Introducción al Procesamiento de Lenguaje Natural

Grupo PLN – InCo

Recuperación de Información

Recuperación de Información

Dada una colección de documentos:

- ¿cómo podemos recuperar información relevante para nosotros contenida en ellos?
- ¿qué cosas serán representativas de cada documento?
- ¿cómo se van a representar los documentos?
- ¿y las consultas?
- ¿importa la estructura del documento o sólo su contenido?
- ¿cuál va a ser la relevancia asociada a cada documento?

Recuperación de Información

- Introducción
- Modelos
- Evaluación de la recuperación
- Indexación
- CLIR
- RI en la web

Introducción

- ¿Qué es la Información?
- ¿Qué es un Sistema de Información?
- ¿Qué es la Recuperación de Información?

Introducción

Información:

(1) **5.** *f. Comunicación o adquisición de conocimientos que permiten ampliar o precisar los que se poseen sobre una materia determinada.* (RAE)

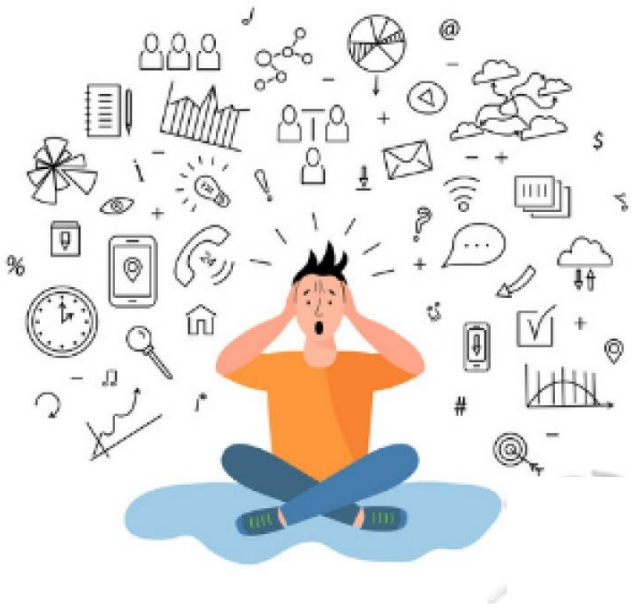
(2) *Es un conjunto organizado de datos procesados, que constituyen un mensaje que cambia el estado de conocimiento del sujeto o sistema que recibe dicho mensaje.* (Wikipedia)

Introducción

Información → *como soporte para la transferencia de conocimientos*

Introducción

Información → *como soporte para la transferencia de conocimientos*



Introducción

Información → *como soporte para la transferencia de conocimientos*

Introducción

Sistema de Información:

(1) Conjunto de funciones o componentes interrelacionados que forman un todo, es decir, obtiene, almacena, procesa y distribuye información para apoyar la toma de decisiones y el control en una organización. (Wikipedia)

(2) Conjunto organizado de elementos, que pueden ser personas, datos, actividades o recursos materiales en general. Estos elementos interactúan entre sí para procesar información y distribuirla de manera adecuada en función de los objetivos de una organización. (http://definicion.de/)

Introducción

Cualidades de un Sistema de Información:

- Precisión
- Oportunidad
- Completitud
- Contenido semántico
- Integridad (coherencia)
- Seguridad

Introducción

Algunos ejemplos:

- Transaccionales / Empresariales / Gestión
- Procesos de negocio (BPM)
- Soporte para la toma de decisiones
- Sistemas Expertos
- ...

Introducción

Se pueden clasificar en:

- Sistemas de Gestión de (Bases de) Datos
- Sistemas de Recuperación de Información

Introducción

Gestión de (Bases de) Datos:

es un conjunto de programas que permiten el almacenamiento, modificación y extracción de la información en una base de datos. Los usuarios pueden acceder a la información usando herramientas específicas de consulta y de generación de informes, o bien mediante aplicaciones. (Wikipedia)

Introducción

Recuperación de Información:

(1) es la ciencia de la búsqueda de información en documentos, búsqueda de los mismos documentos, la búsqueda de metadatos que describan documentos, o, también, la búsqueda en bases de datos, ya sea a través de Internet, intranet, para textos, imágenes, sonido o datos de otras características. (Wikipedia)

(2) es la disciplina encargada de la representación, almacenamiento y organización, y su posterior acceso y recuperación para responder a las necesidades de un usuario. (Salton)

Introducción

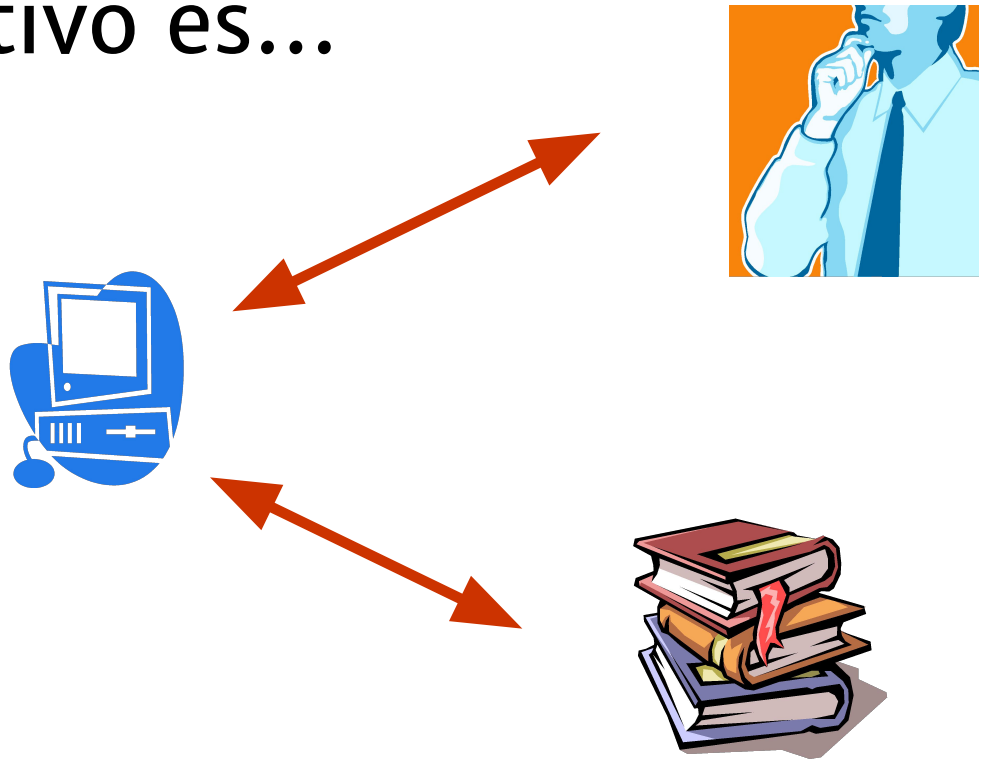
Introducción

Las siete etapas en la vida del hombre (en RI)

