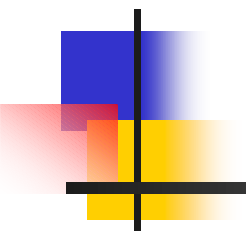
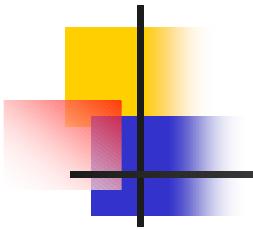


Introducción a los SARI

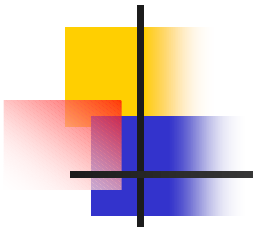
E.S. Informática CR - UCLM

1/Octubre/2002



Índice

1. Objetivos
2. Conceptos básicos
3. Evaluación de la recuperación
4. Perspectiva histórica



Objetivos

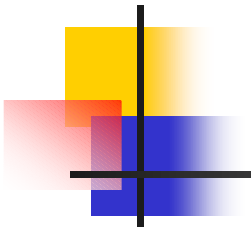
- Concepto de *Almacenamiento y Recuperación de la Información* como proceso
- Máquina *Memex* del Dr. Vannervar Bush
- Parámetros fundamentales en el método de almacenar la información:
  - **Tipo de Información:** Numérica, textos en lenguaje natural, sonidos, fotos,...
  - **Métodos de recuperación:** Lenguaje formal no ambiguo, lenguaje formal ambiguo, métodos exploratorios,...

Definición de recuperación de información

- **Baeza – Yates [1999]:** Parte de la informática que estudia la recuperación de la información (no datos) de una colección de documentos escritos. Los documentos recuperados pueden satisfacer *una necesidad de información de un usuario* expresada normalmente en lenguaje natural.
- **Korfhage [1997]:**La localización y presentación a un usuario de información *relevante* a una necesidad de información expresada como una pregunta.
- **Salton [1989]:** Un sistema de recuperación de información procesa archivos de registros y peticiones de información, e identifica y recupera de los archivos ciertos registros en respuesta a las peticiones de información.

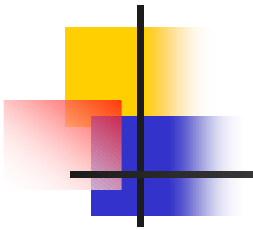
Sinónimos

- Sistema de almacenamiento y recuperación de la información (SARI).
- Sistema de recuperación de información.
- Sistema de recuperación de textos.
- Base de datos documental.
- Base de datos de información no estructurada (Base de datos no estructurada).



Diferencias SARI vs SGBD

- Un SGBD tiene un lenguaje de descripción de datos no ambiguo. Se suele expresar que la información de un SGBD es información estructurada mientras que en un SARI es información no estructurada.
- Un SGBD tiene un lenguaje de consulta formal no ambiguo.
- Hay más información no estructurada (libros, artículos, revistas, notas, e-mails, etc.) que información estructurada.



Conceptos básicos

- Componentes básicos de un SARI:
  - Almacenamiento de información.
  - Caracterización de las preguntas.
  - Identificación de documentos relevantes a las preguntas.

Definiciones

- **Documento:** Item de almacenamiento y recuperación.
- **Documento estructurado:** Documento con una división interna describiendo elementos externos al contenido del documento (fecha de creación, lugar, ...) o describiendo la jerarquía de contenidos (capítulos, secciones, subsecciones,...).
- **Documento de texto completo:** Documento almacenado en un SARI con el contenido completo sin resumir.
- **Palabra clave:** Una palabra elegida por el autor, editor o automáticamente, para representar el contenido del documento. Sinónimo: término.
- **Relevante:** Importante, significativo. Sobresaliente, excelente.

Caracterización de las preguntas del usuario

- La modelización de las preguntas se puede clasificar en dos grandes grupos:
  - Una pregunta es una expresión que deben satisfacer los documentos a recuperar.
  - Una pregunta es otro documento que es cercano "semánticamente" al conjunto de documentos a recuperar.
- Problema: Contexto del usuario (perfil del usuario).

