

Introducción a los SINAI

Sistemas Inteligentes de Acceso a la Información

José María Gómez Hidalgo
<http://www.esp.uem.es/~jmgomez/>

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

Índice

- Acceso y Recuperación de la Información
- Tipos de información
- Tareas y aplicaciones
- Descripción de la tarea
- Evaluación

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

Acceso y Recuperación de la Información

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

Acceso y Recuperación de la Información

- Sistemas Inteligentes de Acceso a la Información (SINAI) = S + IN + AI
- Sistemas (S) = visión práctica = ejemplos y técnicas de sistemas reales
- Inteligentes (IN) = avanzados e.g.
 - *Bruce Croft, Top 10 Research Issues for IR: What do people want from IR? DLib Magazine, Nov. 95*
 - “Magic” (Effective Vocabulary Expansion – expansion efectiva del vocabulario)
astronauta => cosmonauta

Acceso y Recuperación de la Información

- Acceso a la Información (AI)
 - *Marti Hearst, Current Topics in Information Access, 1998*
 - “The process by which users use information technology to seek, organize, and understand information”
 - *Marti Hearst, Context and Structure in Automated Full-Text Information Access, PhD Thesis, Berkeley, 1994*
 - “The term information access is beginning to supercede that of information retrieval since the latter’s implication is too narrow; the field should be concerned with information retrieval, display, filtering, and query facilitation”

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

Acceso y Recuperación de la Información

- Recuperación de Información (RI)
 - *E. Voorhees, D. Harman, Overview of the Eighth Text Retrieval Conference (TREC-8), NIST Special Publication 500-246, 1999*
 - “The ad hoc retrieval task investigates the performance of systems that search a static set of documents using new questions (...). This task is similar to how a researcher might use a library – the collection is known but the questions likely to be asked are not known.”

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

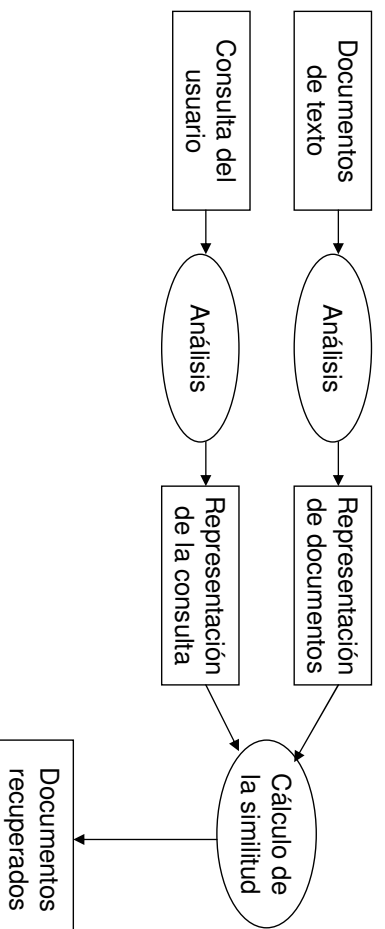
Acceso y Recuperación de la Información

- Aspectos de la tarea (RI)
 - Necesidad de información =? consulta del usuario
 - Colección estática de documentos (textuales)
 - El sistema recupera (presenta) documentos o fragmentos “relevantes” a la necesidad de información =? satisface la necesidad de información

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

Acceso y Recuperación de la Información

- Operativa del un sistema de RI



Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

Acceso y Recuperación de la Información

- Otras tareas TREC (tracks)
 - Cross-Language Track
 - Filtering Track
 - Genome Track
 - Novelty Track
 - Question Answering Track
- Por razones previas *acceso* >> *recuperación*
- Aunque nos centraremos sobre IR y texto

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

Tipos de información

Tipos de información

- Información vs. datos
 - Recuperación de datos (SGBDs)
 - Trabaja con datos con estructura y semántica definidas e inambíguas
 - No tolera fallos/imprecisión
- ```
select *
from employees
where salary > 100
```

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Tipos de información

- Recuperación de información
  - Trabaja con datos poco o nada estructurados y ambiguos (texto en lenguaje natural, video, etc.)
  - Es tolerante a fallos = se puede satisfacer la necesidad de información recuperando algunos documentos no relevantes

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

# Tipos de información

- Ejercicio
  - Describe la siguiente foto con 10 palabras
  - Pongamos en común
- Búsqueda de imágenes Google
- Consulta “r r m g”, resultado 10



Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

La Web

Imágenes

Grupos

Directorio

News | Nuevo!