- Infancia - 1950 Luhn; extracción de concordancia en textos y creación de abstracts. Colecciones pequeñas.
- Escolar - 1960; sistemas comerciales (catálogos), vocabulario controlado
- Adolescencia - 1970; disminuyen los trabajos de investigación; se comienza con el catálogo on-line de la biblioteca del Congreso
- Madurez - 1980; incremento de BD on-line y CDs con catálogos
- Crisis de los 40 - 1990; Internet y buscadores (Mosaic, Yoahoo)
- Vejez - 2000; RI multimedia y textos on-line
- Muerte - 2010; Recuperación sintáctica → Recuperación semántica; CLIR; Calidad en la Recuperación

Introducción

El objetivo es...



Introducción

Recuperar datos vs Recuperar información

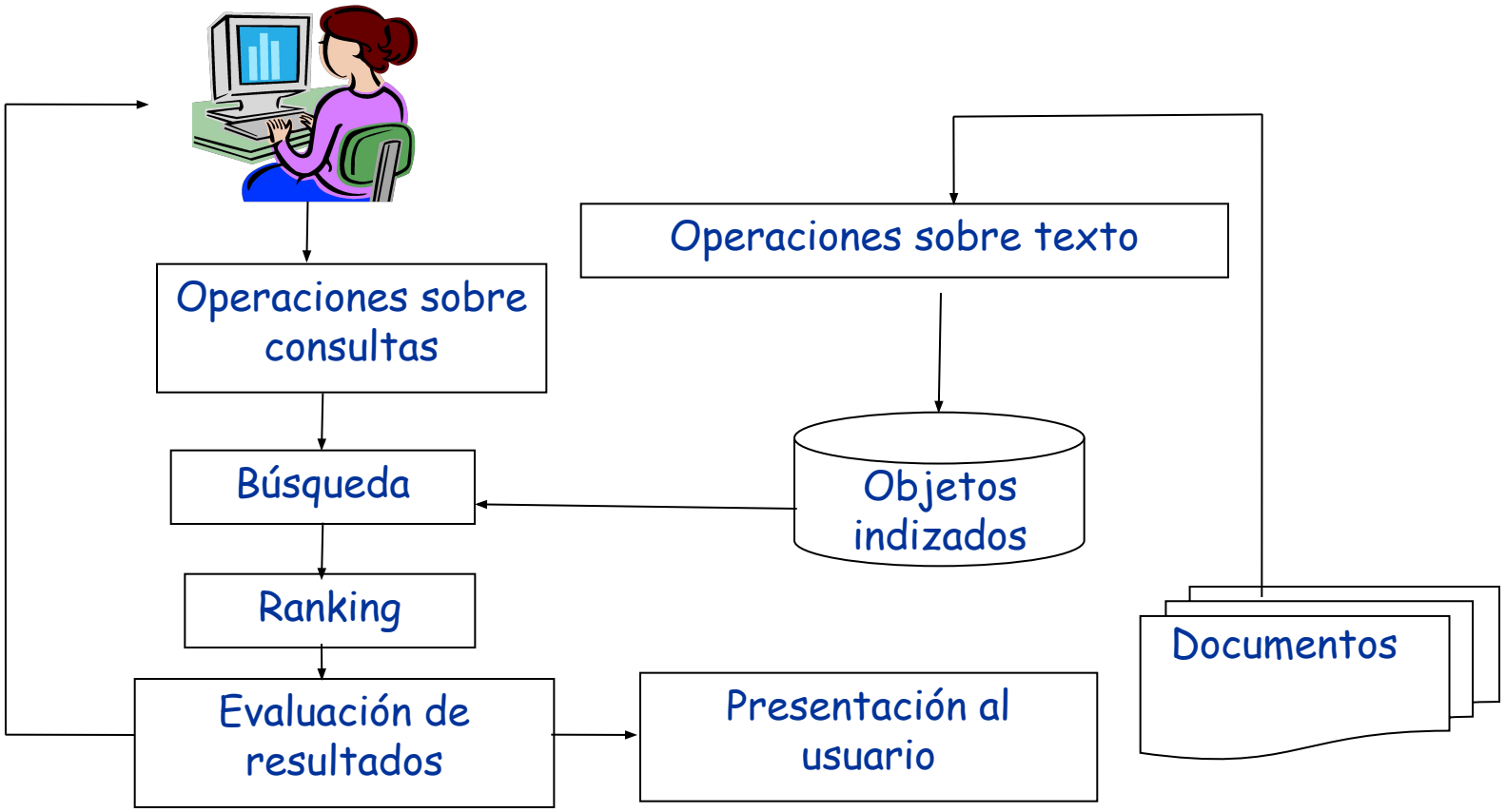
```
SELECT padron, valor
FROM   Padrones
WHERE  valor > 250000
```

Quiero saber equipos uruguayos clasificados a copas internacionales en los últimos años

Introducción

Tipos de Sistema	Objetos de datos	Operaciones básicas	Consultas	Tamaño de la BD
SGBD (Relacional)	Tablas (estructurados)	Recuperación (determinística)	Algebra relacional No ambiguas	Pequeña a muy grande
SRI	Documentos (no estructurados, texto)	Recuperación (probabilística)	Lenguaje Natural Puede ser ambigua	Pequeña a muy grande

Introducción



Introducción

"*Quiero información sobre las batallas de la Segunda Guerra Mundial que no se libraron en Europa*"

procesamiento intuitivo...

batallas Segunda Guerra Mundial Europa

Introducción

Lo básico en RI es muy simple:

- Dado un conjunto de palabras (consulta)
- Encontrar esas palabras en un conjunto de documentos
- Devolver el/los documento/s (si corresponde)

Introducción

Lo básico en RI es muy simple:

- Dado un conjunto de palabras (consulta)
- Encontrar esas palabras en un conjunto de documentos
- Devolver el/los documento/s (si corresponde)

pero....

¿cómo se eligen los documentos a devolver?

Introducción

Por otro lado, las consultas se representan por un conjunto de términos

- ¿cuáles?
- ¿de qué tipo?
- ¿cómo los elijo?

