CC52V
Bases de datos multimedia

Prof. Benjamin Bustos

Capítulo 2
Búsqueda por similitud en bases de datos multimedia

2.1 Conceptos básicos

- Concepto de similitud es inherentemente subjetivo
- Modelos para definir objetivamente "similitud"
  - *Modelo general*:
    - Similitud: parte concordante entre objetos
    - Características concordantes conducen a, por ejemplo, "100% similitud"
    - También se puede considerar como un problema de optimización: "¿Cuánto cuesta transformar un objeto en otro?"

2.1 Conceptos básicos

- Modelos para definir objetivamente “similitud”
  - *Similitud basada en distancias*:
    - Una función de distancia mide la disimilitud entre objetos
      - A mayor distancia, más disímiles los objetos
      - Un objeto *q* tiene (por lo general) distancia 0 a sí mismo
    - Se puede formalizar matemáticamente
      - Espacios métricos
      - Espacios vectoriales

2.1.1 Espacios métricos

- Definición de espacio métrico
  - Universo de objetos válidos: $\mathbb{X}$
  - Función de distancia: $\delta : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}^+$
  - $(\mathbb{X}, \delta)$ representa un espacio métrico ssi δ cumple con las siguientes propiedades:
    i. Positividad estricta $\forall x, y \in \mathbb{X},\ x \neq y \Rightarrow \delta(x, y) > 0$
    ii. Simetría $\forall x, y \in \mathbb{X},\ \delta(x, y) = \delta(y, x)$
    iii. Reflexividad $\forall x \in \mathbb{X},\ \delta(x, x) = 0$
    iv. Desigualdad triangular $\forall x, y, z \in \mathbb{X},\ \delta(x, z) \leq \delta(x, y) + \delta(y, z)$

2.1.1 Espacios métricos

- Objetos de $\mathbb{X}$ son directamente comparados utilizando δ
- δ indica el grado de disimilitud entre dos objetos
- Ejemplo: strings + distancia de edición
  - *String*: secuencia de caracteres
  - *Distancia de edición*: mínimo # de inserciones, borrados o substituciones para transformar un string en otro

5

2.1.2 Espacios vectoriales

- Espacio vectorial: caso particular de espacio métrico
- $\mathbb{R}^d$: *d*-tuplas de números reales (*vectores*)

$$x \in \mathbb{R}^d \Rightarrow x = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_d \end{bmatrix}, \quad x_i \in \mathbb{R}$$

6

2.1.2 Espacios vectoriales

- Métricas para espacios vectoriales:
  - *Distancias de Minkowski*:

$$L_p(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{1/p}, \quad p \geq 1$$

- Ejemplos:
  - *Manhattan (p=1)*: $L_1(x,y) = \sum_{i=1}^{d} |x_i - y_i|$
  - *Euclidiana (p=2)*: $L_2(x,y) = \sqrt{\sum_{i=1}^{d} |x_i - y_i|^2}$
  - *Máximo (p=inf)*: $L_\infty(x,y) = \max_{i=1}^{d} |x_i - y_i|$

7

2.1.2 Espacios vectoriales

- Métricas para espacios vectoriales:
  - *Distancia de Mahalanobis:*

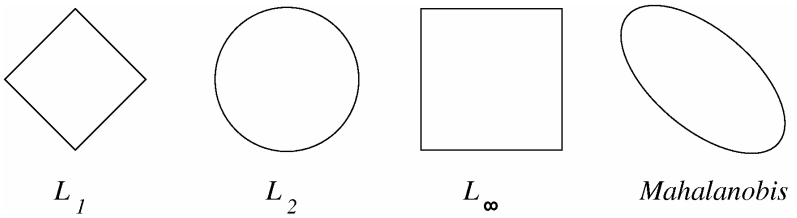
$$\delta(x,y) = \sqrt{(x-y)^T C^{-1} (x-y)}$$

    - *C*: matriz de covarianza de la distribución de los vectores
    - operador T: vector/matriz transpuesto
  - Minkowski con pesos