cual es la montaña más alta de europa

Búsqueda en Google

Voy a Tener Suerte

Buscar en la Web

Buscar sólo páginas en español

Monte Eibruss

• Búsqueda Avanzada  
• Historiales  
• Herramientas del idioma

La Web Imágenes Grupos Directorio News | Nuevo!  
Se buscó **cual es la montaña más alta de europa** en la Web. Resultados 1 - 10 de aproximadamente 20,400. La búsqueda

Ascensiones en Semana Santa  
... de geógrafos y alpinistas la región del Caucaso pertenece a la **Europa** Meridional, lo **cual** convierte al Monte Eibruss como la **montaña** más **alta** de **Europa**. ...  
www.vanguardia.com/Ascensiones%20en%20Semana%20Santa.htm - 13K - [En caché](#) - [Páginas similares](#)

## En Almería?

DIVERSIDAD DE LA ENDEMOFLORA EN LA ALTA MONTAÑA DE ALMERÍA  
... y la superficie de la **alta montaña**, de nuevo resulta ... más importante del SE peninsular y de **Europa**. ... una gran diversificación mitroclimática, la **cual** se suma ...  
www.gem.es/MATERIALES/DOCUMENT/DOCUMENT/08/08211/0808211.htm - 54K - [En caché](#) - [Páginas similares](#)

## En los Alpes?

Desnivel.com - ENLACES ... Una **montaña** de enlaces de **montaña** ...  
... y la intenta colocar a la primera mujer sudamericana en el techo del mundo. Cordilleras de los Alpes Una cita ineludible con la **montaña** más **alta** de **Europa**. ...  
www.desnivel.com/textos/enlaces/index.php?node\_id=15 - 80K - [En caché](#) - [Páginas similares](#)

## Everest?

LA VANGUARDIA DIGITAL  
... **Europa** lanza su primera misión a la Luna JOSEP CORBELLA - 27/09/2003. ... GALERÍA DE FOTOS: Everest, la **montaña** más **alta**. GALERÍA DE FOTOS. ...  
www.lavanguardia.es/canalciencia/ - 84K - 1 Oct 2003 - [En caché](#) - [Páginas similares](#)

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Tipos de información

- Múltiples facetas
  - Medio y formato
    - Papel vs. electrónica, video vs. texto, GIF vs. JPEG
  - Estructura
    - Implícita, explícita, etiquetada, corrección
  - Tiempo
    - Importa?
  - Interdependencias entre unidades de información
    - Explícitas, implícitas

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

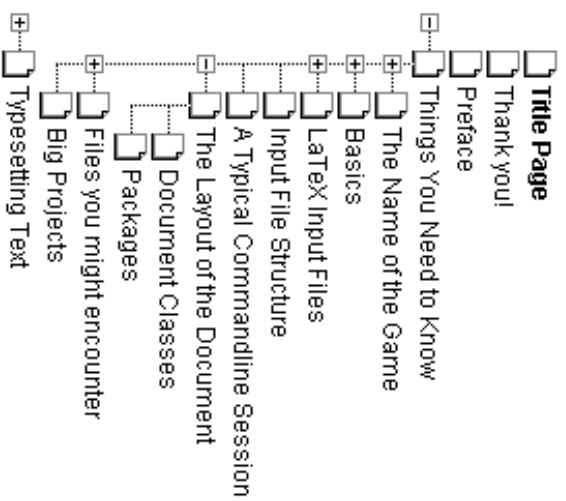
## Tipos de información

- Estructura implícita
  - Artículo (23 párrafos), revista *Discover*
  - Tema principal: *exploration of Venus by the space probe Magellan*
    - 1-2 *Intro to Magellan space probe*    12-15 *Styx channel*
    - 3 *Atmosphere obscures view*    16-17 *Aphrodite Highland*
    - 4 *Climate*    18 *Gravity readings*
    - 5- 7 *Meteors*    19-21 *Recent volcanic activity*
    - 8-11 *Volcanic activity*    22-23 *Future of Magellan*

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

# Tipos de información

- Estructura explícita
  - Vista Bookmarks (Acrobat Reader®)
  - *Tobias Oetiker et al., The Not So Short Introduction to LATEX2e, Version 3.20, 9 Agosto, 2001*



Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

# Tipos de información

- Estructura, etiquetado y corrección

```
<HTML>
<HEAD>
<TITLE>El título</TITLE>
</HEAD>
<BODY>
<H1>Introducción</H1>
...
<ADDRESS>
 Resúmen
</ADDRESS>
</BODY>
</HTML>
```

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Tipos de información

- **Carácter temporal**
  - Versiones de documentos técnicos
  - Novedad de noticias y seguimiento de eventos
- **Interdependencias**
  - Implícitas (manual para novatos, guía de referencia, temas avanzados)
  - Explícitas (hiperenlaces)

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Tareas y aplicaciones

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Tareas y aplicaciones

- Tareas TREC (ver arriba)
- Categorización de documentos
  - Clasificación de documentos en categorías predefinidas
- Agrupamiento de documentos
  - Clasificación de documentos en grupos coherentes

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Tareas y aplicaciones

- Extracción de resúmenes
- Traducción automática
- Agrupamiento de términos
  - Creación de clases de palabras coherentes (thesauri – diccionarios de sinónimos)
- ...
- Todas ***dan soporte*** al Acceso a la Información

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

# Tareas y aplicaciones

ACM Digital	Google	Páginas	SquidGuard
Library	Google News	blancas y	Teoma
Altavista	ICTnet	amarillas	Tucows
Amazon	KanguroNet	PhpNuke	WIKI
Blog	Kazaa	POESIA	Yahoo
Citeseer	Kelkoo	SIT	Yahoo Groups
EDonkey	Monster	Sofonic	Yahoo Maps
Emule	My Yahoo	SpamAssasin	Yahoo News
Froogle	ODP	Springer Link	

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Descripción de la tarea

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

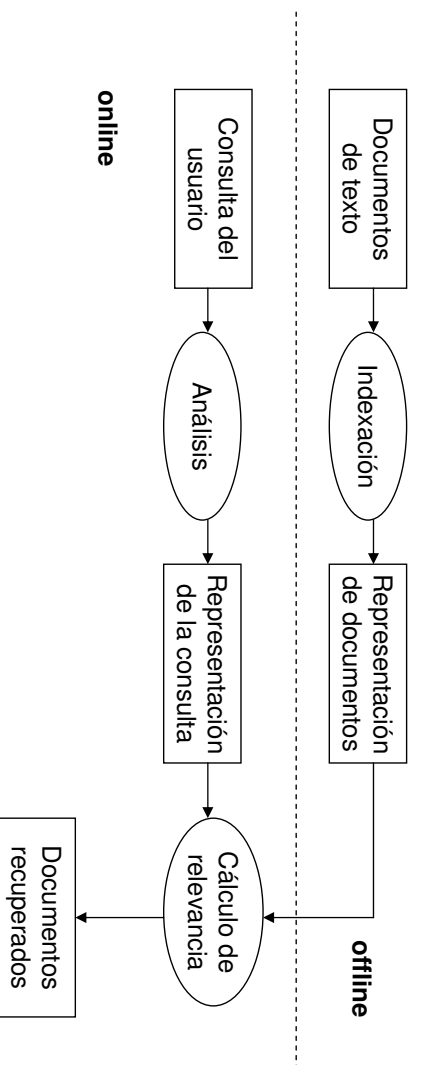
## Descripción de la tarea

- Nos concentramos en “ad-hoc retrieval”
1. Se parte de una colección de documentos
  2. Un usuario tiene una necesidad de información y plantea una consulta (query) al sistema de RI
  3. El sistema devuelve los documentos “relevantes” = satisfacen la necesidad, posiblemente en forma de “ranking”

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Descripción de la tarea

- **Procesos**



Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Descripción de la tarea

- Debemos proporcionar soporte para
  1. Representar documentos y consultas (*lenguaje de representación*)
  2. Procesar documentos (*indexación*) y consultas
  3. Obtener una aproximación de la relevancia de los documentos a la consulta (*cálculo de relevancia o similitud*)

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Evaluación

## Evaluación

- **Crítica para comparar sistemas/enfoques**
- **Usualmente evaluados en términos de**
  - **Efectividad**
    - La capacidad del sistema de satisfacer las necesidades del usuario = capacidad de recuperar exclusivamente documentos relevantes
  - **Eficiencia**
    - Teórica (complejidad)
    - Empírica (tiempos de respuesta)
  - **Nos centramos en efectividad**

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Evaluación

- **Se ignora e.g. el informe EAGLES (Expert Advisory Groups for Language Engineering Standards - Evaluation Working Group) – Estándar (ISO 9126)**

Características	Subcaracterísticas
Funcionalidad	Adecuación, precisión, interoperabilidad, conformidad, seguridad
Fiabilidad	Madurez, tolerancia a fallos, recuperabilidad
Facilidad de uso	Comprensibilidad, facilidad de aprendizaje, facilidad de operación
Eficiencia	Comportamiento respecto al tiempo y memoria
Facilidad de mant.	Analizabilidad, capacidad de modificación y prueba, estabilidad
Transportabilidad	Adaptabilidad, facilidad de instalación y mantenimiento, capacidad de ajuste