Introducción

Problemas:

- muchas formas de decir lo mismo
- un mismo término puede tener distintos significados en diferentes contextos
- predecir cuáles documentos son los más relevantes

Introducción

2 aspectos claves:

- Elegir un Modelo
  - una representación de los documentos y de las necesidades de información del usuario (consulta)
  - una función de ranking, que defina el orden de relevancia de los documentos respecto a una consulta concreta
- Diseñar algoritmos y estructuras de datos que lo implementen eficientemente

Modelos

Definición:

(1) abstracción de un proceso real

(2) esquema teórico de un sistema o realidad compleja que se elabora para facilitar su comprensión y estudio de su comportamiento

Modelos

Caracterización formal de un modelo en RI

$\{ D, Q, F, R(q_i, d_j) \}$ donde

- **D** es un conjunto de representaciones lógicas de los documentos de la colección
- **Q** es un conjunto de representaciones lógicas que modelan las consultas de los usuarios del sistema
- **F** es un esquema operacional que modele las representaciones de documentos y consultas y las correspondencias entre ellos
- **$R(q_i, d_j)$** como función de ranking que vincula un número real a la relación entre una consulta q_i de Q y un documento d_j de D

Modelos

- Clásicos
  - Booleano
  - Vectorial
  - Probabilístico
- Alternativos
  - Redes Neuronales
  - Redes Bayesianas
  - LSI (Latent Semantic Indexing)

Modelo Booleano

- Un documento se representa como un conjunto de términos (principalmente ***sustantivos***)
- Un documento es relevante o no (si contiene o no un término)
- Regla que lo rige:
  - *Vocabulario similar → contenido similar*
- Se utilizan los operadores clásicos para expresar las consultas:
  - AND → intersección
  - OR → unión
  - NOT → complemento

Modelo Booleano

"*Quiero información sobre las batallas de la Segunda Guerra Mundial que no se libraron en Europa*"

(batallas **OR** combates) **AND** "Segunda Guerra Mundial" **NOT** Europa

Modelo Booleano

Ejemplo

D1: En el jardín hay plantas y flores todo el año
D2: A mi mamá le gusta que le regalen flores y bombones

Términos = {jardín, plantas, flores, año, mamá, bombones}

Modelo Booleano

Ejemplo

D1: En el jardín hay plantas y flores todo el año
D2: A mi mamá le gusta que le regalen flores y bombones

Términos = {jardín, plantas, flores, año, mamá, bombones}
D1 = (1,1,1,1,0,0)
D2 = (0,0,1,0,1,1)

Modelo Booleano

Ejemplo

D1: En el jardín hay plantas y flores todo el año
D2: A mi mamá le gusta que le regalen flores y bombones

Términos = {jardín, plantas, flores, año, mamá, bombones}
D1 = (1,1,1,1,0,0)
D2 = (0,0,1,0,1,1)

q1: (flores **OR** plantas) **AND** mamá

q2: plantas **AND** jardín

Modelo Booleano

Ejemplo

D1: En el jardín hay plantas y flores todo el año
D2: A mi mamá le gusta que le regalen flores y bombones

Términos = {jardín, plantas, flores, año, mamá, bombones}
D1 = (1,1,1,1,0,0)
D2 = (0,0,1,0,1,1)

q1: (flores **OR** plantas) **AND** mamá
q1: (flores **AND** mamá) **OR** (plantas **AND** mamá)
(0,0,1,0,1,0) **OR** (0,1,0,0,1,0)

q2: plantas **AND** jardín
q2: (1,1,0,0,0,0)

Modelo Booleano

Ejemplo

D1: En el jardín hay plantas y flores todo el año
D2: A mi mamá le gusta que le regalen flores y bombones

Términos = {jardín, plantas, flores, año, mamá, bombones}
D1 = (1,1,1,1,0,0)
D2 = (0,0,1,0,1,1)

q1: (flores **OR** plantas) **AND** mamá
q1: (flores **AND** mamá) **OR** (plantas **AND** mamá)
(0,0,1,0,1,0) **OR** (0,1,0,0,1,0) → ***D2***

q2: plantas **AND** jardín
q2: (1,1,0,0,0,0) → ***D1***

Modelo Booleano

Críticas:

- No discrimina entre documentos más y menos relevantes
- Es indiferente que un documento contenga una o cien veces las palabras de la consulta
- Da lo mismo que cumpla una o todas las cláusulas de un OR
- No considera el "*casi* " de un documento (documento que cumpla casi todas las cláusulas de un AND)
- Puede no ser intuitivo de aplicar para el "usuario común"

Por ejemplo:

"Quiero investigar sobre los Charrúas y los Guaraníes"

¿Cuál es la consulta?

Modelo Booleano

A favor:

- Es de las primeras ideas que a uno se le ocurren
- Fue adoptado por los primeros SRI documentales
- Simple de formalizar

Modelo Vectorial

- Clásico en RI (Salton, 1968 – SMART)
- Se selecciona un conjunto de palabras (o términos) útiles para discriminar
- Maneja el concepto de "*grado de similitud*" entre consultas y documentos
- Se expresa la relación entre cada documento d_i de una colección de N documentos, con el conjunto de los k términos elegidos para indexar

Modelo Vectorial

- Los documentos y las consultas son ***vectores*** en un **espacio n-dimensional**

  Sea $\{t_1, t_2, \ldots, t_k\}$ el conjunto de términos

  Sea $\{d_1, d_2, \ldots, d_N\}$ el conjunto de documentos

- Un documento d_i se modela como un vector:

  $$d_i = (w_{i1}, w_{i2}, \ldots, w_{ik})$$

  donde a cada término j en un documento d_i se le asigna un peso w_{ij}

Modelo Vectorial

Asignación de pesos

df_i = cantidad de documentos que contienen el término ***i***

idf_i = frecuencia inversa del término ***i***

$= \log_2 (N/ df_i)$ (N: total de documentos)

- máximo = $\log_2 N$ → ocurre en un solo documento
- mínimo = 0 → ocurre en todos los documentos

Modelo Vectorial

Asignación de pesos (cont.)

- Los términos **más** frecuentes en un documento serían mejores indicadores del tema del documento

- Sin embargo, por otro lado, los términos que aparecen en muchos documentos diferentes no son buenos indicadores

f_{ij} = cantidad de ocurrencias del término *i* en el documento *j*

$tf_{ij} = f_{ij} / max\{f_{ij}\}$

- "Atenuamos" el crecimiento.
- Normalizamos, para ser más justos con los documentos más largos

Modelo Vectorial

Asignación de pesos (cont.)