8

2.1.2 Espacios vectoriales

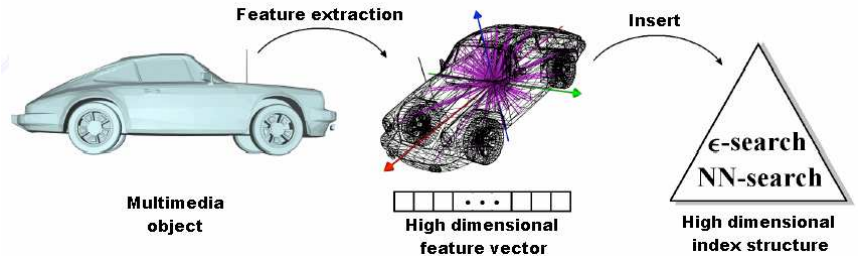
- Formas inducidas por las distintas métricas



9

2.1.2 Espacios vectoriales

- Método de vectores característicos



- Búsqueda se reduce a buscar puntos cercanos en un espacio vectorial

10

2.1.2 Espacios vectoriales

- Función de transformación
  - Dependiente del tipo de dato multimedia
  - Definida por un experto
- Dimensión del espacio: parámetro de la función de transformación
  - Representaciones más finas se obtienen usando más dimensiones
  - En general existe un punto de saturación
- Usualmente la transformación es irreversible

11

2.2 Consultas por similitud

- Consulta por rango
  - Objeto de consulta: $q \in \mathbb{X}$
  - Radio de tolerancia: $r \in \mathbb{R}^+$

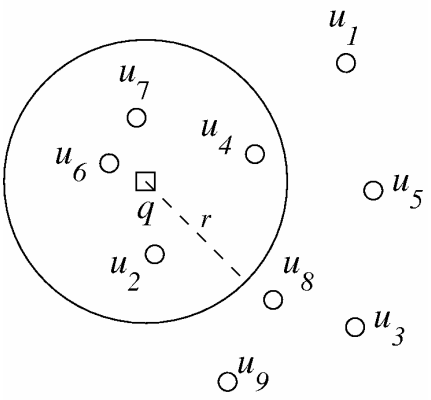
$$(q, r) = \{u \in \mathbb{U},\ \delta(u, q) \leq r\}$$

  - Bola de consulta: subespacio definido por *q* y *r*

12

2.2 Consultas por similitud

- Ejemplo de consulta por rango en $(\mathbb{R}^2, L_2)$

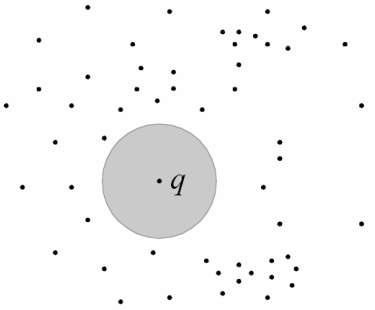


$$(q, r) = \{u_2, u_4, u_6, u_7\}$$

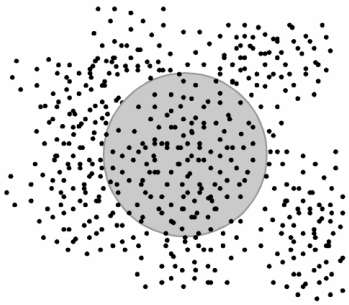
13

2.2 Consultas por similitud

- Problema de la consulta por rango: ¿Qué valor debe tener el radio de tolerancia?



Muy pequeño: no retorna nada



Muy grande: retorna demasiado

14

2.2 Consultas por similitud

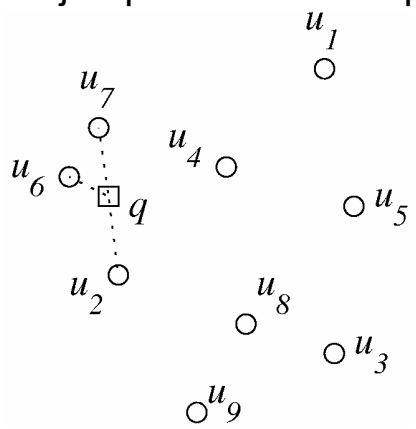
- Consulta por vecinos más cercanos (*k*-NN)
  - Objeto de consulta: $q \in \mathbb{X}$
  - Número de vecinos: $k \in \mathbb{N}$
  - Retorna conjunto $\mathbb{C}$, $|\mathbb{C}| = k$ tal que

$$\forall\, x \in \mathbb{C},\ y \in \mathbb{U} - \mathbb{C},\ \delta(x,q) \leq \delta(y,q)$$