Interacción del usuario con el sistema

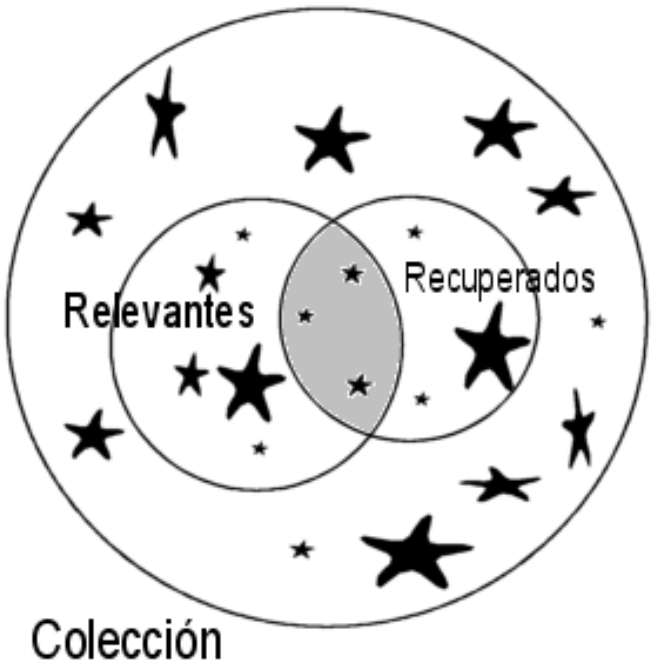
- Recuperación inmediata: El usuario desea recuperar los documentos o referencias en una sección.
  - Navegación: El sistema ofrece enlaces entre documentos para que el usuario determine los documentos a recuperar.
  - Recuperación "ad hoc": El usuario expresa en un lenguaje de consulta su necesidad y el sistema devuelve un conjunto de documentos. Variante: *Realimentación del usuario.*
- Recuperación diferida: El usuario desea recibir de forma continua los nuevos documentos que lleguen al sistema y que concuerden con sus necesidades.

Evaluación de la recuperación

- **Eficiencia:** Grado de rendimiento de un sistema referido al uso de sus recursos computacionales.
- **Eficacia:** Grado de rendimiento de un sistema referido a la consecución del objetivo de recuperación medido por el usuario.

Tabla de contingencia

	Recuperados	No Recuperados	
Relevante	RvRc	RvNRc	Rv = RvRc + RvNRc
No Relevante	NRvRc	NRvNRc	
	Rc = RvRc+NRvRc		N



Métricas simples

- **Precisión:** Proporción entre documentos recuperados relevantes y documentos recuperados.

$$P = RvRc/Rc$$

- **Índice de recuperación (recall):** Proporción entre documentos recuperados relevantes y documentos relevantes en la colección.

$$R = RvRc/Rv$$

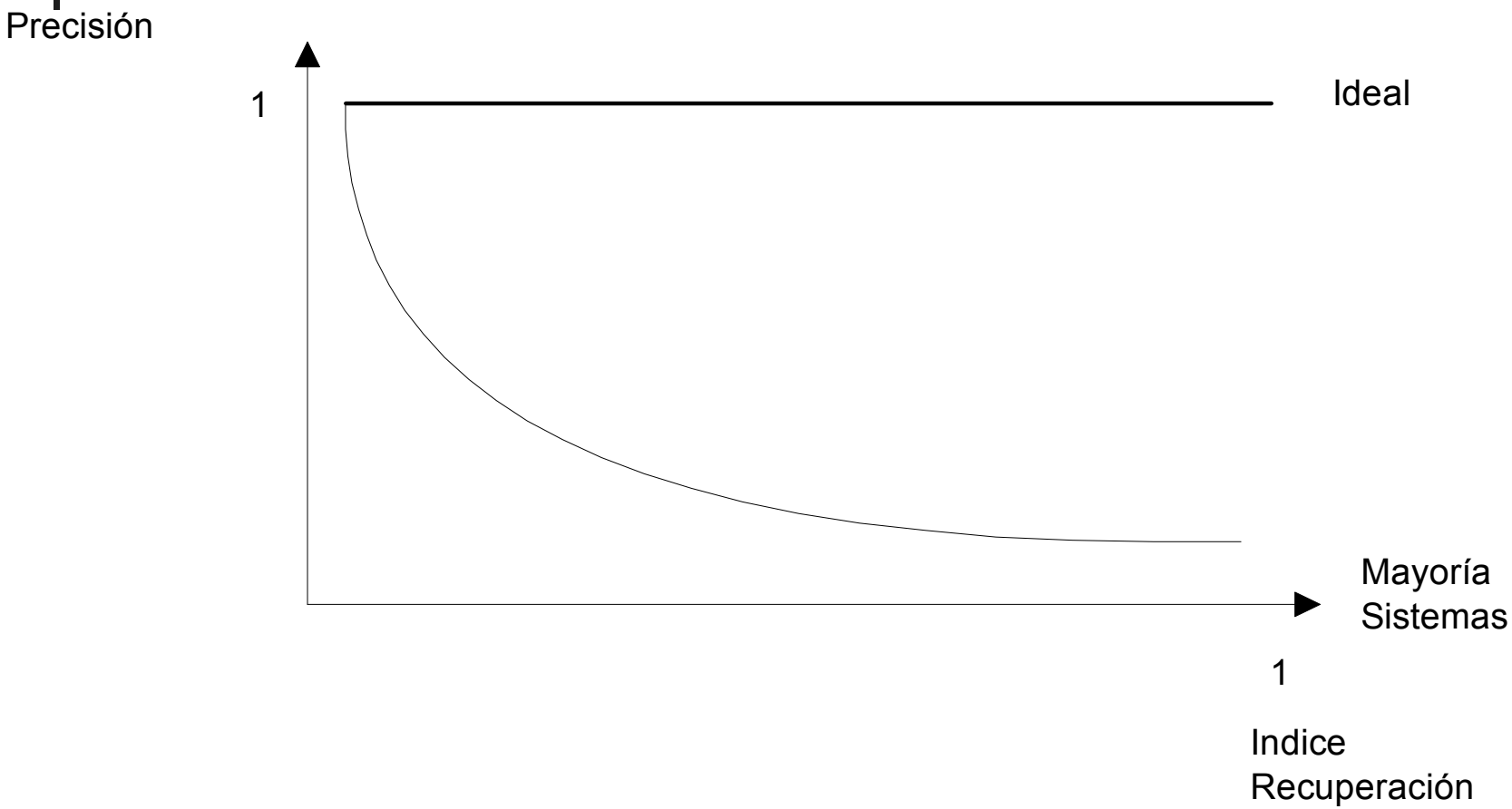
- **Índice de relevancia (generality):** Proporción entre documentos relevantes y el tamaño de la colección.