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Evaluación

- Múltiples métricas
  - Tasa de recuperación o cobertura (*recall*)
  - Precisión (*precision*)
- Cálculo en términos de documentos clasificados correcta e incorrectamente por el sistema para
  - Una serie de consultas tipo
  - Una colección de documentos estándar

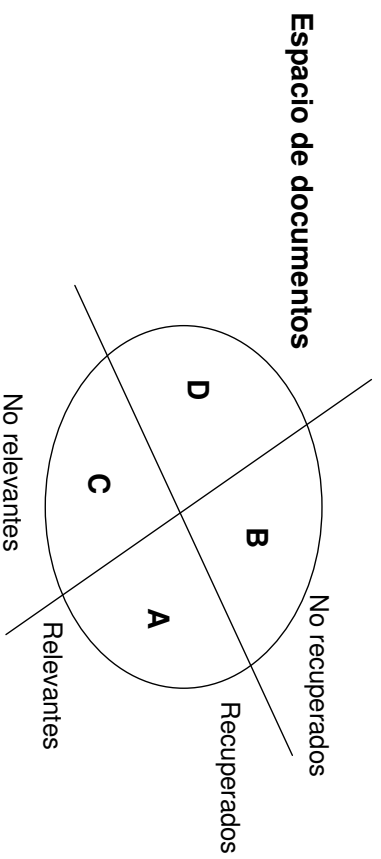
Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Evaluación

- Colecciones disponibles (documentos + consultas + juicios de relevancia)
  - Clásicas (CACM, ISI, etc) => minúsculas
  - Actuales, + realistas
    - OHSUMED
      - 400k documentos de medicina, inglés
    - TREC (e.g. TREC6)
      - + 600k documentos diversos (noticias, patentes, publicaciones oficiales, inglés)

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

# Evaluación



Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

# Evaluación

- Tabla de contingencia o confusión

	Recuperados	No recuperados
Relevantes	A	B
No relevantes	C	D

$$recall = \frac{A}{A + B} \quad precision = \frac{A}{A + C}$$

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

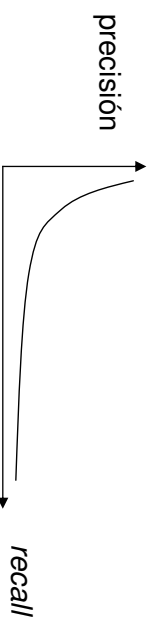
## Evaluación

- *Recall* = proporción entre documentos relevantes recuperados y documentos relevantes
- *Precisión* = proporción entre documentos relevantes recuperados y documentos recuperados

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Evaluación

- Generalmente, *recall* y *precisión* son inversamente proporcionales



- Se suele buscar un equilibrio entre ellas o primar la preferida por el usuario tipo

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Evaluación

- Dos factores adicionales
  - Varias consultas => necesidad de promediar
  - *Ranking* de documentos => necesidad de normalizar
- Promedio
  - Calcular para cada consulta y promediar => macro-media (*macroaveraging*) => todas las consultas tiene igual importancia
  - Sumar tablas para todas las consultas y calcular un sólo valor => micro-media (*microaveraging*) => las consultas con más documentos recuperados tienen mayor peso en la evaluación

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Evaluación

- Si no hay un *ranking* de documentos, sólo hay un valor de *recall* y precisión
- Si hay un *ranking* de documentos, se suele calcular la precisión en once niveles de *recall* = 0.0, 0.1, 0.2, ..., 0.9, 1.0  
=> se puede obtener una gráfica para comparar sistemas visualmente

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

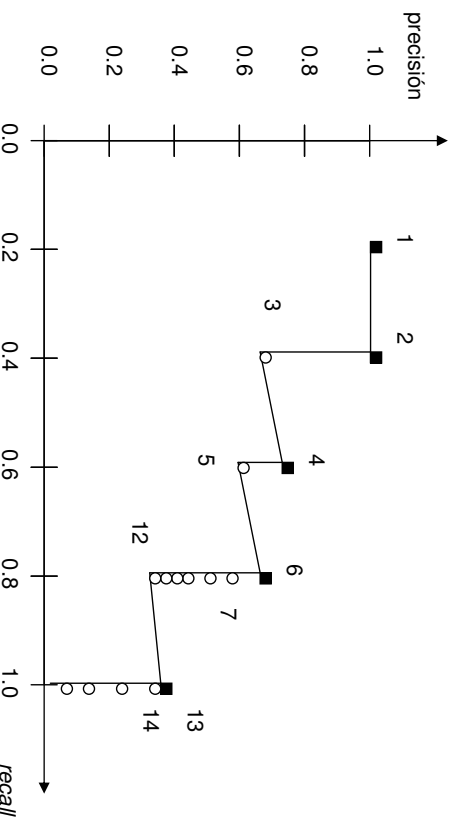
# Evaluación

reca//precisión tras haber recuperado 14 documentos (5 relevantes)

posición	núm. del documento. (x = relevante)	reca//	precisión
1	345 x	0.2	1.0
2	456 x	0.4	1.0
3	342	0.4	0.67
4	563 x	0.6	0.75
5	872	0.6	0.60
6	307 x	0.8	0.67
7	867	0.8	0.57
8	153	0.8	0.50
9	561	0.8	0.44
10	789	0.8	0.40
11	592	0.8	0.36
12	457	0.8	0.33
13	767 x	1.0	0.38
14	191	1.0	0.36

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

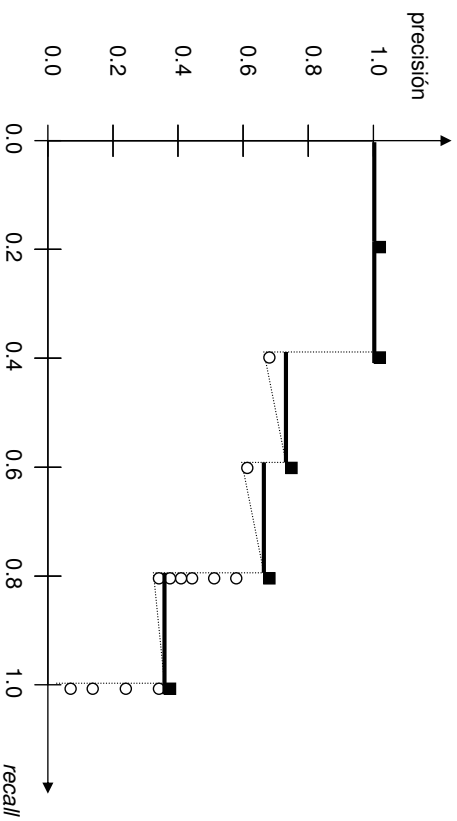
# Evaluación



Gráfica reca//precisión real

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

# Evaluación



Gráfica *recall*//*precisión* interpolada

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

# Evaluación

Tabla *recall*//*precisión* interpolada

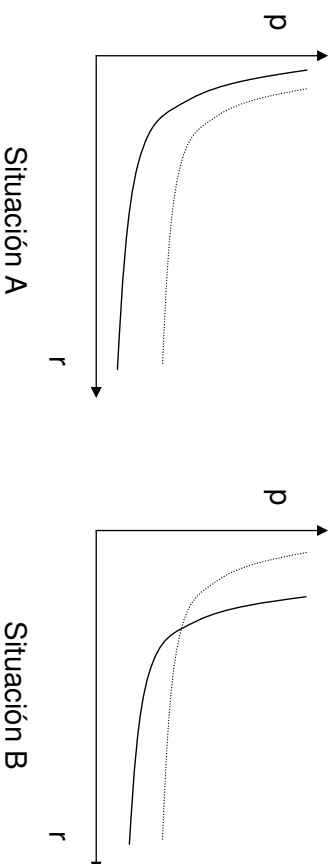
<i>Recall</i>	<i>Precisión</i>
0,0	1,0
0,1	1,0
0,2	1,0
0,3	1,0
0,4	1,0
0,5	0,75
0,6	0,75
0,7	0,67
0,8	0,67
0,9	0,38
1,0	0,38

Después se puede promediar sobre las consultas => macro-averaging

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Evaluación

- Comparación de dos técnicas o sistemas



Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo – U. Europea Madrid

## Evaluación

- Para obtener una sola medida de efectividad
  - Si se hace una evaluación basada en *ranking*, se calcula la media sobre los once valores de precisión
  - Si se ha obtenido un solo valor de *recall* y precisión se utiliza la  $F_{\beta}$  de van Rijsbergen

$$F_{\beta} = \frac{(\beta^2 + 1)pr}{\beta^2 p + r}$$

$F_1$  = igual importancia a  $r$  y  $p$   
 $F_0$  = sólo precisión  
 $F_{\infty}$  = sólo *recall*