15

2.2 Consultas por similitud

- Ejemplo de consulta por 3-NN en $(\mathbb{R}^2, L_2)$



$$3 - NN(q) = \{u_2, u_6, u_7\}$$

16

2.2 Consultas por similitud

- Consulta por ranking incremental
  - Motivación: A menudo uno no conoce ni un radio de tolerancia adecuado ni un valor de *k* razonable
  - Ejemplo: búsqueda en Internet
  - Se desea un resultado ordenado por distancias al objeto de consulta: ranking
  - También conocida como consulta give-me-more

17

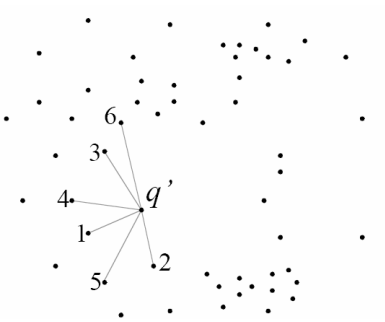
2.2 Consultas por similitud

- Método de consulta por ranking incremental
  - Al empezar, especificar objeto de consulta *q*
  - Repetir llamado de función *getnext*(k_i), la cual retorna los siguientes k_i objetos relevantes, hasta que se alcance la cantidad deseada de objetos

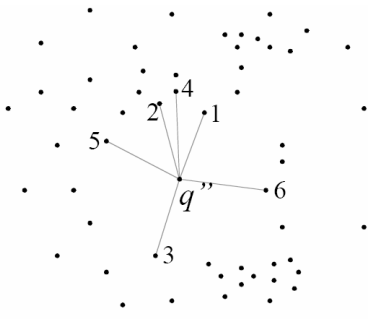
18

2.2 Consultas por similitud

- Ejemplo de consulta por ranking incremental



6x *getnext(1)* para *q'*



6x *getnext(1)* para *q''*

19

2.3 Efectividad y eficiencia

- Eficiencia
  - Se relaciona con el costo de búsqueda
  - Qué se mide: tiempo de CPU y tiempo de E/S
- Algoritmo ingenuo de búsqueda por similitud: búsqueda secuencial
- Estructuras de datos para agilizar búsquedas
  - Índices multidimensionales (*spatial access methods*) [BBK01]
  - Índices métricos (*metric access methods*) [CNB01]

20

2.3 Efectividad y eficiencia

- Efectividad
  - Calidad de la respuesta retornada por el sistema
  - Ejemplo en espacios vectoriales: función de transformación efectiva mapea dos objetos similares en puntos cercanos
- Descriptores finos se obtienen con resoluciones altas
  - No necesariamente implican mejor efectividad
- Efectividad = Eficacia

21

2.3.1 Evaluación de la efectividad

- Evaluación de la efectividad de un sistema de búsqueda por similitud
  - Medir su habilidad de recuperar objetos relevantes de la BD y de evitar los objetos no relevantes
- Medidas de efectividad
  - “Ground truth”: Colección de prueba
  - Medida de evaluación: cuantifica *similitud* entre objetos recuperados y objetos relevantes

22

2.3.1 Evaluación de la efectividad

- Objetos relevantes vs. no relevantes

	Deseado	No deseado
Encontrado	*Positivo correcto (RP)*	*Falso positivo (FP)*
No encontrado	*Falso negativo (FN)*	*Negativo correcto (RN)*

23

2.3.1 Evaluación de la efectividad

- Precision y recall [BR99]
  - Precision: ¿Cuántos de los objetos recuperados son relevantes?

