformed (cleaned, sanitized, scrubbed) and loaded into an output data container. The data can be collated from one or more sources and it can also be outputted to one or more destinations. ETL processing is typically executed using software applications but it can also be done manually by system operators. ETL software typically automates the entire process and can be run manually or on reoccurring schedules either as single jobs or aggregated into a batch of jobs.*
>
> *A properly designed ETL system extracts data from source systems and enforces data type and data validity standards and ensures it conforms structurally to the requirements of the output."*

2. Describa dos diferencias entre una base de datos relacional tradicional y una base de datos basada en blockchain, haciendo referencia al siguiente texto. Explique en

qué contextos de uso resultan especialmente valiosas las características distintivas de blockchain.

> "*A blockchain is a type of distributed ledger technology (DLT) that consists of growing list of records, called blocks, that are securely linked together using cryptography. Each block contains a cryptographic hash of the previous block, a timestamp, and transaction data. The timestamp proves that the transaction data existed when the block was created. Since each block contains information about the previous block, they effectively form a chain (compare linked list data structure), with each additional block linking to the ones before it. Consequently, blockchain transactions are irreversible in that, once they are recorded, the data in any given block cannot be altered retroactively without altering all subsequent blocks.*"

3. En este texto se describen las diferencias entre una base de datos tradicional y un almacén de datos:
   "*Database*
   - *Used for Online Transactional Processing (OLTP) but can be used for other purposes such as Data Warehousing. This records the data from the user for history.*
   - *The tables and joins are complex since they are normalized (for RDMS). This is done to reduce redundant data and to save storage space.*
   - *Entity – Relational modeling techniques are used for RDMS database design.*
   - *Optimized for write operation.*
   - *Performance is low for analysis queries.*

   Data Warehouse
   - *Used for Online Analytical Processing (OLAP). This reads the historical data for the Users for business decisions.*
   - *The Tables and joins are simple since they are de-normalized. This is done to reduce the response time for analytical queries.*
   - *Data – Modeling techniques are used for the Data Warehouse design.*
   - *Optimized for read operations.*
   - *High performance for analytical queries.*
   - *Is usually a Database.*
   - *It's important to note as well that Data Warehouses could be sourced from zero to many databases.*"

   Describa un escenario en el que un almacén de datos sea más adecuado que una base de datos tradicional, y otro escenario donde sea a la inversa, justificando las razones.

4. Explique la funcionalidad de un data mart:
   "*A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area), hence they draw data from a limited number of sources such as sales, finance or marketing. Data marts are often built and controlled by a single*

department within an organization. The sources could be internal operational systems, a central data warehouse, or external data."

5. Explique dos ventajas y dos desventajas de una base de datos distribuida.