- Existen varias fórmulas para la asignación de pesos
  - 1era. aproximación ... binaria
    si aparece el término i en el documento d_j el valor de w_{ij} es 1,
    en caso contrario es 0
  - Pero ... el término i puede aparecer más de una vez en el mismo documento d_j, o puede considerarse como más "significativo" que otro, entonces el w_{ij} es más "sofisticado" que una asignación binaria

- Un indicador típico de la "importancia" de un término sigue el esquema *tf . idf* :

$$w_{ij} = tf_{ij} * idf_i = tf_{ij} * \log_2 (N/ df_i)$$

Modelo Vectorial

Ejemplo (para asignación de pesos):

Sea un documento con términos A, B, C con frecuencias:

A(3), B(2), C(1)

10000 documentos en la colección y la cantidad de documentos que contienen cada uno de los términos es :

A:50, B:1300, C:250

Entonces:

A: tf = 3/3; idf = $\log_2(10000/50)$ = 5.3; tf * idf = 5.3

B: tf = 2/3; idf = $\log_2(10000/1300)$ = 2.0; tf * idf = 1.3

C: tf = 1/3; idf = $\log_2(10000/250)$ = 3.7; tf * idf = 1.2

Modelo Vectorial

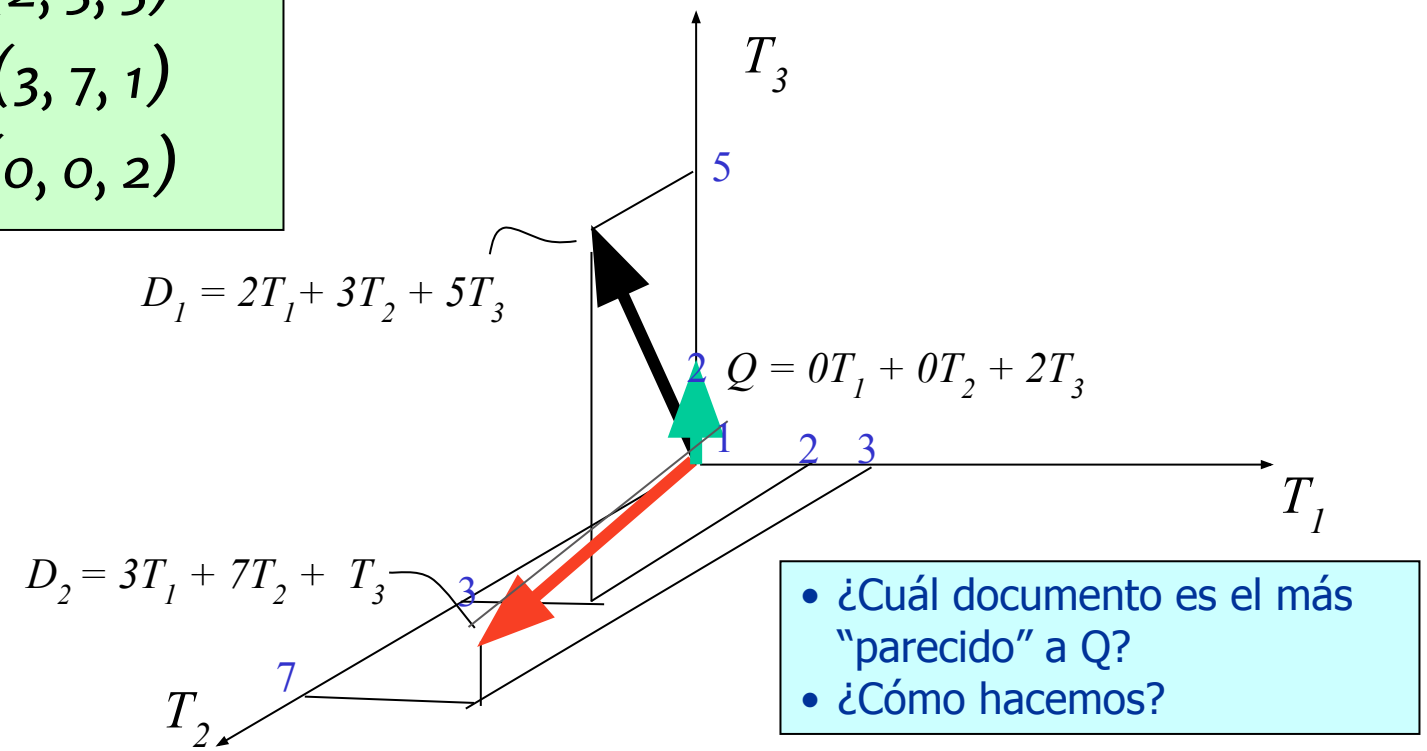
Una vez que:

- cada documento
- la consulta

tienen su "vector de términos" con los correspondientes pesos, hay que calcular la "*similitud*" entre cada documento y la consulta

Modelo Vectorial

$D_1 = (2, 3, 5)$
$D_2 = (3, 7, 1)$
$Q = (0, 0, 2)$



- ¿Cuál documento es el más "parecido" a Q?
- ¿Cómo hacemos?

Modelo Vectorial

- Una *medida de similitud* es una función que computa el ***grado de similitud*** entre 2 vectores (su distancia)

- Usando el grado de similitud entre la consulta y cada documento es posible:

  - "Ranquear" los documentos recuperados en orden de relevancia

  - Fijar un umbral por debajo del cual se considera que el documento no es relevante