$$Precision = \frac{RP}{RP+FP}$$

  - Recall: ¿Cúantos de los objetos relevantes fueron encontrados?

$$Recall = \frac{RP}{RP+FN}$$

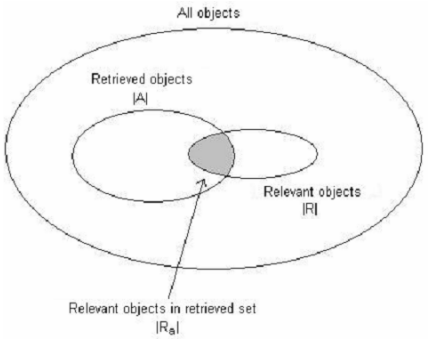
24

2.3.1 Evaluación de la efectividad

- Precision y recall

$$Precision = \frac{|R_a|}{|A|}$$

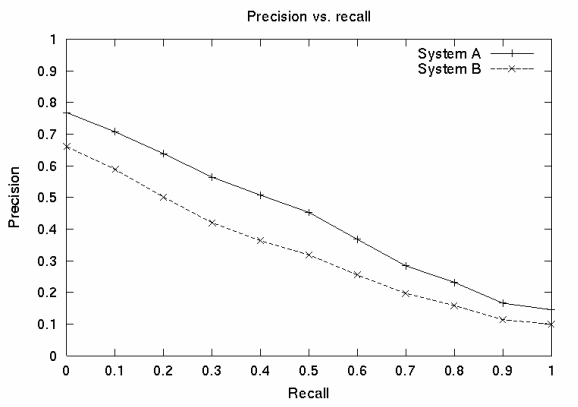
$$Recall = \frac{|R_a|}{|R|}$$



25

2.3.1 Evaluación de la efectividad

- Gráfico precision vs. recall



Sistema A es más efectivo que el sistema B

26

2.3.1 Evaluación de la efectividad

- Medidas de un solo valor
  - R-precision ("first tier") [BR99]: precision calculada cuando

$$|A| = |R|$$

  - Bull-Eye Percentage ("second tier") [ZP01]: recall calculado cuando

$$|A| = 2|R|$$

27

2.3.1 Evaluación de la efectividad

- Medidas alternativas
  - Harmonic mean [SBP97]; E measure [Rij79]

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}}$$

$$E(j) = 1 - \frac{1+b^2}{\frac{b^2}{r(j)} + \frac{1}{P(j)}}$$

$r(j)$: recall j^{th} objeto en el ranking

$P(j)$: precision j^{th} objeto en el ranking

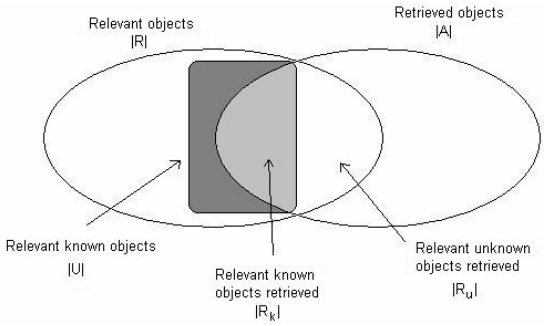
b: parámetro del usuario

28

2.3.1 Evaluación de la efectividad

- Coverage y novelty [Kor97,BR99]

$$Coverage = \frac{|R_k|}{|U|} \qquad Novelty = \frac{|R_u|}{|R_u|+|R_k|}$$



29

2.3.1 Evaluación de la efectividad

- Relative recall, relative effort [Kor97]

$$Relative\ recall = \frac{\#\ \text{objetos relevantes recuperados}}{\#\ \text{objetos relevantes deseados}}$$

$$Relative\ effort = \frac{\#\ \text{objetos relevantes deseados}}{\#\ \text{objetos examinados para encontrarlos}}$$

30

2.3.1 Evaluación de la efectividad

- Satisfacción y frustración [Kor97]
  - Objetos juzgados en una escala de 5 puntos
  - {0,1}: no relevante; {2,3,4}: relevante

