NUEVAS PROPUESTAS PARA BÚSQUEDAS POR SIMILITUD EN BASES DE DATOS MÉTRICAS

Dra. Nora Reyes
Dra. Karina Figueroa
Verónica del Rosario Ludueña
Patricia Roggero

UNIVERSIDAD MICHOACANA
DE SAN NICOLÁS DE HIDALGO
Cuna de héroes, crisol de pensadores

Contenido del curso (1/2)

- Conceptos Fundamentales de Espacios Métricos
  - Introducción y motivación
  - Definición de espacios métricos.
  - Funciones de Distancia: propiedades.
  - Tipos de búsquedas pos similitud más comunes.
  - Maldición de la dimensión.
- Índices para Bases de Datos Métricas
  - Taxonomía de los índices
  - Principales referentes de índices basados en particiones compactas.
  - Principales referentes de índices basados en pivotes.
  - Principales referentes de índices basados en permutaciones
  - Índices estáticos y dinámicos. Ejemplos.
  - Índices para memoria secundaria. Ejemplos.
- Algoritmos Exactos y Aproximados
  - Algoritmos Exactos.
  - Algoritmos Aproximados.
  - Medidas de evaluación de calidad de respuesta.

Contenido del curso (2/2)

- Otras operaciones de Interés sobre Bases de Datos Métricas:
  - Join por Similitud, variantes.
  - Algoritmos para Join: con índices y sin índices.
  - Ejemplos de soluciones existentes.
  - Medidas de evaluación de la dimensionalidad.

El poder de los datos

- Datos vs información
- Los datos son percibidos a través de los sentidos y, una vez que se integran, terminan por generar la información que se necesita para producir el conocimiento.
- La información es un recurso que **otorga significado o sentido** a la realidad.
- La **información** está constituida por un **grupo de datos ya supervisados y ordenados**.
  - La información permite resolver problemas y tomar decisiones.

- Para llegar a la información hay que resolver muchos problemas primero

Datos en el pasado

- 1908, en la isla de Creta descubrieron un disco de arcilla que data del año 2000 AC
- Un misterio! Pero…. así funcionaba la sociedad
- Cómo buscaban!!! ????

En nuestros días

- Podemos copiar
- Reutilizar la información
- Guardarla
- Distribuirla

La nueva era de los datos

- Los datos pasaron de ser un ente estático y almacenado a ser algo que es un fluido y además dinámico
- Hace 4000 años, los datos eran pesados, no almacena mucha información y sobre todo, no cambiaban!
- Hoy en día, en una memoria del tamaño de una uña caben miles de documentos que pueden compartirse a velocidad de la luz

Datos digitales

- Todo lo que se puede ver, escuchar, leer escribir y medir, ahora puede estar en forma digital
- En los siguientes años, la cantidad de información producida será gigantesca
- Estimaciones
  - 95% de los datos producidos son digitales
  - Textos digitales son importantes
  - Los datos multimedia, científicos, obtenidos de sensores, etc, ahora son muy frencuentes

Cuánto es mucho?

- 1 TB (250,000 canciones)
- 20 TB (fotos subidas a facebook al mes)
- 120 TB (Datos recogidos por el telescopio espacial Hubble)
- 460 TB (Datos del tiempo climático EEUU compilados por el National Climatic Data Center)
- 530 TB (Videos de youtube)
- 600 TB (censos EEUU 1970-2000 y base de datos genealogía)
- **1 PB Datos procesados por google cada 75 minutos**

Definiciones

Unidades de información (del byte)			
Sistema Internacional (decimal)		ISO/IEC 80000-13 (binario)	
Múltiplo (símbolo)	SI	Múltiplo (símbolo)	ISO/IEC
kilobyte (kB)	10^3	kibibyte (KiB)	2^{10}
megabyte (MB)	10^6	mebibyte (MiB)	2^{20}
gigabyte (GB)	10^9	gibibyte (GiB)	2^{30}
terabyte (TB)	10^{12}	tebibyte (TiB)	2^{40}
petabyte (PB)	10^{15}	pebibyte (PiB)	2^{50}
exabyte (EB)	10^{18}	exbibyte (EiB)	2^{60}
zettabyte (ZB)	10^{21}	zebibyte (ZiB)	2^{70}
yottabyte (YB)	10^{24}	yobibyte (YiB)	2^{80}
Véase también: nibble • byte • sistema octal			