Modelo Vectorial

Cálculo de la similitud → *producto escalar de vectores*

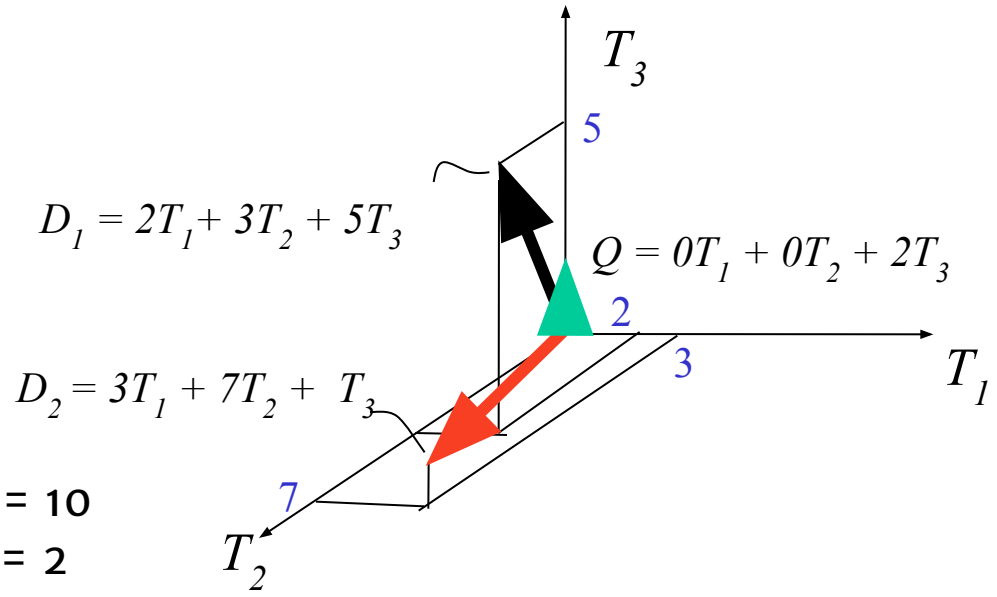
$$\text{sim}(\boldsymbol{d}_j,\boldsymbol{q}) = \boldsymbol{d}_j \cdot \boldsymbol{q} = \sum_{i=1}^{N} \text{wij} \cdot \text{wiq}$$

donde w_{ij} es el peso del término *i* en el documento d_j y w_{iq} es el peso del término *i* en la consulta *q*

Modelo Vectorial

En el ejemplo que teníamos antes …

$D_1 = (2, 3, 5)$
$D_2 = (3, 7, 1)$
$Q = (0, 0, 2)$



$sim(D_1,Q) = 2*0 + 3*0 + 5*2 = 10$
$sim(D_2,Q) = 3*0 + 7*0 + 1*2 = 2$

De donde, D_1 es "*más similar*" a Q qué D_2

Modelo Vectorial

Algunas ventajas ...

- Es simple y tiene una base matemática
- Mejora en la recuperación gracias a la asignación de pesos a los términos
- Considera tanto ocurrencias locales (*tf*) como globales (*idf*) de los términos
- Un documento puede ser recuperado con una coincidencia parcial
- En lugar de predecir si un documento es o no relevante, proporciona un *grado* de relevancia
- Resultados ordenados por "relevancia"
- Empíricamente adecuado

Modelo Vectorial

Ejemplos

D1: “Un perro ataca a un niño con un virus”

D2: “Virus ataca al perro de un niño”

Estos documentos tendrán 100% de similitud

D3: “El banco de la plaza está pintado de rojo”

D4: “En rojo quedó la plaza luego que el banco quebró”

….. y estos también !!

Modelo Vectorial

Algunos problemas...

- Carece de información sintáctica:
  - orden de las palabras, las considera independientes
  - proximidad
  - relaciones entre términos, etc.
- Existencia de la ambigüedad semántica
- Ignora sinonimia

Modelo Probabilístico

- Nuevo paradigma teórico para los SRI en los ‘70
- Presupone la existencia de exactamente un conjunto *ideal* de documentos *relevantes* para una consulta dada, así como una *consulta ideal*
- Lo crítico es:
  - determinar inicialmente las propiedades que marcan la relevancia del conjunto “ideal” de recuperación, ya que no están definidas
  - lograr una aproximación al conjunto por probabilidades a partir de una suposición inicial de las propiedades
  - ir refinando interactuando con el usuario luego de la consulta

Modelo Probabilístico

- Definición de probabilidad:
  - la probabilidad ***p*** de aparición de un suceso ***S*** de un total de ***n*** casos posibles igualmente factibles es el cociente entre el número de ocurrencias ***h*** de dicho suceso (casos favorables) y el número total de casos posibles ***n***

$$p = P\{S\} = h/n$$

- Idea general:
  - obtener el conjunto de documentos relevantes que maximiza la probabilidad de que los documentos que contenga sean relevantes para la consulta formulada por el usuario

Modelo Probabilístico

Objetivo:

- reformular la consulta sucesivamente usando ponderación de términos

Problemas:

- dada una consulta ***q*** y un documento ***dj*** de la colección, el modelo probabilístico estima la probabilidad de que el documento ***dj*** sea relevante para el usuario

- el modelo asume que esta probabilidad de relevancia depende sólo del documento y la consulta

- asume que hay un subconjunto **R** de documentos que el usuario prefiere como respuesta para la consulta ***q***

Simplificaciones:

- Los pesos son binarios {0,1} para todos los términos tanto de la consulta como de los documentos

Modelo Probabilístico

- Sea **R** el conjunto de documentos conocidos (o inicialmente supuestos) como relevantes y sea **R'** el complemento de **R**

- Sea $P(\mathbf{R}|d_j)$ la probabilidad de que el documento d_j sea relevante a la consulta q y $P(\mathbf{R'}|d_j)$ la probabilidad de que d_j no sea relevante a q

Entonces, la similitud del documento con la consulta se define como:

$$sim(d_j, q) = \frac{P(R \mid d_j)}{P(R' \mid d_j)} = \frac{\frac{P(d_j \mid R) \times P(R)}{P(d_j)}}{\frac{P(d_j \mid R') \times P(R')}{P(d_j)}} \approx \frac{P(d_j \mid R)}{P(d_j \mid R')}$$

Modelo Probabilístico

Ventajas...

- asignación de pesos a los términos, permitiendo recuperar los documentos que probablemente sean relevantes
- maneja retroalimentación del usuario (*relevance feedback*)
- capacidad para construir una función de ranking que ordene los documentos de manera decreciente según la probabilidad de ser relevantes a una consulta dada

Modelo Probabilístico

Contras...

- No toma en cuenta el factor ***tf*** (frecuencia del término)
- No tiene en cuenta todos los términos del documento
- Es poco intuitivo, con alta capacidad de cómputo y complejo de implementar
- Necesita una hipótesis inicial que no siempre es acertada de separar documentos relevantes de no relevantes
- Obtuvo pobres resultados

Modelos

Un problema común...

Consulta: *guerra fría*

Documento: *la crisis de los misiles cubanos*

El sistema no tiene idea que ambas cosas están relacionadas.

Solución → **análisis lingüístico**

Modelos

Nos van a ayudar cosas como...