$Retrieved = \{3, 0, 4, 2, 1\}$

$Ideal = \{4, 3, 2, 1, 0\}$

(a) $Satisfaction = \{3, 3, 7, 9, 9\}$

(b) $Frustration = \{0, 2, 2, 2, 3\}$

$Total = \alpha Satisfaction - \beta Frustration$

$\alpha = 1 \wedge \beta = 1 \Rightarrow \{3, 1, 5, 7, 6\}$

$Total\ ideal = \{4, 7, 9, 8, 6\}$

31

2.3.2 Colecciones de referencia

- Colección de referencia
  - Colección de documentos usados para probar modelos de RI y algoritmos [BR99]
  - Usualmente incluye:
    - Conjunto de objetos
    - Conjunto de consultas
    - Conjunto de objetos relevantes a cada consulta

32

2.3.2 Colecciones de referencia

- TREC collection
  - Text REtrieval Conference, empezó en 1992.
  - Colección de documentos de TREC
    - Varios gigabytes de datos
    - Documentos provienen de fuentes diversas
    - Conjunto de documentos relevantes se obtienen vía *método de pooling*.
  - URL: http://trec.nist.gov/

33

2.3.2 Colecciones de referencia

- TREC video retrieval evaluation
  - Objetivo: promover el progreso en content-based retrieval de video digital vía evaluación abierta basada en métricas"
  - 2001 and 2002: video "track" en TREC dedicada a investigar segmentación automática, indexamiento y content-based retrieval de video digital
  - 2003: Evaluación independiente (TRECVID)

34

2.3.2 Colecciones de referencia

- TRECVID
  - Cuatro tareas principales:
    - Determinación de los bordes de un “shot”
    - Segmentación de historias
    - Extracción de características de alto nivel
    - Búsqueda
  - Datos en video:
    - 120 horas de ABC World News Tonight y CNN Headline News (fines de enero - junio 1998)
    - 13 horas de programación de C-SPAN (entre 1998 – 2001).
  - URL: http://www-nlpir.nist.gov/projects/trecvid/

35

2.3.2 Colecciones de referencia

- Base de datos de fibrosis cística [SWW+91]
  - 1,239 documentos publicados entre 1974 y 1979 discutiendo aspectos de la fibrosis cística
  - Un conjunto de 100 consultas con su respectivo set de documentos relevantes como respuesta
  - Conjunto de scores de relevancia generados por expertos (0 a 8 puntos)
  - URL: http://www.sims.berkeley.edu/~hearst/irbook/cfc.html

36

2.3.2 Colecciones de referencia

- Princeton Shape Benchmark [SMK+04]
  - Base de datos y herramientas para recuperación de objetos 3D
  - 1,814 modelos 3D:
    - Clasificación base para entrenamiento, 90 clases, 907 modelos
    - Clasificación base para pruebas, 92 clases, 907 modelos
  - URL: http://shape.cs.princeton.edu/benchmark/index.cgi

37

2.4 Referencias

- [BBK01] C. Böhm, S. Berchtold, and D. Keim. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. *ACM Computing Surveys*, 33(3):322—373, 2001
- [BR99] R. Baeza-Yates, B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999
- [CNB01] E. Chávez, G. Navarro, J. Marroquín, and R. Baeza-Yates. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273—321, 2001
- [Kor97] R. Korfhage. Information Storage and Retrieval. John Wiley & Sons, Inc., New York, 1997
- [Rij79] C. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979
- [SBP97] W. Shaw, R. Burgin, and P. Howell. Performance standards and evaluations in IR test collections: Cluster-based retrieval models. *Information Processing & Management*, 33(1): 1—14, 1997
- [SMK+04] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser. The Princeton Shape Benchmark. Shape Modeling International, 2004
- [SWW+91] W. Shaw, J. Wood, R. Wood, and H. Tibbo. The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research*, 13:347—366, 1991
- [ZP01] T. Zaharia and F. Prêteux. 3D-shape-based retrieval within the MPEG-7 framework. In *Proc. SPIE Conference on Nonlinear Image Processing and Pattern Analysis XII*, volume 4304, pages 133—145, 2001

38