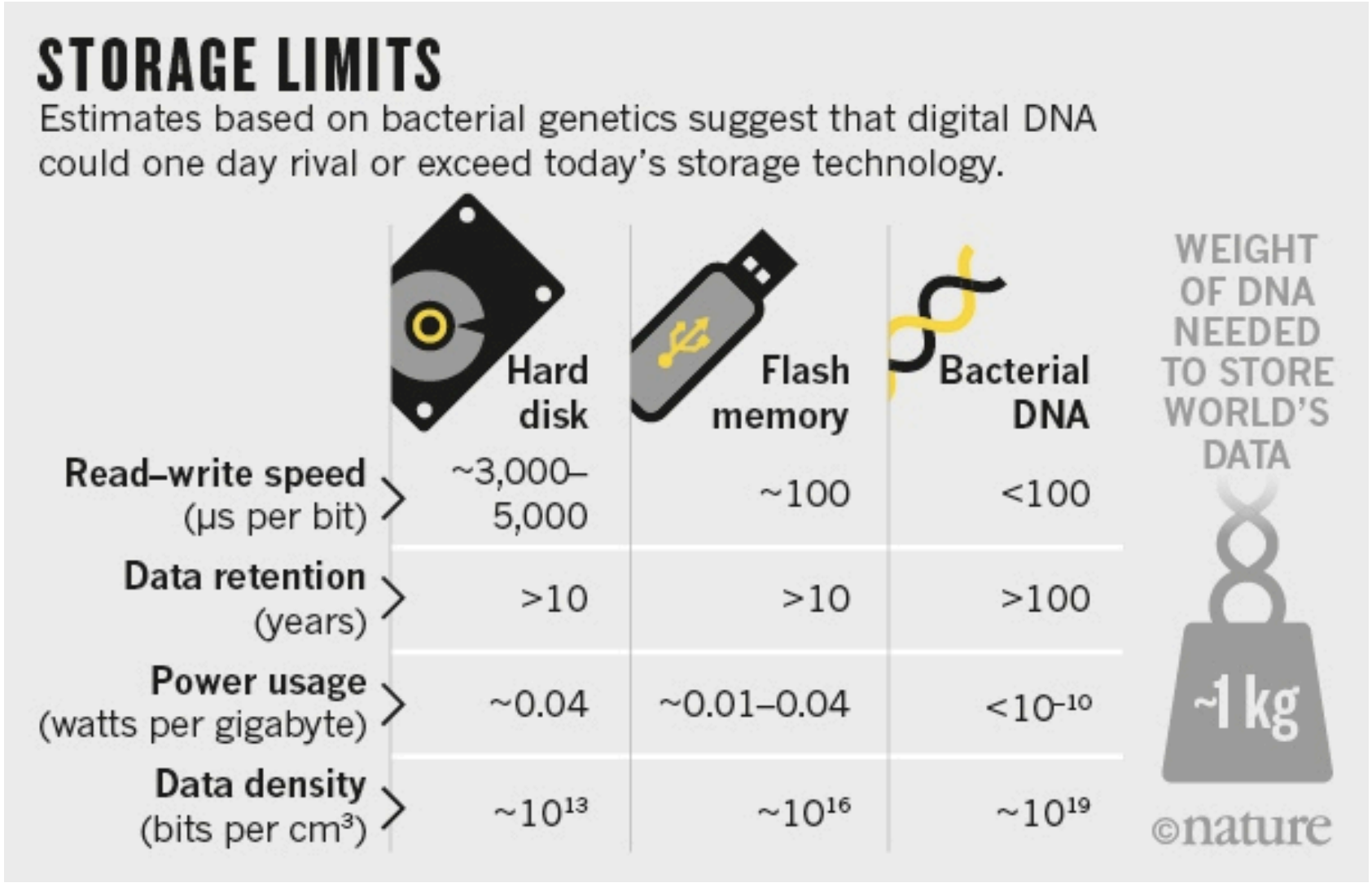
Dato curioso

1998, 1GB costaba $228

2003, $3.88

2007, $0.88

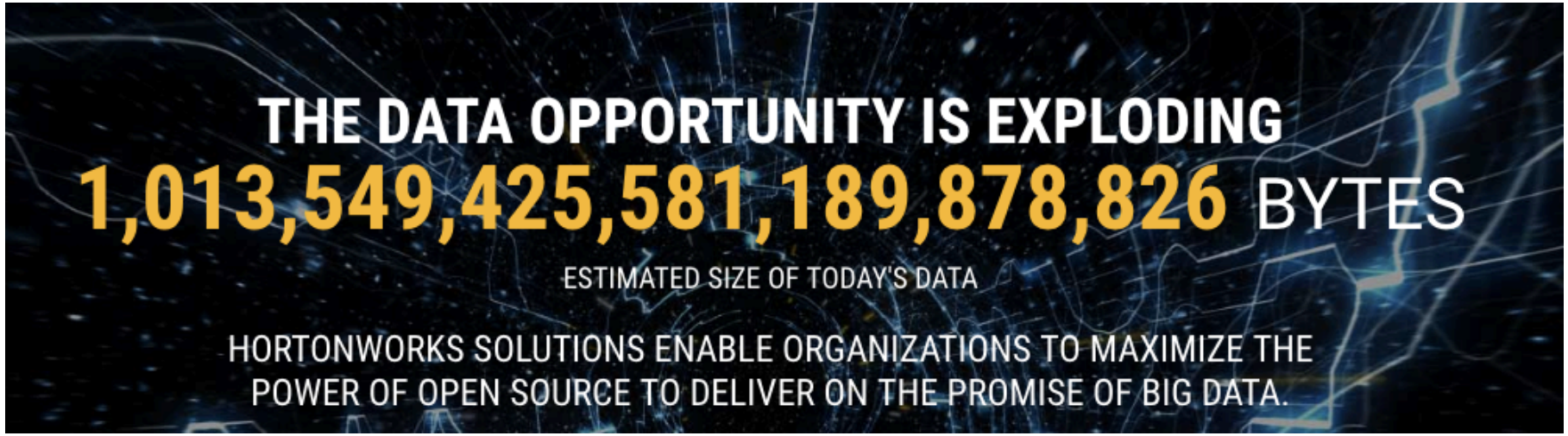
Almacenamiento de datos



Tamaño de los datos

- 2008 se habla de petabytes a zettabytes
- 2012 cerca de 2.5 trillones de bytes de datos (2.5×10^{18})
- IBM estima que cada día se producen 2.5 quitillones de datos
- 90% de los datos se ha hecho en los últimos 2 años!

Un dato curioso de Hortonworks



Octubre 2016

Motivación


IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON URBAN DICTIONARY
1,600+ READS ON Scribd
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW Linked in ACCOUNTS
1 NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WORDPRESS DOWNLOADS
125+ PLUGIN DOWNLOADS
695,000+ facebook STATUS UPDATES
1,700+ Firefox DOWNLOADS
79,364 WALL POSTS
510,040 COMMENTS
694,445 SEARCH QUERIES
168 MILLION EMAILS ARE SENT
60+ NEW BLOGS
1,500+ BLOG POSTS
70+ DOMAINS REGISTERED
600+ NEW VIDEOS
25+ HOURS TOTAL DURATION
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
13,000+ iPhone APPLICATIONS DOWNLOADED
20,000+ NEW POSTS ON tumblr.
LARGEST
GO-Globe.com

Búsquedas

¿Qué y cómo buscamos?

Desordenadas

Búsqueda exacta:

- Cromático
- piña
- Cuatequil
- río
- coche
- Ria
- comarcante
- concatedralidad
- conexión
- Pinar
- cosidura
- Debeladora
- contraclave
- Contrariadora
- cordojo
- coronizar
- confusión
- coxa
- Componedora
- cuellidegollado
- Collazo
- curiel
- daguerrotipo
- consiliario
- copropietario
- Pino
- Conversivas
- cochambrera
- Piso
- Concordato

ritmo

amigo

Ordenadas

Búsqueda exacta:

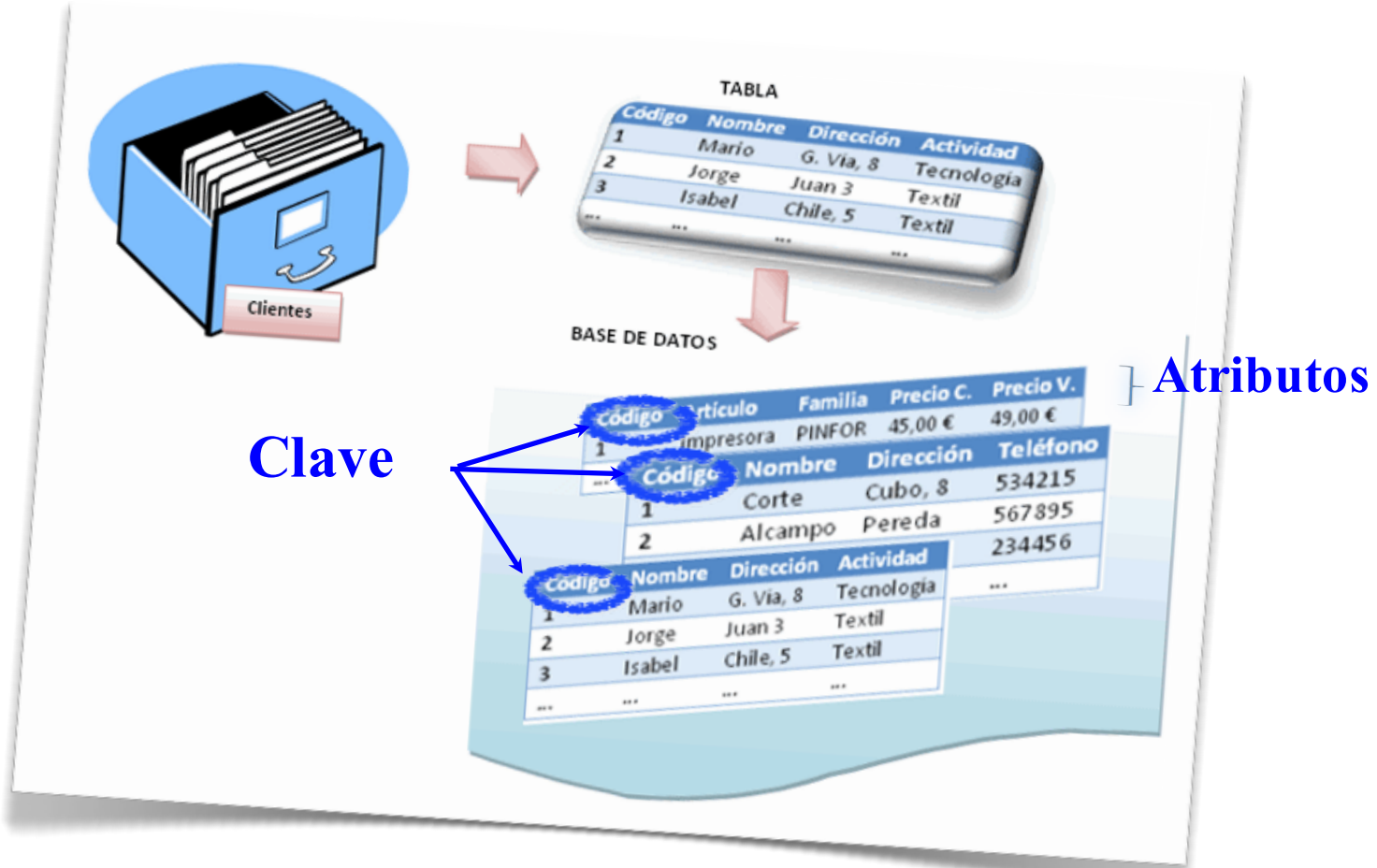
- Componedora
- Concordato
- Cuatequil
- cochambrera
- coche
- collazo
- comarcante
- concatedralidad
- conexión
- confusión
- consiliario
- contraclave
- contrariadora
- conversiva
- copropietario
- cordojo
- coronizar
- cosidura
- coxa
- cromático
- cuellidegollado
- curiel
- daguerrotipo
- debeladora
- pinar
- pino
- pinos
- piso
- piña
- ria
- río

Mate

Fácil !!!!!

Introducción

★Las Bases de Datos convencionales se basan en estructurar la información:



Información

Tablas

Registros

Campos

Introducción

- Sobre ellas las consultas más comunes pueden ser:
  - Búsquedas exactas
  - Búsquedas parciales
  - Búsquedas extremales
  - Búsquedas secuenciales
  - etc.

Todas suponen que los objetos se pueden comparar por igualdad y que existe un orden entre ellos.

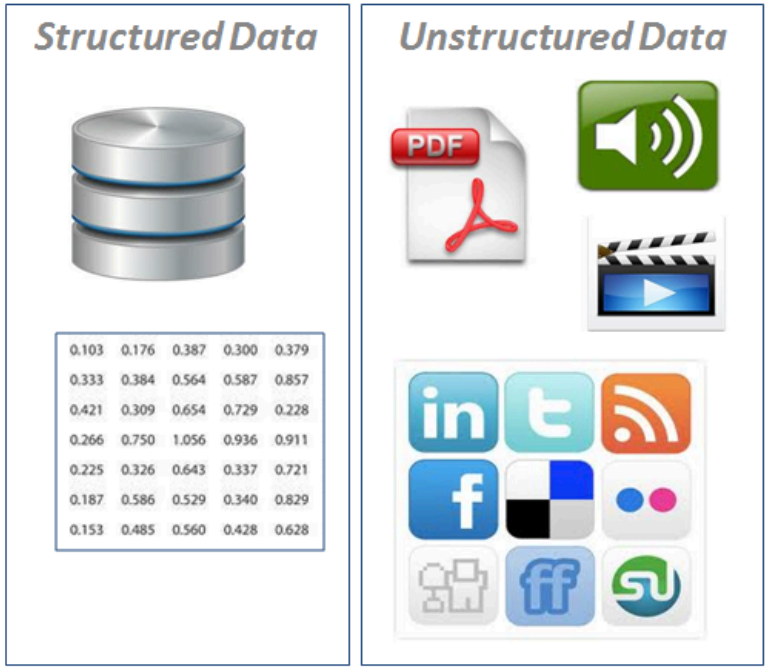
Introducción

- Además, sobre este tipo de base de datos existen otras operaciones de interés:
  - Selección
  - Join
  - Proyección
  - Producto Cartesiano
  - etc ...

Algunas de ellas también necesitan igualar exactamente los objetos, o en algunos casos la existencia de un orden sobre los objetos.

Tipos de datos

- Datos estructurados
  - Bases de datos relacionales
- Datos no estructurados
  - PDF, emails, texto en general
- Datos semi-estructurados
  - Datos con marcas, como XML

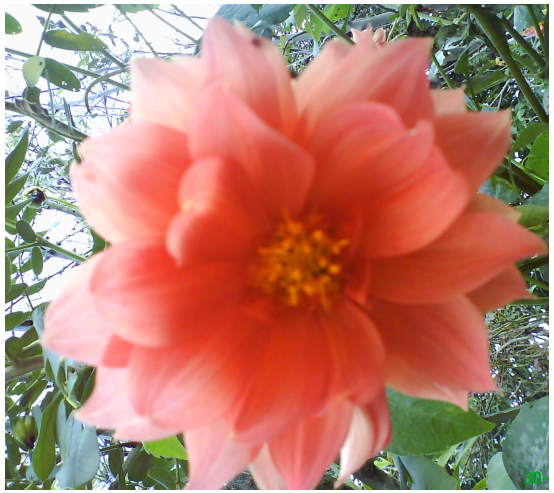


- Cómo buscamos?

Búsquedas por similaridad

★Actualmente se almacenan otros tipos de datos no estructurados, que no son necesariamente comparables por igualdad:

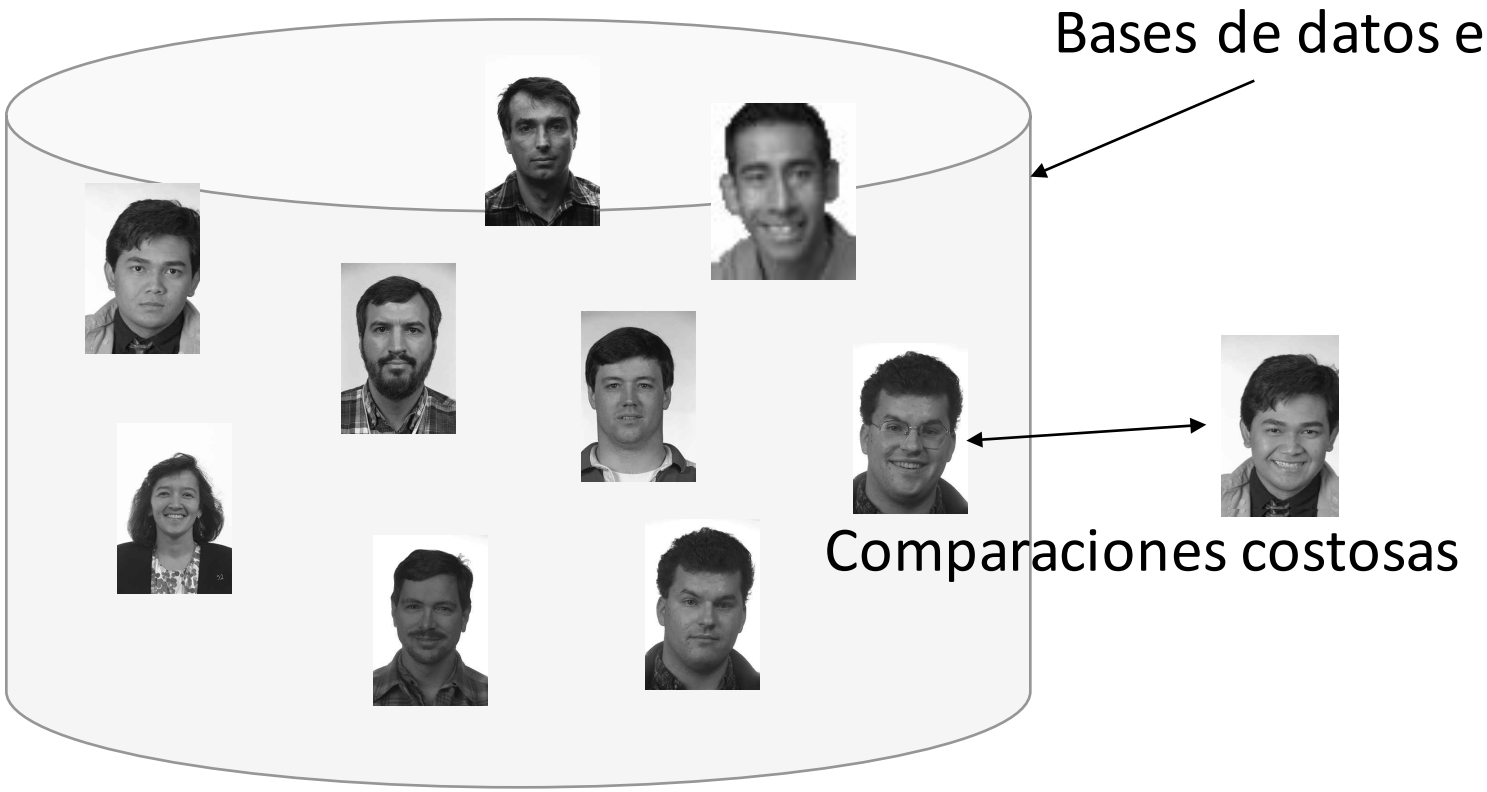
Búsqueda en imágenes



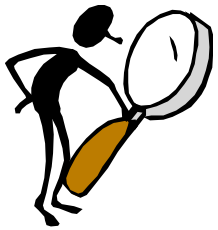
Búsqueda por similitud!!

Respuesta de google

Motivación



- No es posible la búsqueda exacta



El verdadero problema

Existirá el orden total para todos los tipos de datos?

Cómo establecer un orden?

Encontrar una aguja en el pajar

Aplicaciones

- Recuperación de información
- Problemas de Clasificación
- Finanzas. Qué economía tuvo un comportamiento similar?
- Biología. Imágenes de resonancia magnética, IoT.
- El corazón de la inteligencia artificial tiene que ver con resolver un problema de búsquedas

- Importante: la función de distancia
  - Permite que el usuario pueda definir un problema de percepción sobre lo que significa similiaridad en el dominio que se va a usar

Espacios métricos

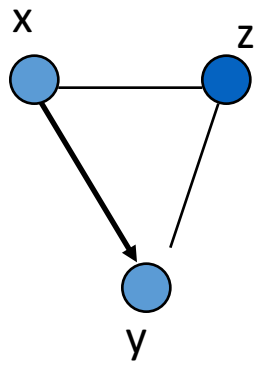
Son una alternativa de solución en muchas aplicaciones y en otras, quizá la única.

Búsqueda por distancia

Definición del problema

Sea $\mathbb{X}$ el universo de objetos y d una función de distancia. Dado un conjunto $\mathbb{U} \subseteq \mathbb{X}$ de n elementos, preprocesar o estructurar los datos de manera que se puedan resolver las búsquedas por similitud.

Espacios métricos



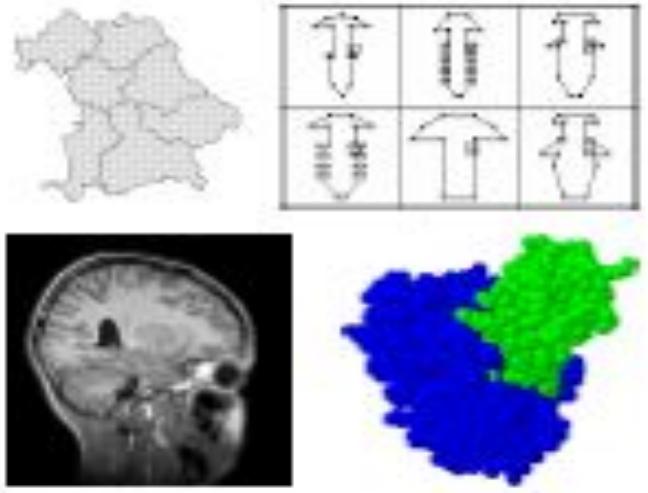
Espacio métrico consiste de un universo de objetos $\mathbb{X}$ y una función de distancia d

$$d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}^+$$

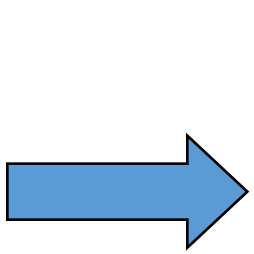
Con las siguiente propiedades:

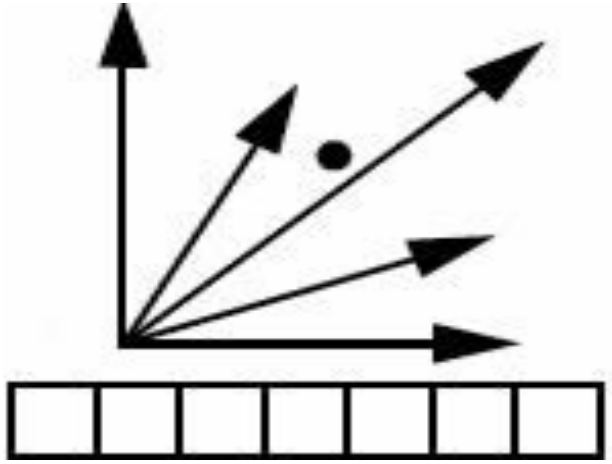
Positividad estricta:	$\forall x, y \in \mathbb{X},$	$d(x, y) \geq 0$
Simetría:	$\forall x, y \in \mathbb{X},$	$d(x, y) = d(y, x)$
Identidad:	$\forall x, y \in \mathbb{X},$	$x = y \Leftrightarrow d(x, y) = 0$
Desigualdad triangular:	$\forall x, y, z \in \mathbb{X},$	$d(x, z) \leq d(x, y) + d(y, z)$