$$G = Rv/N$$

- **Índice de fallos (fallout):** Proporción entre documentos no relevantes recuperados y documentos no relevantes en la colección.

$$F = NRvRc/N\text{-}Rv$$

Métricas Simples



Relación entre métricas

$$\frac{R}{F} = \frac{\frac{P}{(1-P)}}{\frac{G}{(1-G)}}$$

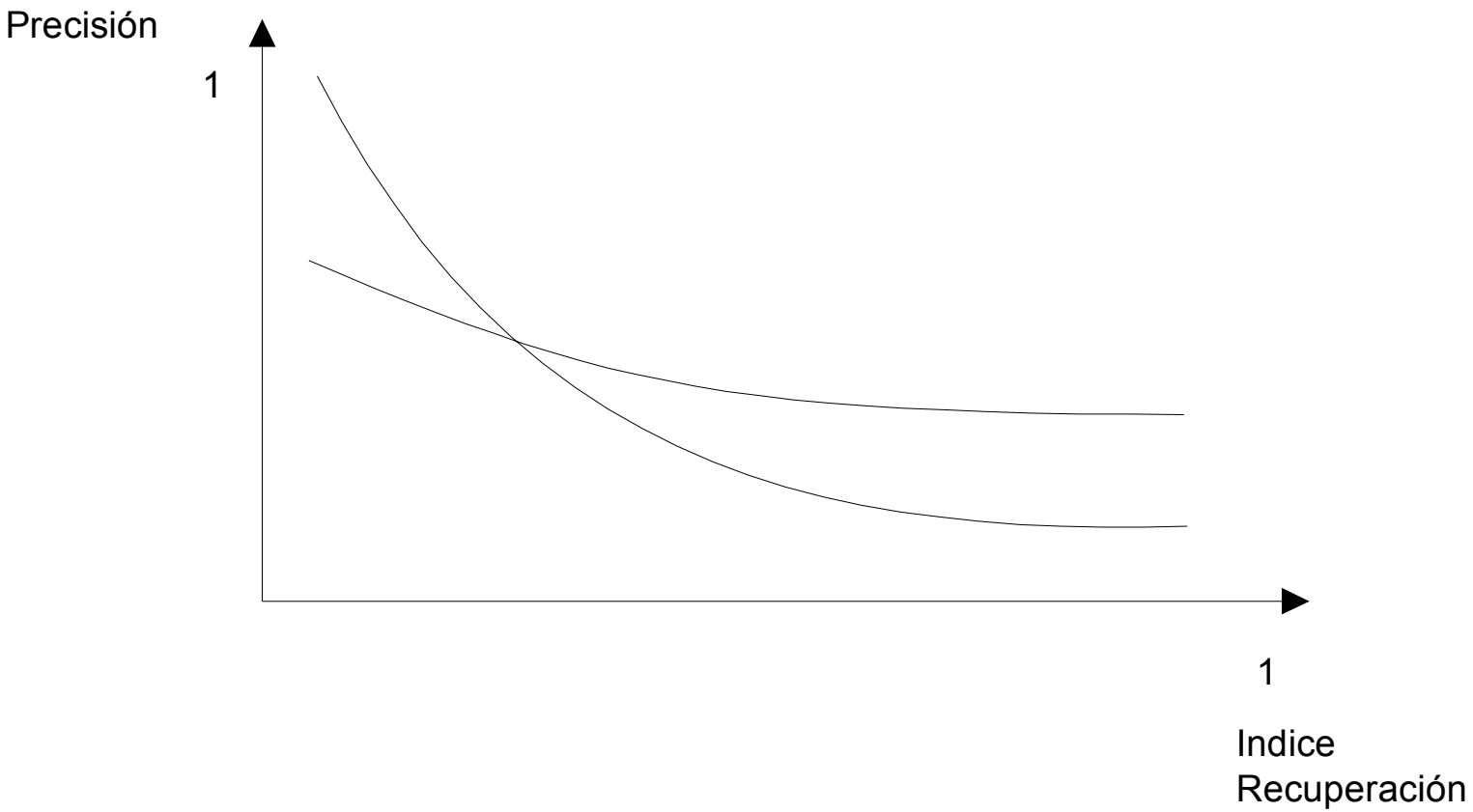
$$RG(1-P) = FP(1-G)$$

- **P/(1-P)** es el ratio entre documentos relevantes recuperados y documentos no relevantes recuperados
- **G/(1-G)** es el ratio entre documentos relevantes en la colección y documentos no relevantes en la colección
- **R/F** es el ratio de la eficacia en la recuperación de los documentos relevantes respectos de los documentos no relevantes.

Problemas de las métricas

- La precisión puede ser medida exactamente pero el índice de recuperación no.
- No es claro que el índice de recuperación y la precisión sean significativos para el usuario.
- La precisión y el índice de recuperación están relacionados pero por separado no son significativos (Dos medidas dificultan la comparación).

Problemas de las métricas



Medidas combinan precisión e índice de recuperación

$$F = \frac{2}{\frac{1}{R} + \frac{1}{P}}$$

- **Media armónica:** La función F tiene un valor entre [0,1].
  - F vale 0 cuando no ha recuperado ningún documento relevante
  - F vale 1 cuando recupera todos los documentos relevantes.
  - La media armónica tiende a ponderar por igual P y R.

Medidas combinan precisión e índice de recuperación

$$E = \frac{1 + b^2}{\frac{b^2}{R} + \frac{1}{P}}$$

- **Medida E:** El usuario especifica si está más interesado en precisión o en el índice de recuperación.
  - b es un parámetro que indica la importancia relativa entre P y R.
  - Si b=1 es igual a F.
  - Si b>1 pondera P
  - Si b<1 pondera R

Métricas orientadas al usuario

- **Ratio de cobertura:** Proporción entre los documentos relevantes ya conocidos por el usuario y todos los documentos relevantes de la colección.
- **Ratio de novedad:** Proporción entre los documentos relevantes recuperados nuevos y todos los documentos relevantes recuperados.

Métricas orientadas al usuario

- **Índice de recuperación relativo:** Proporción entre los documentos relevantes recuperados y el número de documentos relevantes deseados por el usuario.
  - Se puede medir
  - No tiene por que llegar a 1 si no hay suficientes documentos relevantes en la colección.

Métricas orientadas al usuario

- **Esfuerzo de recuperación:** Proporción entre el número de documentos relevantes deseados y el número de documentos examinados para conseguir ese número de documentos relevantes.
  - (d_{728},d_{45}*,d_{32},d_{1024}, d_{72}*, d_{7}*)
    - Para obtener 3 docs relevantes (deseados) tengo que leer 6 docs. Sería (3/6 = 0'5)
    - El ideal es 1.
    - Se puede medir
- La bondad de un sistema se mide con una batería de preguntas.

Precisión media

- Es la media de precisiones tomadas en unos valores concretos de índices de recuperación para un conjunto de preguntas:
  - Medias de 3 puntos (0.25, 0.5 y 0.75 de R)
  - Medias de 11 puntos (0, 0.1, 0.2, ..., 0.9,1)
  - Si no se tiene la precisión para un valor concreto del índice de recuperación se realiza una interpolación. Habitualmente se toma el valor máximo entre el valor anterior y el posterior

Perspectiva histórica

- **Sistemas preinformáticos:** Creación manual de índices para clasificar el contenido de libros (Bibliotecas).
- **1ª generación:** Mecanización de las fichas bibliográficas.
- **2ª generación:** Búsquedas más sofisticadas por palabras claves, etc...
- **3ª generación:** Interfaces gráficos, hipertexto, sistemas distribuidos, almacenamiento de documentos de texto completo.
- **Futuro:** Bibliotecas digitales
  - Ventajas: Bajo coste, acceso generalizado, libertad de publicación.
  - Problemas a resolver: Protección de copyright, pago por acceso, interoperabilidad entre bibliotecas digitales.