- Lematización
- Etiquetar (POS)
- Detectar frases comunes
- Tesauros
- Relaciones entre términos
- Uso de la estructura del texto

Evaluación de la recuperación

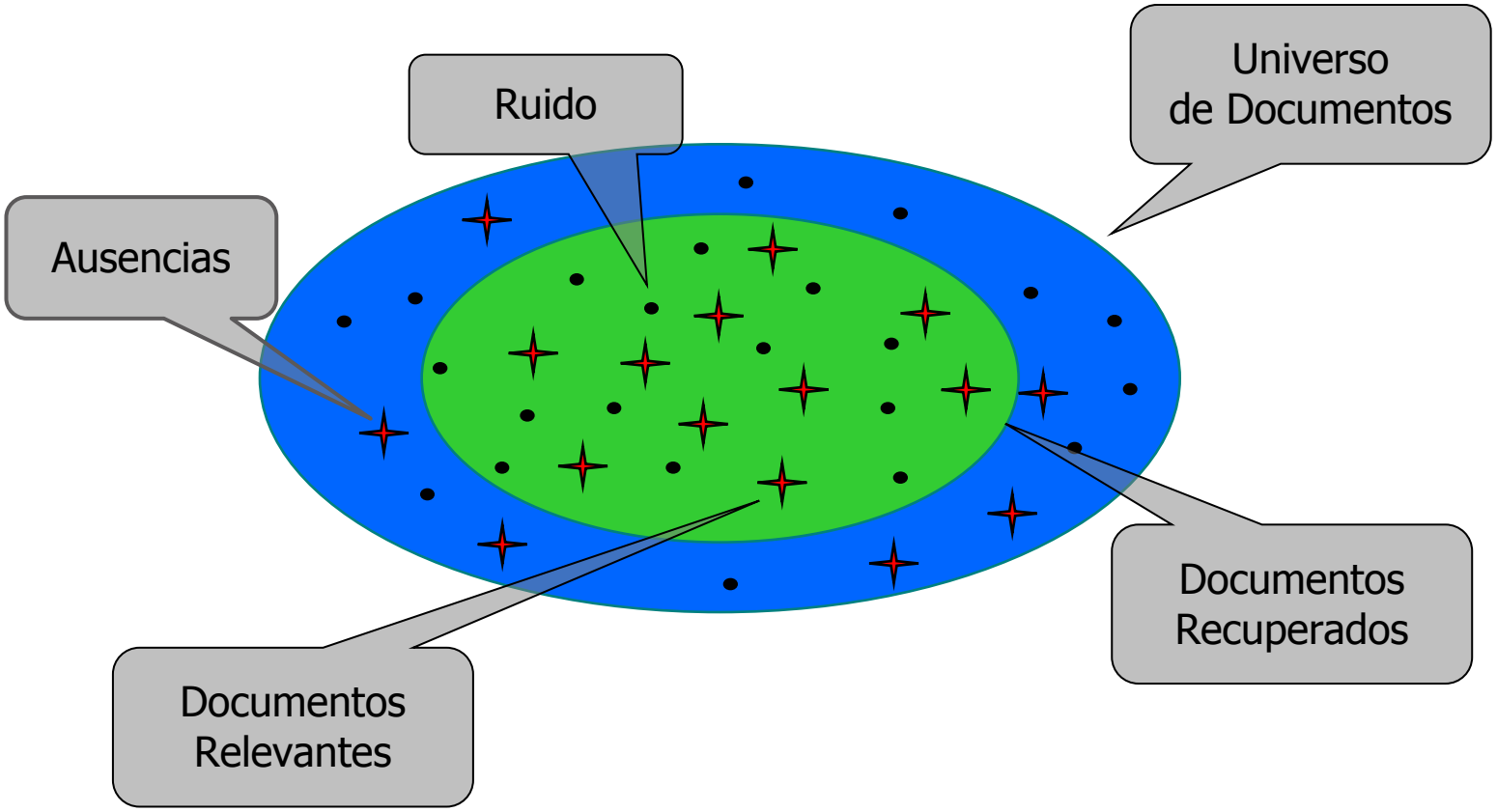
- ¿Cómo medimos lo bueno/malo de un sistema de RI?
- ¿Comparando documentos recuperados entre sí?
- ¿Cuánto hay que mirar en la lista de resultados para encontrar los resultados relevantes?
- ¿Qué costo tiene? ¿Es posible?
- Puede no haber acuerdo entre evaluadores
- Puede cambiar con el tiempo

Evaluación de la recuperación

- ¿Pero... qué significa ***relevancia*** ?
  - *Cualidad o condición de relevante, importancia, significación (RAE)*
  - *Relevante: Importante, destacado, significativo*
- En Recuperación de Información:

  *un documento es **relevante** cuando el contenido del mismo posee alguna importancia en relación con la consulta de un usuario*