Etapas del proceso



Objetos complejos





Espacios y distancias

Tipos de espacios

- Vectores
  - Caras
  - Imágenes
  - Etc, etc….
- Palabras
- Documentos

Medidas de distancia

- Tipos de distancias
  - Vectores
  - palabras
- Ejemplos de distancias
  - Distancia Minkowski
  - Distancia de edición

Medidas de distancia

- Discretos
  - Función que regresa sólo un pequeño conjunto de valores (predefinidos)

- Continuos
  - Funciones en las cuales la cardinalidad del conjunto de valores es muy grande o infinita

Distancia Minkowski

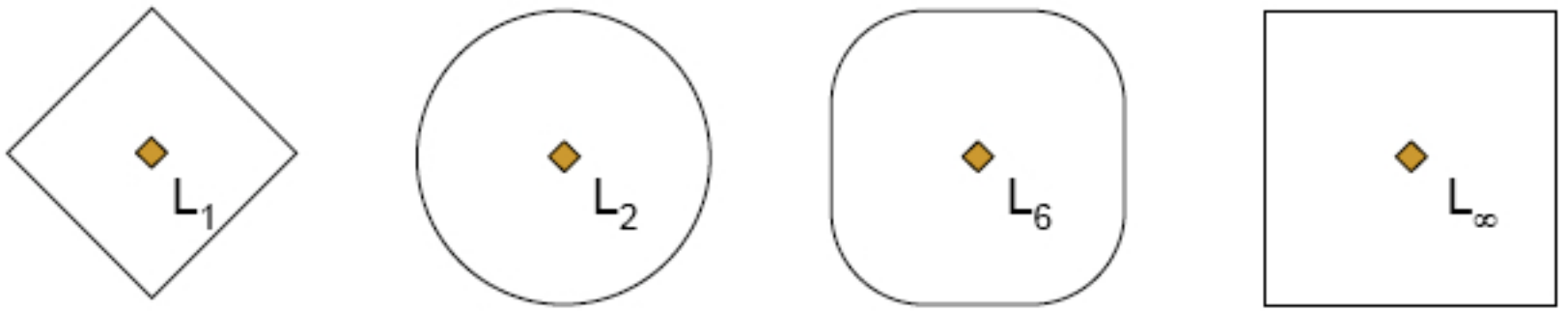
- También conocidos como métricas Lp
- Definida para vectores n-dimensionales

$$L_p[(x_1,\cdots,x_n),(y_1,\cdots,y_n)] = \sqrt[p]{\sum_{i=1}^{n} | x_i - y_i |^p}$$

Casos especiales

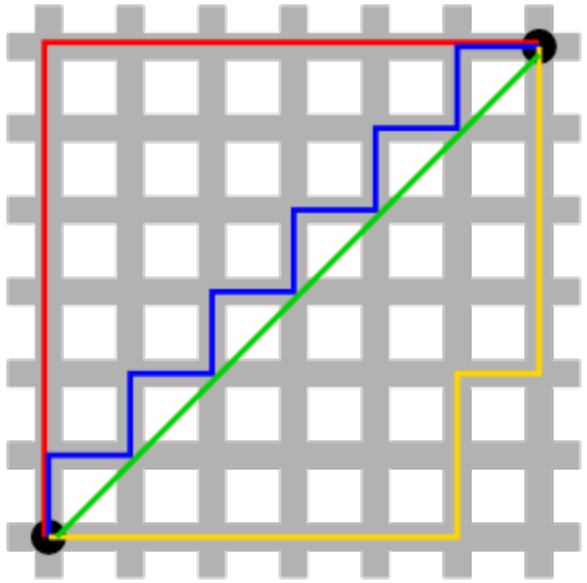
- L_1 Distancia Manhattan (city-block)
- L_2 Distancia euclidiana
- L_∞ Distancia máxima (infinito)

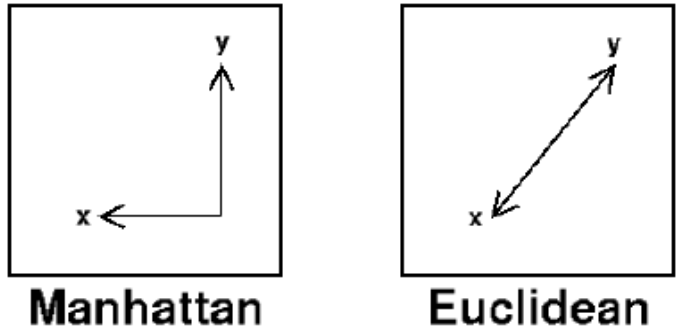
$$L_\infty = \max_{i=1}^{n} | x_i - y_i |$$



Medidas de distancia

Distancia de Minkowski





- Distancia de Manhattan = 12
- Distancia Euclídiana ≅ 8.5

Distancia de edición

- También llamada distancia Levenstein
  - Mide el menor número de operaciones para transformar una cadena *x* en una cadena *y*
- ***Insertar*** un caracter c en una cadena ***x*** en la posición ***i***

$$ins(x,i,c) = x_1 x_2 \cdots x_{i-1} c x_i \cdots x_n$$

- ***Borrar*** el caracter en la posición ***i*** en la cadena ***x***

$$del(x,i) = x_1 x_2 \cdots x_{i-1} x_{i+1} \cdots x_n$$

- **Remplazar** el caracter en la posición ***i*** en la cadena ***x*** con ***c***

$$replace(x,i,c) = x_1 x_2 \cdots x_{i-1} c x_{i+1} \cdots x_n$$

Medidas de distancia

Distancia de edición = Distancia de Levenshtein

Número de operaciones necesario
para transformar una cadena en otra.

d(“data mining”, “data minino”) = 1

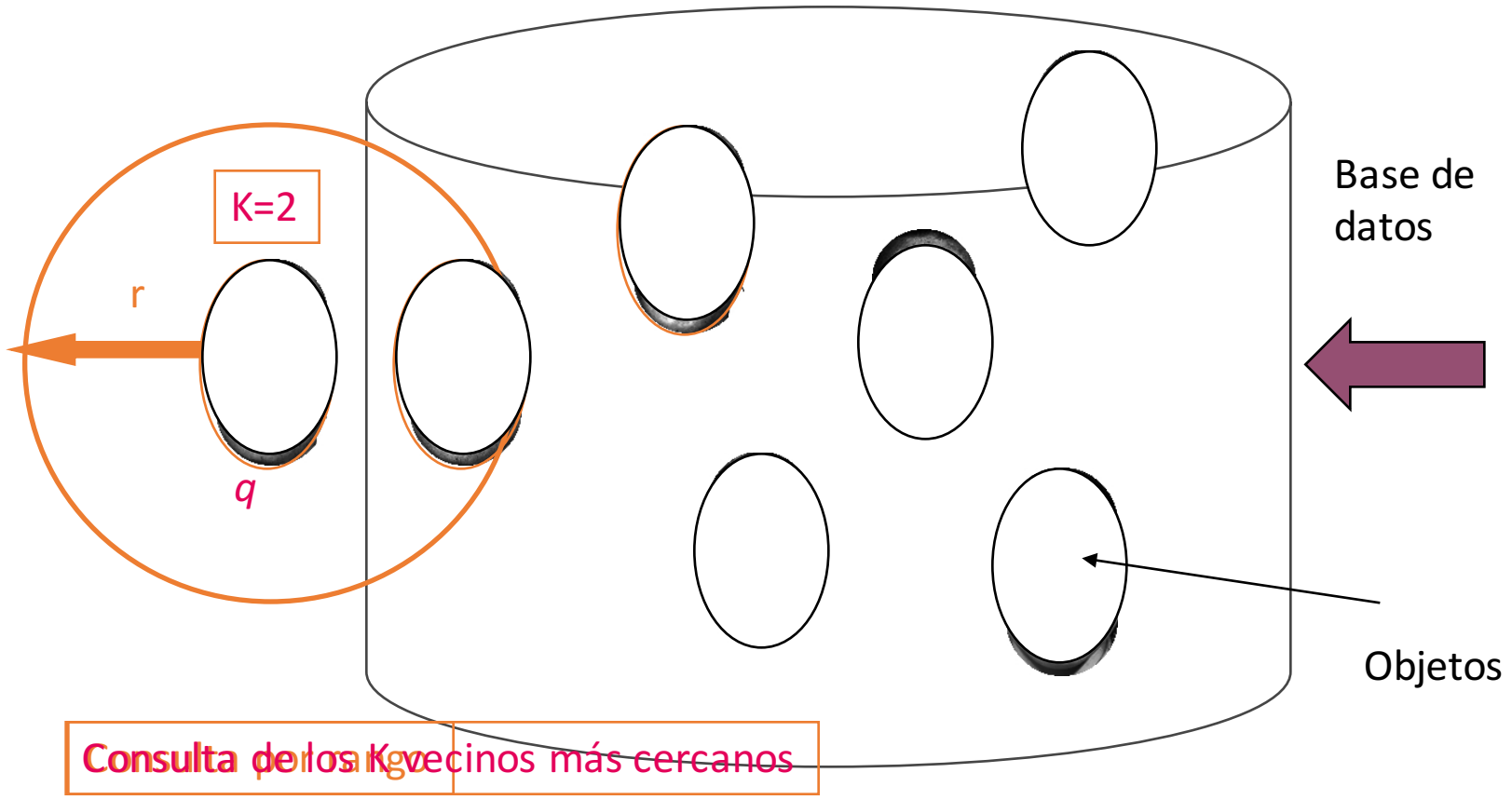
d(“efecto”, “defecto”) = 1

d(“poda”, “boda”) = 1

d(“night”,”natch”) = d(“natch”,”noche”) = 3

Aplicaciones: Correctores ortográficos, reconocimiento de voz, detección de plagios, análisis de ADN…

Tipos de consultas



Tipos de consulta

Asumimos que tenemos $\mathbb{U} \subseteq \mathbb{X}$, de tamaño n, los tipos básicos de consultas son:

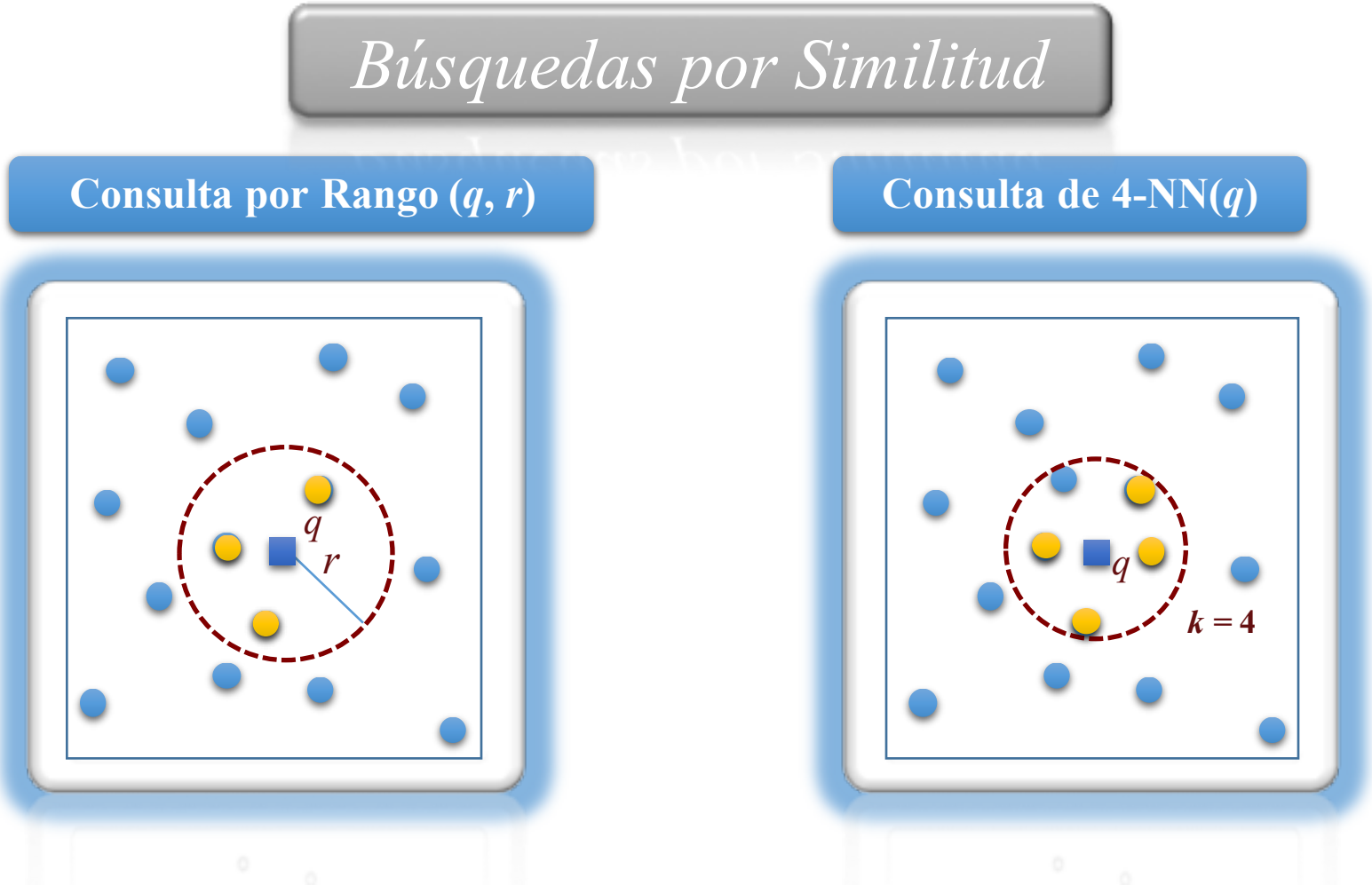
- Consultas de rango: recuperar todos los elementos a distancia r de q.

$$R(q, r) = d(u, q) \leq r \mid \forall u \in \mathbb{X}$$

- Consultas de vecinos mas cercanos: **k-Nearest Neighbor (kNN)** recupera los K elementos mas cercanos a q en $\mathbb{U}$.

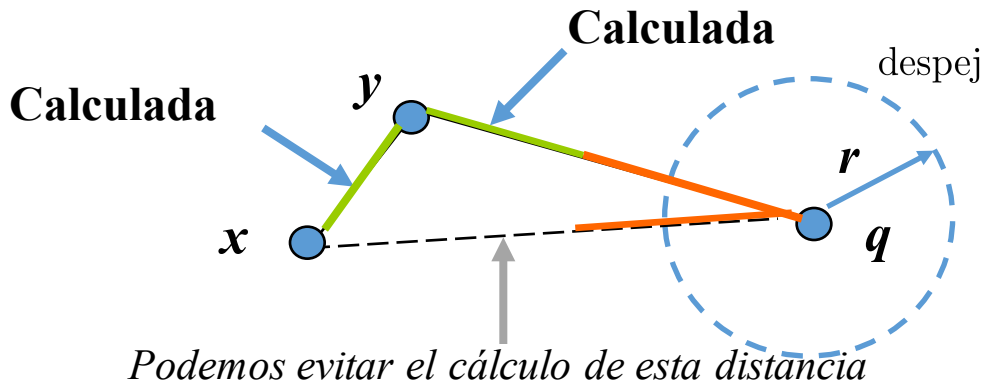
$$kNN(q) = \{R \subseteq \mathbb{U}, |R| = k \ \wedge \ \forall x \in R, y \in \mathbb{U} - R : d(q, x) \leq d(q, y)\}$$

Tipos de consultas



Detalle

- Uso de la desigualdad triangular para evitar cálculos de distancia en las búsquedas por rango:



$$d(x,y) \leq d(x,q) + d(q,y)$$

$$d(q,y) \leq d(q,x) + d(x,y)$$

despejando desde estas desigualdades y aplicando propiedades

$$d(x,y) - d(q,y) \leq d(x,q)$$

$$d(q,y) - d(x,y) \leq d(x.q)$$

combinándolas

$$|d(q,y) - d(x,y)| \leq d(x,q)$$

Complejidad y costos

- Tiempo de resolución de una búsqueda:

$$\boldsymbol{T} = \#\text{evaluaciones}(d) \times \text{complejidad}(d) + \text{Tpo CPU} + \text{Tpo. E/S}$$

- Resolver las búsquedas por similitud trivialmente $\Rightarrow$ $O(N)$ evaluaciones de distancia.

- ***Algoritmo de indexación***: construye una ***estructura de datos o índice*** diseñada para ahorrar cálculos de $\boldsymbol{d}$ en la búsqueda.

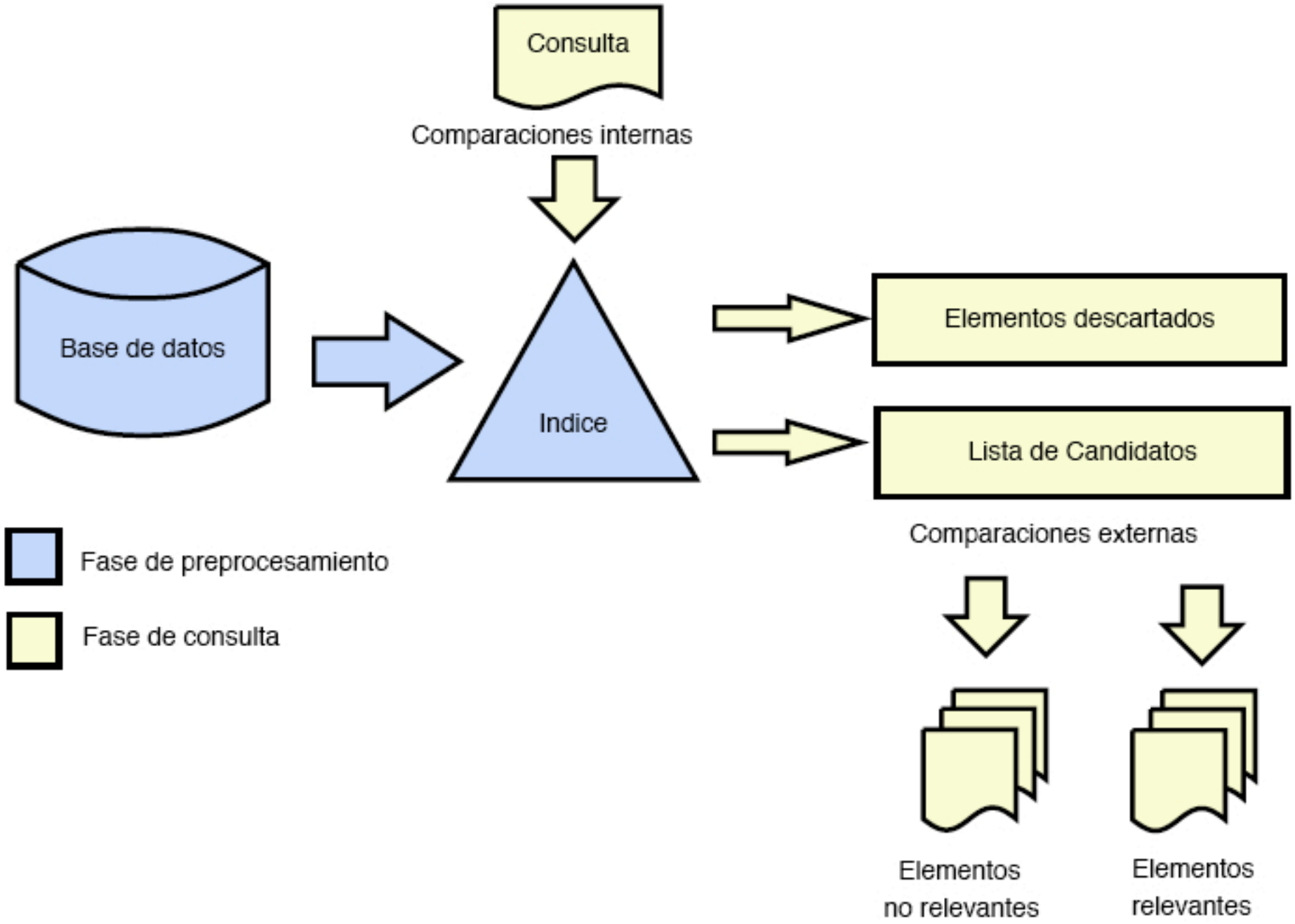
Conceptos

☐*Modelo Unificado*

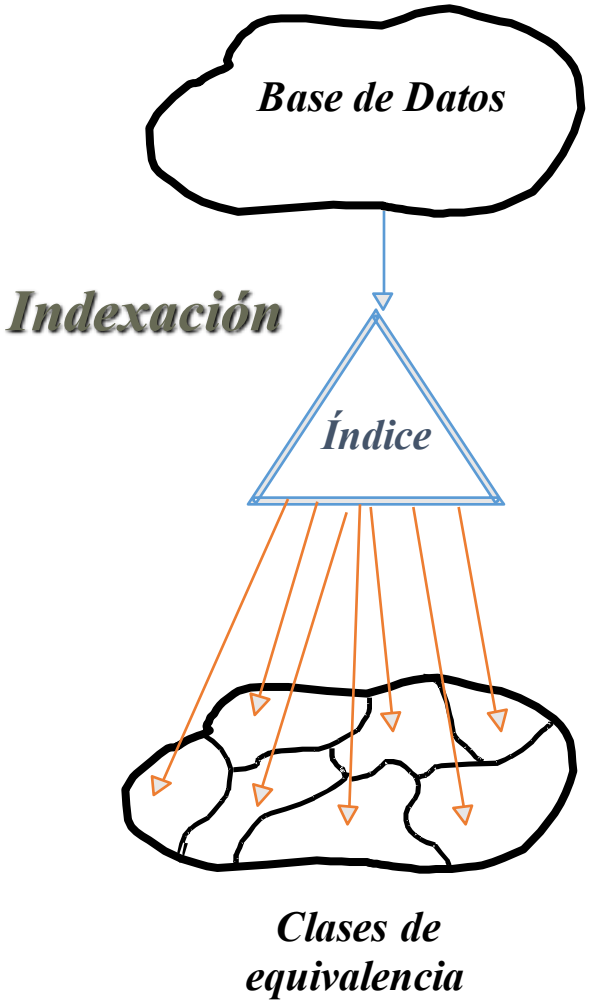
- Todos los algoritmos de indexación particionan el conjunto o base de datos U en subconjuntos.

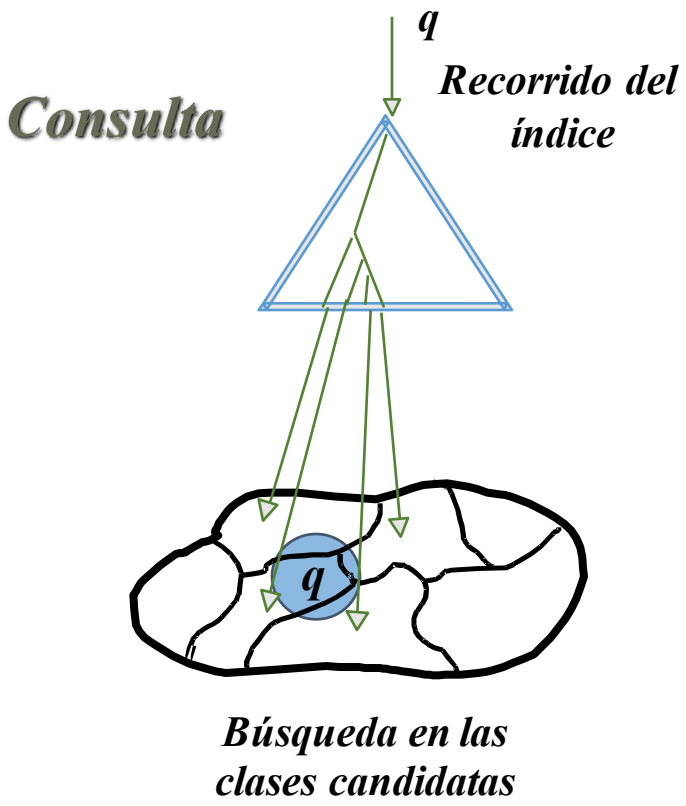
☐Pretendemos que el índice permita determinar un conjunto de subconjuntos candidatos que pueden contener elementos relevantes a la consulta.

☐Se busca en el índice para encontrar los subconjuntos relevantes (***complejidad interna***) y luego se verifican exhaustivamente (***complejidad externa***).


Consulta
Comparaciones internas
Base de datos
Indice
Elementos descartados
Lista de Candidatos
Comparaciones externas
Fase de preprocesamiento
Fase de consulta
Elementos
no relevantes
Elementos
relevantes

Modelo Unificado





¿Cuáles son los problemas?

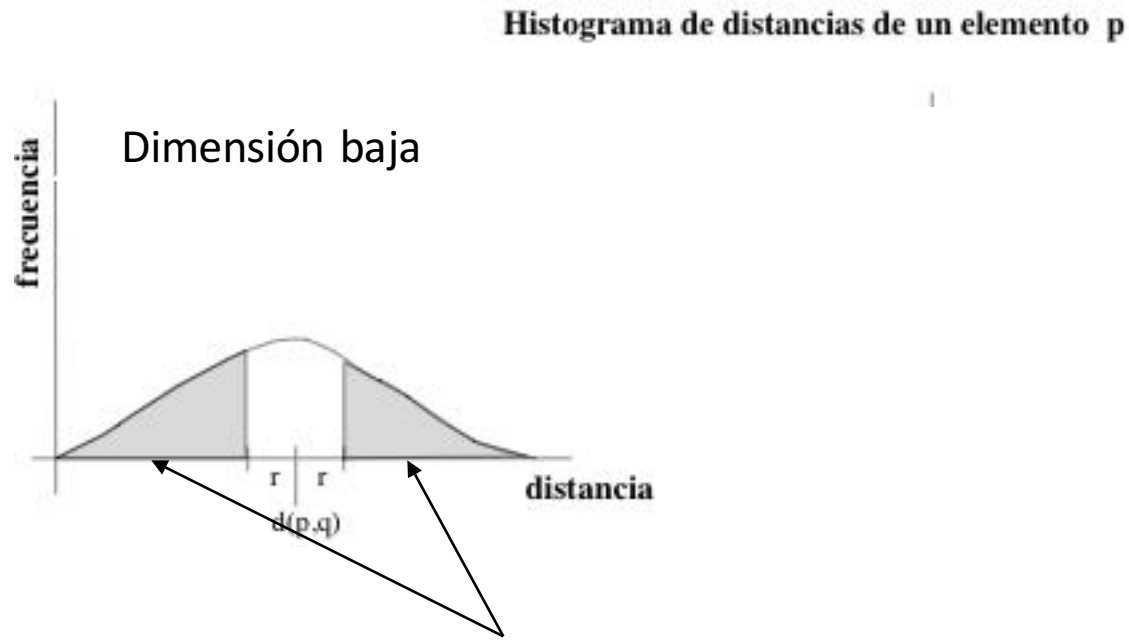
- El número de cálculos de distancia requeridos para resolver la consulta
- La calidad de una consulta aproximada (recuperación)
- Tiempo de CPU y memoria requerida para procesar la consulta
- Aplicabilidad, (a qué espacios se puede aplicar?)
- Modalidad de consulta: rango, kNN, vecinos reversos, join, etc
- Escalabilidad, qué tan independiente es el algoritmo del tamaño de la BD
- Propiedades dinámicas.
- Operaciones de entrada/salida de los discos
- Capacidad para correr en paralelo o en ambientes distribuidos

Maldición de la dimensión

Dimensión de los datos

- Usualmente los datos son representados como vectores característicos con alta dimensión
- El problema es que los índices se deterioran rápidamente cuando la dimensión de los datos aumenta
- Básicamente el problema es el acelerado deterioro de los datos al seguir la tendencia de estar equidistantes

Problema:
La dimensión de los datos



Elementos descartados en una consulta usando *p*

¿Preguntas?

Conocen algún problema donde puedan aplicarse los índices de búsqueda por similaridad

Referencias

- Además de los artículos mencionados
- http://www.nmis.isti.cnr.it/amato/similarity-search-book/SAC-07-tutorial.pdf
- Similarity search: The metric space approach. P Zezula, G Amato, V Dohnal, M Batko. Springer-Verlag New York Inc.
- Searching in metric spacesE Chávez, G Navarro, R Baeza-Yates, JL Marroquín. ACM computing surveys (CSUR) 33 (3), 273-321