Evaluación de la recuperación

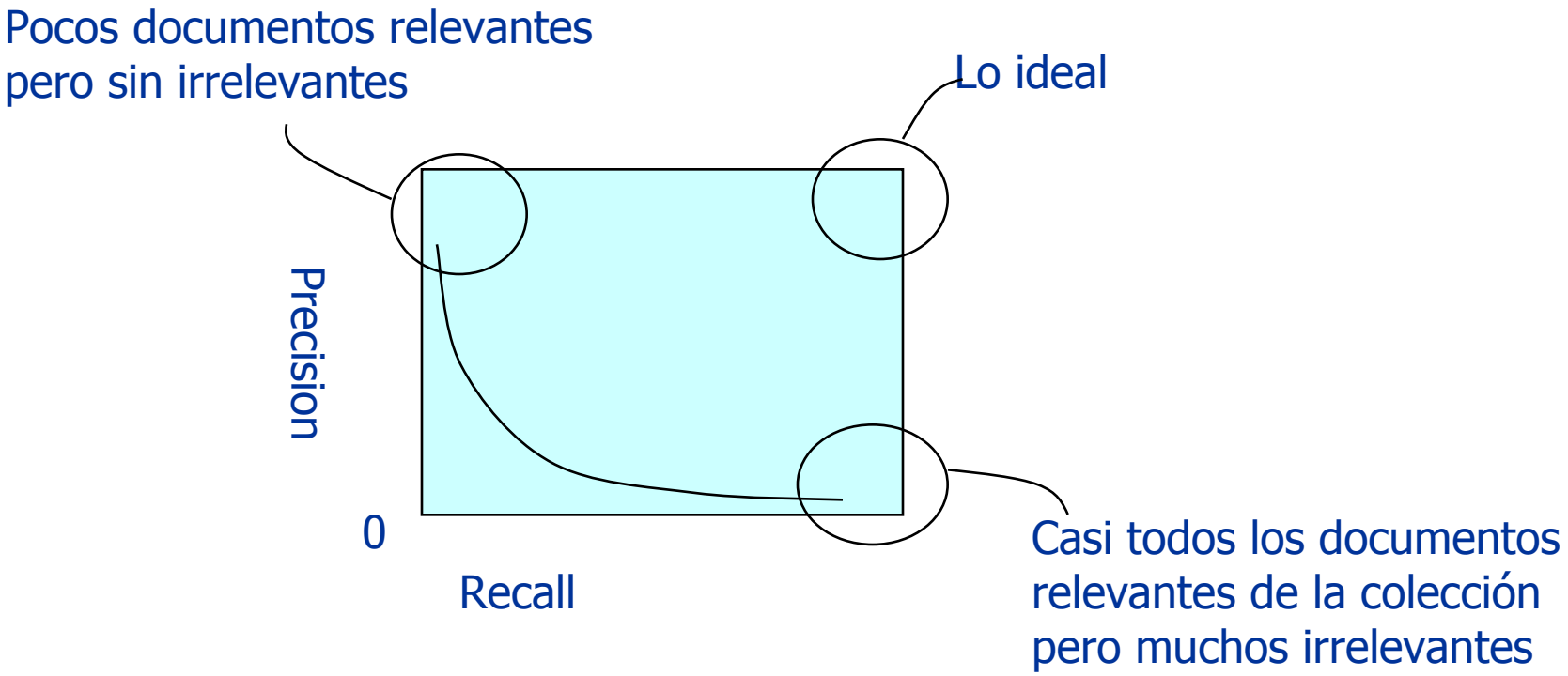


Evaluación de la recuperación

$$\text{Precision} = \frac{|\text{documentos recuperados relevantes}|}{|\text{documentos recuperados}|}$$

$$\text{Recall} = \frac{|\text{documentos recuperados relevantes}|}{|\text{documentos relevantes}|}$$

Evaluación de la recuperación



Evaluación de la recuperación

Combinación de medidas

- Medida ***F1*** = $\frac{2PR}{P+R}$

- Medida ***F*** = $\frac{(1+\beta^2)PR}{\beta^2 P+R}$

Evaluación de la recuperación

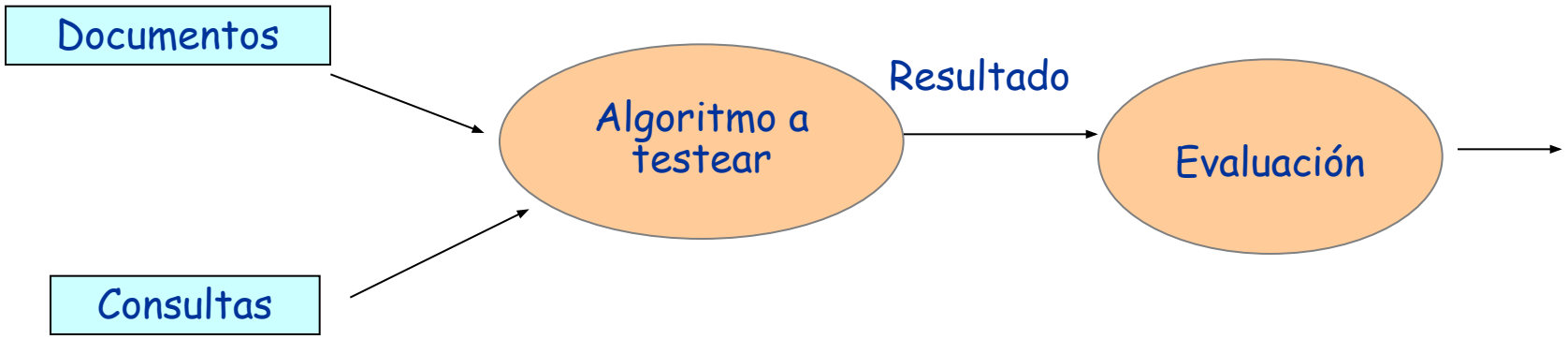
Entonces, para evaluar un SRI:

- La efectividad de un sistema se mide mediante una "colección de testeo"
- Se fija una colección de documentos y un conjunto de consultas
- Se realiza un etiquetado (humano) de cada documento según cada consulta
- En general, juicios binarios de relevancia (no se consideran grados de relevancia)

Evaluación de la recuperación

Una colección de testeo contiene:

- Un conjunto de documentos
- Un conjunto de consultas/búsquedas
- Un conjunto de documentos relevantes para cada consulta



Evaluación de la recuperación

Colecciones de testeo

Colección	Tema	Documentos	Preguntas
ADI	Ciencia de la Información	82	35
CACM	Informática	3200	64
CISI	Biblioteconomía	1460	76
CRAN	Aeronáutica	1400	225
MED	Medicina	1033	30
NLM	Medicina	3078	155
NPL	Ingeniería eléctrica	11429	100
TIME	Artículos generales	423	83

Evaluación de la recuperación

Conferencias TREC (Text REtrieval Conference)

- Co-patrocinadas por el NIST y DARPA
- Existen desde 1992
- Promueven la investigación dentro de la comunidad dando la infraestructura (documentos)
- Fomentan el intercambio entre academia, industria y gobierno

Indexación

- Una decisión importante es:

  ¿Qué términos como índices se va a usar para el esquema de indexación?

- Hipótesis:

  Tomar como término de índice, cada una de las palabras del documento

- ¿Problemas?

Indexación

Acciones previas:

- Eliminación de palabras vacías
- Transformaciones sobre el texto
- Lematización (Stemming)
- Análisis conceptual (elegir términos asociados a conceptos)

Indexación

Índices invertidos

- Estructura más simple
- Básicamente:
  - Vocabulario: conjunto de términos distintos del texto
  - Lista de *posteo* (ocurrencias): para cada término, la lista de documentos en que aparece

Indexación

Ejemplo

d1: Vendo autos y camionetas

d2: Autos usados

d3: Excelente oferta de camionetas

d4: Autos de segunda mano

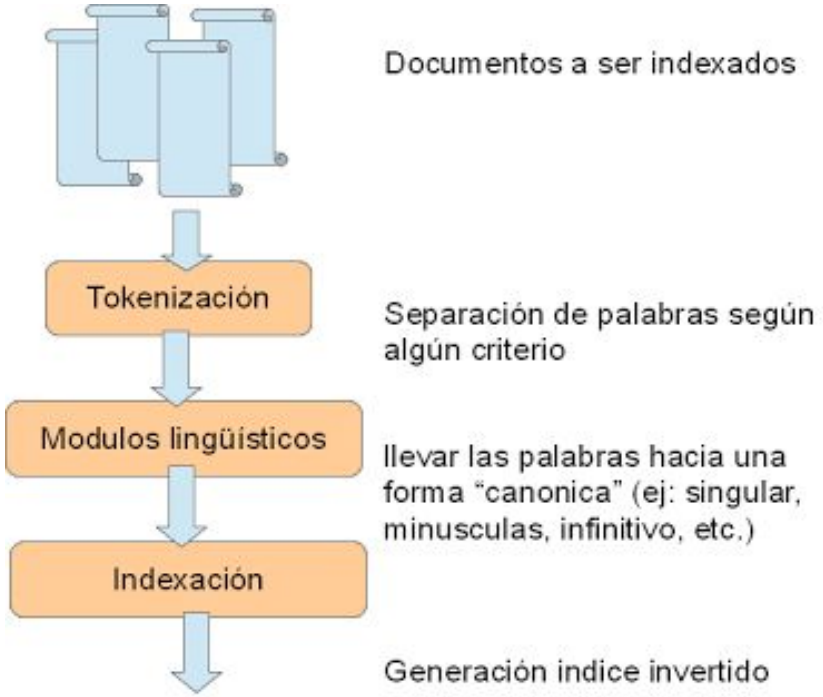
d5: Autos y camionetas de ocasión

d6: Permuto auto por camioneta

Vocabulario	Posteo
vend	d1
auto	d1,d2,d4,5,d6
camioneta	d1,d3,d5,d6
usad	d2
excelente	d3
oferta	d3
segund	d4
mano	d4
ocasión	d5
permut	d6

Indexación

Esquema general de construcción de un índice invertido



CLIR

- ¿qué es CLIR?

  Cross **L**anguage **I**nformation **R**etrieval

- ¿qué significa?

  todos los documentos relevantes para una pregunta deben ser encontrados, sin importar los idiomas particulares de cada documento y el de la consulta

CLIR

- ¿por qué surge?
  - hay mucha información en diversos idiomas
  - hay interés en recuperar la mayor cantidad de documentos relevantes
- ¿para qué sirve?
  - traducción de documentos
  - acceder a mayor cantidad de información

CLIR

- La consulta puede estar escrita en una lengua distinta a la de los documentos

- ¿qué se traduce: la consulta o los documentos?

  Necesitamos:

  - diccionarios
  - corpus paralelos
  - analizadores morfológicos y pos-taggers
  - tesauros
  - ontologías

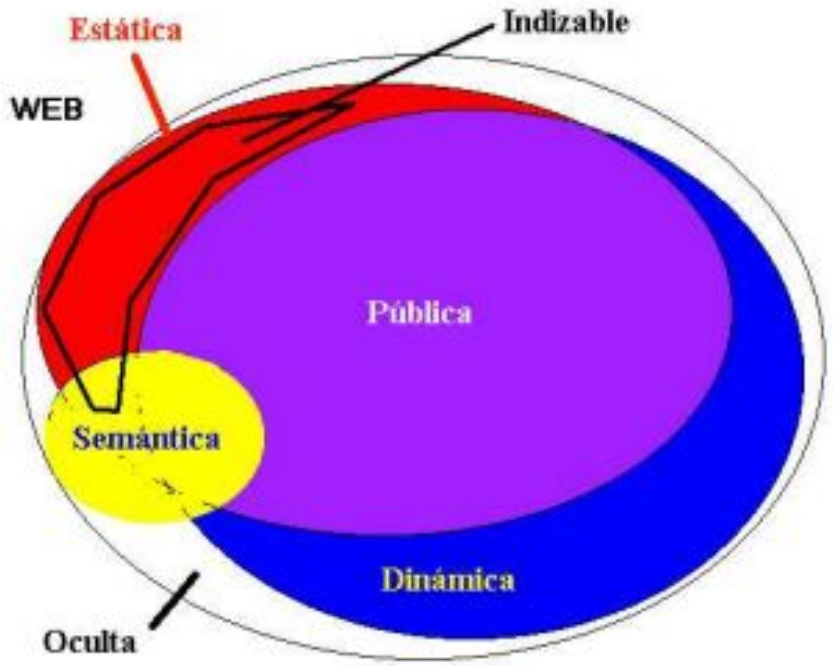
RI en la Web

- más de 1000: de sitios web (cifras a 2016)
- más de 50: se agregan por año
- más de 200: de blogs (cifras a 2015)
- más de 2:000.000: de archivos en Drive (cifras a 2017)

Determinar la relevancia de un documento era el problema central en los SRI tradicionales, en el contexto de la web es más que difícil

1997 uso de los "clicks" como señal de relevancia
1998 ***PageRank*** algoritmo basado en los enlaces

RI en la Web


Estática
Indizable
WEB
Pública
Semántica
Dinámica
Oculta

RI en la Web

RI tradicional	RI en la Web
• documentos similares	• documentos diferentes
• control terminológico	• carencia de control terminológico
• interfaz homogénea de consulta	• interfaces diferentes
• conjunto de documentos estable	• documentos cambian de forma y lugar
• volumen de información **grande**	• volumen de información **enorme**

RI en la Web

Desafíos para los buscadores:

- calidad de la relevancia en los resultados
- eficiencia en la búsqueda
- información mal estructurada y redundante
- contenido muy volátil
- conocer la intención del usuario al hacer la búsqueda (consulta)
- ver a la web no como conjunto de páginas sino como conjunto de objetos con atributos

RI en la Web

Estadísticas de la forma de consultar:

- 80% de las consultas no utilizan operadores
- 25% utiliza una sola palabra y en promedio 2
- 80% no utiliza *relevance feedback*
- 85% mira las primeras 2 páginas de resultados
- en promedio se navegan 2 páginas

RI en la Web

Buscar en la web NO es un problema de recuperar documentos, sino que es un mecanismo para mediar entre las personas y las necesidades que se encuentran detrás de sus objetivos Ricardo Baeza-Yates (2011)

Bibliografía

- Ricardo Baeza-Yates y Berthier Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley Longman Limited, ISBN:0-201-39829-X, 1999
- Salton, G. y McGill, M. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1986, ISBN: 0070544840
- Sparck Jones, K. (1972). *A statistical interpretation of term specificity and its application in retrieval*. Journal of Documentation, 28(1), 11–21
- Salton, G. (1971). T*he SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall