

Repositorios de Información

Recuperación de información

Daniel Gayo Avello

¿Qué vamos a ver?

1. Introducción
2. Indexado y TF-IDF
3. Evaluación
4. Modelos de recuperación de información:
 - Modelo booleano, vectorial, probabilístico
 - PageRank
5. Conclusiones

Introducción

- Terminología
- Tareas IR
- Modelo conceptual de la recuperación de información
- Documentos y representación de documentos
- Consultas
- *Best-match retrieval*
- Historia
- Temas en recuperación de información
- Recuperación de información vs Extracción de información vs Búsquedas Web

Terminología

- **General:** recuperación de información, necesidad de información, consulta, modelo de recuperación, motor de búsqueda, buscador, relevancia, evaluación, búsqueda de información, comunicación persona-máquina, navegación, interfaces, búsquedas *ad hoc*, filtrado
- **Experta:** frecuencia de términos, frecuencia de documentos, *inverse document frequency*, modelo vectoral, modelo probabilístico, BM25, PageRank, estematización, precisión, exhaustividad

Terminología

Necesidad de información

Ejemplo de una necesidad de información:

Encontrar documentos en los que se hable sobre la censura y la libertad de expresión en Internet. Los documentos en los que se discutan asuntos como la pornografía o el racismo en Internet, sin mencionar el tema de la censura o libertad de expresión, no se considerarán relevantes.

Una necesidad de información debe trasladarse a una consulta

Ejemplo de una posible consulta:

libertad expresión internet

Terminología

Recuperación de información (definición informal)

Estudio de sistemas automáticos que permitan a un usuario determinar la existencia o inexistencia de documentos (esto es, textos) relativos a una necesidad de información formulada habitualmente como una consulta.

Posibles objetivos de un sistema de RI:

- Exhaustividad (*recall*): recuperar todos los documentos relevantes
- Precisión: recuperar los documentos más relevantes
- Balance entre P y R:
 - Recuperar tan pocos documentos no relevantes como sea posible
 - Recuperar los documentos relevantes antes de los no relevantes

Terminología

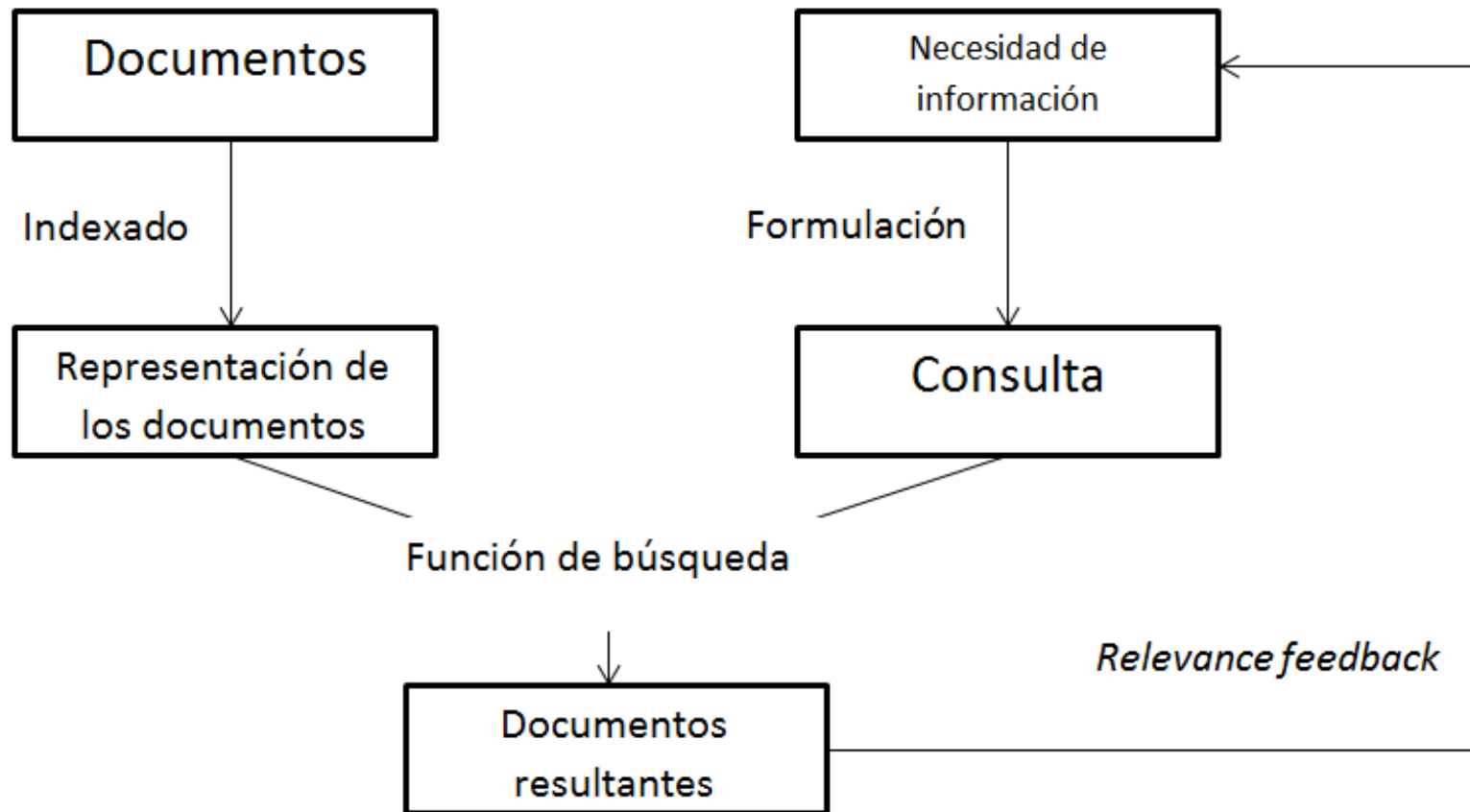
Recuperación de información vs Recuperación de datos

	Recuperación de información	Recuperación de datos
Matching	Difuso	Exacto
Modelo	Probabilista	Determinista
Lenguaje de consulta	Natural	Artificial
Especificación de la consulta	Incompleta	Completa
Objetos buscados	Relevantes	Todos los que hagan <i>matching</i>

Tareas de RI

- Búsquedas *ad hoc* (consultas de usuario)
- Filtrado de documentos
- Categorización de documentos
- Agrupamiento de documentos
- Búsqueda mediante exploración

Modelo conceptual de RI



Documentos y representación de documentos

- Unidad básica de trabajo
- Un pasaje de texto libre
 - Compuesto de texto, es decir, cadenas de caracteres de un alfabeto
 - Escrito en lenguaje natural: artículos periodísticos, artículos académicos, posts, tuits, correos electrónicos, ...
 - Tamaño de los documentos: arbitrario (pueden ser muy largos o muy cortos)

Documentos y representación de documentos

- Representación de texto libre: extraída directamente del texto, buen rendimiento en muchos contextos.
 - Representación con vocabulario controlado: más concisa, buen rendimiento en contextos muy especializados
-
- Representación de texto completo: la opción más deseable, requiere muchos recursos
 - Representación parcial (reducida): eliminar *stopwords*, lematizar términos, ...
-
- Representación estructurada: aprovechar la organización del texto (p.ej. en capítulos, secciones y párrafos)

Consultas

- Son la "traducción" de una necesidad de información
- Consultas simples: dos o tres keywords, quizás una docena.
Típicas en búsquedas web
- Consultas booleanas: "redes neuronales" AND "reconocimiento del habla". Típicas en búsquedas en catálogos
- Consultas contextuales: búsqueda de frases, con comodines, o por proximidad

Habitualmente podemos combinar todos los tipos anteriores en consultas más ricas

Best-match retrieval

- Se comparan los términos en documento y consulta
- Se calcula la "similitud" entre cada documento de la colección y la consulta en base a los términos que tienen en común
- Se ordenan los documentos por similitud (con la consulta) decreciente
- El resultado es una lista ordenada de documentos que se ofrece al usuario. Los primeros son más relevantes (según el sistema)

Historia (hasta Google)

- 1950s

- Primera descripción de un sistema IR automático. Utilización de la frecuencia de aparición de un término para determinar su relevancia, uso de stoplists. Luhn, H.P. 1957, “A Statistical Approach to Mechanized Encoding and Searching Information”, IBM Journal of Research and Development, vol. 1, no. 4, pp. 309-317.
- Primera propuesta para un sistema de resumen automático. Luhn, H.P. 1958, “The Automatic Creation of Literature Abstracts”, IBM Journal of Research and Development, vol. 2, no. 2, pp. 159-165.

Historia (hasta Google)

- 1960s

- Primera alternativa “aritmética” a la búsqueda booleana. Maron, M.E. y Kuhns, K.L. 1960, “On relevance, probabilistic indexing and information retrieval”, Journal of the ACM, vol. 7, no. 3, pp. 216-244.
- Primer esfuerzo para la evaluación experimental de sistemas IR. Cleverdon, C.W. 1962, Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems, College of Aeronautics, Reino Unido.
- Se propone el modelo vectorial de documentos y medida coseno de similitud. Salton, G. y Lesk, M.E. 1965, “The SMART Automatic Document Retrieval System – An Illustration”, Communications of the ACM, vol. 8, no. 6, pp. 391-398.

Historia (hasta Google)

- 1970s

- Se propone la cluster hypothesis, documentos estrechamente asociados tienden a ser relevantes para las mismas peticiones. Jardine, N. y van Rijsbergen, C.J. 1971, “The use of hierarchic clustering in information retrieval”, Information Storage and Retrieval, vol. 7, pp. 217-240.
- Introducción del concepto idf (inverse document frequency). Spärck-Jones, K. 1972, “A statistical interpretation of term specificity and its application in retrieval”, Journal of Documentation, vol. 28, no. 1, pp. 11-21.
- Se propone el modelo probabilista de IR. Robertson, S.E. y Spärck-Jones, K. 1976, “Relevance weighting of search terms”, Journal of the ASIS, vol. 27, no. 3, pp. 129-146.
- Por primera vez se señala la naturaleza interactiva de los sistemas IR. Swanson, D.R. 1977, “Information retrieval as a trial-and-error process”, Library Quarterly, vol. 47, no. 2.
- Primera colección moderadamente grande, NPL (11.500 documentos). SpärckJones, K. y Webster, C.A. 1979, Research in Relevance Weighting, Informe técnico, University of Cambridge.

Historia (hasta Google)

- 1980s

- Se inventa el primer algoritmo de stemming. Porter, M.F. 1980, “An algorithm for suffix stripping”, Program, vol. 14, no. 3, pp. 130-137.
- Probabilidad de coincidencia entre dos individuos en el uso de la misma palabra para identificar un concepto está entre el 10 y el 20%. Furnas, G.W., Landauer, T.K., Gómez, L.M. y Dumais, S.T. 1987, “The vocabulary problem in human system communication”, Communications of the ACM, vol. 30, no. 11, pp. 964-971.
- Se inventa la Semántica Latente. Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S. y Harshman, R. 1988, “Using Latent Semantic Analysis to improve access to textual information”, en Human Factors in Computing Systems, CHI’88 Conference Proceedings, pp. 281-285.
- Se inventa la Web. Berners-Lee, T. 1989, Information Management: A Proposal, Informe técnico, CERN.

Historia (hasta Google)

- 1990s

- Se desarrollan los primeros buscadores web... Koster, M. 1994, “ALIWEB – Archie-Like Indexing in the WEB”, Computer Networks and ISDN Systems, vol. 27, no. 2, pp. 175-182. Pinkerton, B. 1994, “Finding what people want: Experiences with the WebCrawler” Mauldin, M.L. y Leavitt, J.R.R. 1994, “Web Agent Related Research at the Center for Machine Translation”
- ...Y los primeros índices Filo, D. y Yang, J. 1994, Yahoo!
- Desarrollo de sistemas IR “tolerantes” por medio de n-gramas. Cavnar, W.B. 1994, “Using an n-gram-based document representation with a vector processing retrieval model”, en Proceedings of TREC-3, pp. 269-277.
- Primeros sistemas con pseudo-relevance feedback. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M. y Gatford, M. 1994, “Okapi at TREC-2”, en Text REtrieval Conference, pp. 21-34. Buckley, C., Salton, G., Allan, J. y Singhal, A. 1995, “Automatic Query Expansion Using SMART: TREC-3”, en Text REtrieval Conference, pp. 69-80.
- Primeros pasos hacia la Web Semántica. Luke, S., Spector, L. y Rager, D. 1996, “Ontology-Based Knowledge Discovery on the World-Wide Web”, en Working Notes of the Workshop on Internet-Based Information Systems at the 13th National Conference on Artificial Intelligence (AAAI96).
- **Se inventa Google**

Algunos temas en RI

- Modelos de recuperación (funciones de ranking, *learning to rank*, aprendizaje automático)
- Procesamiento de texto (para el indexado): PLN (modelos de lenguaje)
- CPM en relación con RI
- Eficiencia, compresión, escalabilidad
- Multimedia: imagen, video, sonido, música, habla
- Evaluación
- Búsquedas web, búsqueda en tiempo real, búsqueda en medios sociales
- RI multilingüe
- Búsqueda en documentos estructurados
- Bibliotecas digitales (legales, patentes, biosanitaria, etc.)

Recuperación de información vs Extracción de información

- Recuperación de información: dado un conjunto de términos y un conjunto de documentos selecciona los más relevantes y preferiblemente todos los relevantes.
- Extracción de información: dado un conjunto de documentos extrae del texto lo que "significan" los documentos (en realidad los esquemas de tablas implícitos en ese texto, [véase](#))

La RI puede encontrar documentos sin necesidad de
"entenderlos" en modo alguno

Recuperación de información vs Búsquedas en la Web

- El común de los mortales los considera términos intercambiables (**error**).
- La recuperación en la Web es muy distinta:
 - Cantidad de documentos es muchísimo mayor.
 - Mucha mayor heterogeneidad
 - Entorno adversarial
 - Es preciso explotar la estructura de hiperenlaces (p.ej. PageRank)
 - Puede explotarse el comportamiento agregado de los usuarios (búsquedas recomendadas, similares, *learning to rank*)

Indexado y TF-IDF

- Generación de representaciones de documentos
- Ponderación de términos
- Fichero invertido

Generación de representaciones de documentos

- Lenguaje de indexado
- Identificación de palabras
- Eliminación de palabras vacías (*stop words*)
- Situaciones especiales (fórmulas, cantidades, medidas, fechas, ...)
- Estematización
- Utilización de tesauros (para desambiguar términos)

Lenguaje de indexado

- Lenguaje que usa para representar los documentos y las consultas
 - Los términos de dicho lenguaje tienden a ser un subconjunto del vocabulario que aparece en los documentos
 - Puede derivarse del texto (lo más habitual) u obtenerse por otros medios
- Búsqueda por palabras clave (*keywords*):
 - análisis estadístico de los documentos en base a la frecuencia de aparición de los términos.
 - Automático, muy eficiente, pero no exento de problemas.
- Búsqueda usando vocabularios controlados:
 - más preciso pero requiere mucho tiempo si hay que indexar manualmente los documentos (ejemplos: catálogos bibliotecarios; ejemplos con vocabularios producidos por usuarios: flickr o delicious)

Identificación de palabras

- No es una tarea trivial, ni siquiera en idiomas con separadores de palabras (hay idiomas que no los tienen, p.ej. chino o japonés)
- Es preciso definir los separadores de palabras (generalmente los espacios en blanco)
- Es preciso ignorar la puntuación (p.ej. . , ; :)
 - Excepto cuando tenemos números, medidas y cantidades, p.ej. 3.141592, 17:30, 3.55€
- Hay que decidir qué hacer con guiones y subrayados, ¿son separadores? ¿Forman parte de la palabra? ¿Son un operador dentro de una fórmula?
- ¿Qué hacer con comillas y apóstrofes?
- ¿Qué hacer con los números?
- ¿Cómo trabajar con frases? P.ej. café con leche, Unión Europea, Universidad de Oviedo
- ¿Se mantiene la diferencia entre mayúsculas y minúsculas o se pasa todo a minúsculas?

Decisiones, decisiones...

Palabras vacías

- Se denominan stop words o palabras vacías aquellas palabras que, a pesar de un uso frecuente, aportan por sí solas poco significado a un texto
- Eliminarlas no siempre es una buena idea. *Riloff, E. 1995, “Little words can make a big difference for text classification”, en Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 130-136.*
- Además, ¿qué es una palabra vacía? Por ejemplo, ser
 - Verbo (palabra vacía)
 - Cadena SER (no es palabra vacía)
 - *SER Society for Ecological Restoration* (no es castellano)
- Listas de palabras vacías:
 - <http://snowball.tartarus.org/algorithms/english/stop.txt>
 - <http://snowball.tartarus.org/algorithms/spanish/stop.txt>

Detección de patrones especiales

- Pueden usarse expresiones regulares para detectar precios, teléfonos, URLs, direcciones de correo electrónico, fechas, horas, números de tarjeta de crédito, etc.
- Más difícil para extraer nombres propios de personas, lugares o empresas
- Otra opción: utilizar sistemas de reconocimiento de entidades (basados en un entrenamiento sobre material etiquetado). P.ej. <http://www.opencalais.com/> En Elasticsearch habría que usar plugins

Estematización

- Reducción de palabras a su raíz (que no su lema). P.ej **Universidad, universitarios, universitarias, universitaria, universitario.**
- Ventajas:
 - Reduce el número de términos que conforman el lenguaje de indexado
 - Aglutina términos que están relacionados semánticamente
- Problema principal:
 - Términos no relacionados entre sí pueden reducirse al mismo stem (p.ej. universidad y universo, libro y librar)
- [Más sobre estematización en Elasticsearch](#)

Uso de tesauros

- Permiten reducir la ambigüedad al poder asignar un sentido a un término. P.ej. banco o servidor son polisémicos
- Puede asociarse un término con el sentido apropiado en tiempo de indexado.
- Dificultades:
 - Desambiguar en tiempo de indexado para determinar cuál es el sentido apropiado
 - Desambiguar una consulta puede ser aún más difícil pues se carece de contexto. P.ej. apple
- De interés:
 - uso de sinónimos en Elasticsearch
 - English Wordnet, Wordnet del castellano

Ponderación de términos

- No todos los términos son iguales
 - de, que, no, a, la
 - estás, sé, tú, nada, nos
 - comer, lista, necesitas, creer, haga
 - colegas, leyenda, paraíso, limpieza, jeremy
 - embustero, mendoza, corregidos, sucesivamente, aceleración
 - grandotes, setos, tita, maniática, atraparía

¿Cuál crees que es la diferencia entre unos y otros?

Ponderación de términos

La frecuencia con que un término aparece en un documento es muy importante (Luhn, 1957): *term frequency*, TF

Ponderación de términos

Este documento parece que trata de pelé (sea lo que sea pelé)

W Pelé - Wikipedia

Es seguro | https://en.wikipedia.org/wiki/Pelé

Aplicaciones Blackboard Collabora WebSem ImpactStory: Daniel C 2011 Editorial Calend Escuela Ingeniería Inf Ministerio de Educaci redis - Project Hostin Otros marcadores


WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)
[Current events](#)
[Random article](#)
[Donate to Wikipedia](#)
[Wikipedia store](#)

[Interaction](#)
[Help](#)
[About Wikipedia](#)
[Community portal](#)
[Recent changes](#)
[Contact page](#)

[Tools](#)
[What links here](#)
[Related changes](#)
[Upload file](#)
[Special pages](#)
[Permanent link](#)
[Page information](#)
[Wikidata item](#)
[Cite this page](#)

[Print/export](#)
[Create a book](#)
[Download as PDF](#)
[Printable version](#)



Wiki Loves Monuments: Photograph a monument, help Wikipedia and win!

Pelé

From Wikipedia, the free encyclopedia

This article is about the retired Brazilian footballer. For other uses, see [Pelé](#) (disambiguation).

"O Rei" redirects here. For the Portuguese footballer of the same nickname, see [Eusébio](#).

This name uses [Portuguese naming customs](#). The first or maternal family name is Arantes and the second or paternal family name is Nascimento.


Edson Arantes do Nascimento (Brazilian Portuguese: [ˈɛtsõ (w)ɐˈɾẽtʃiz du nɐsiˈmẽtu]; born 23 October 1940), known as **Pelé** (Brazilian Portuguese: [pɐˈlɛ]), is a retired Brazilian professional footballer who played as a forward. He is widely regarded as the greatest football player of all time. In 1999, he was voted [World Player of the Century](#) by the [International Federation of Football History & Statistics](#) (IFFHS). That same year, [Pelé](#) was elected Athlete of the Century by the [International Olympic Committee](#). According to the IFFHS, [Pelé](#) is the most successful league goal-scorer in the world, scoring 1281 goals in 1363 games, which included unofficial friendlies and tour games. During his playing days, [Pelé](#) was for a period the best-paid athlete in the world.

[Pelé](#) began playing for [Santos](#) at age 15 and the [Brazil national football team](#) at 16. During his international career, he won three [FIFA World Cups](#): 1958, 1962 and 1970, being the only player ever to do so. [Pelé](#) is the [all-time leading goalscorer](#) for Brazil with 77 goals in 92 games. At club level he is also the record goalscorer for Santos, and led them to the 1962 and 1963 [Copa Libertadores](#). [Pelé](#)'s "electrifying play and penchant for spectacular goals" made him a star around the world, and his club team Santos toured internationally in order to take full advantage of his popularity. Since retiring in 1977, [Pelé](#) has been a worldwide ambassador for football and has made many acting and [commercial](#) ventures. In 2010, he was named the Honorary President of the [New York Cosmos](#).


[Pelé](#) has also been known for connecting the phrase "[The Beautiful Game](#)" with football. A prolific goalscorer, [Pelé](#) was known for anticipating his opponents' movements in the field, and being able to shoot strong and accurate shots with both feet. Early in his career, he played in a variety of [attacking formations](#). In his later career, he played in a [playmaking](#) role behind offensive strikers. In Brazil, he is hailed as a national hero for his accomplishments in football and for his outspoken support of policies that improve the social conditions of the poor. Throughout his career and in his retirement, [Pelé](#) received several individual and team awards for his

pelé

1 de 337



Pelé




Ponderación de términos

Este documento trata de pe l é mucho menos

W Diego Maradona - Wikip

Es seguro | https://en.wikipedia.org/wiki/Diego_Maradona

Aplicaciones Blackboard Collabora WebSem ImpactStory: Daniel C 2011 Editorial Calend Escuela Ingeniería Inf Ministerio de Educaci redis - Project Hostin Otros marcadores



WIKIPEDIA
The Free Encyclopedia

Main page

Contents

Featured content

Current events

Random article

Donate to Wikipedia

Wikipedia store

Interaction

Help

About Wikipedia

Community portal

Recent changes

Contact page

Tools

What links here

Related changes

Upload file

Special pages

Permanent link

Page information


Wikidata item

Cite this page


Article Talk

Read Edit View history

Search Wikipedia



Wiki Loves Monuments: Photograph a monument, help Wikipedia and win!



Diego Maradona

From Wikipedia, the free encyclopedia

"Maradona" redirects here. For other uses, see [Maradona \(disambiguation\)](#).


Diego Armando Maradona (Spanish pronunciation: [ˈdjeɣo maɾaˈðona], born 30 October 1960) is an Argentine retired professional [footballer](#). He has served as a manager and coach at other clubs as well as the national team of [Argentina](#). Many in the sport, including football writers, players, and fans, regard Maradona as the greatest football player of all time.^{[7][8][9][10]} He was joint FIFA [Player of the 20th Century](#) with [Pelé](#).^{[11][12]}

An advanced [playmaker](#) who operated in the [classic number 10 position](#), Maradona is the first player in football history to set the [world record transfer fee](#) twice, first when he transferred to [Barcelona](#) for a then world record £5 million, and second, when he transferred to [Napoli](#) for another record fee £6.9 million.^[13] He played for [Argentinos Juniors](#), [Boca Juniors](#), Barcelona, Napoli, [Sevilla](#) and [Newell's Old Boys](#) during his club career, and is most famous for his time at Napoli, where he won numerous accolades. In his international career with [Argentina](#), he earned 91 [caps](#) and scored 34 goals.

Maradona's vision, passing, ball control, [dribbling](#) skills, speed, reflexes and reaction time was combined with his small size (1.65 m or 5 ft 5 in tall) giving him a low center of gravity which allowed him to maneuver better than most other football players; he would often dribble past multiple opposing players on a run. His presence on the pitch had a great effect on his team's general performance, while he would often be singled out by the opposition. A precocious talent, Maradona was given the nickname "[El Pibe de Oro](#)" ("The Golden Boy"), a name that stuck with him throughout his career.^[14]

Maradona played in four [FIFA World Cups](#), including the [1986 World Cup](#) in Mexico where he captained Argentina and led them to

Diego Maradona



Ponderación de términos


Pero de lo que va verdaderamente este documento es de the

W Pelé - Wikipedia

Es seguro | https://en.wikipedia.org/wiki/Pelé

Aplicaciones Blackboard Collabora WebSem ImpactStory: Daniel C 2011 Editorial Calend Escuela Ingeniería Inf Ministerio de Educaci redis - Project Hostin Otros marcadores

the 2 de 608



WIKIPEDIA
The Free Encyclopedia

Main page

Contents

Featured content

Current events

Random article

Donate to Wikipedia

Wikipedia store

Interaction

Help

About Wikipedia

Community portal

Recent changes

Contact page

Tools

What links here

Related changes

Upload file

Special pages

Permanent link

Page information


Wikidata item

Cite this page


Article Talk

Read View source View history

Search Wikipedia



Wiki Loves Monuments: Photograph a monument, help Wikipedia and win!



Pelé

From Wikipedia, the free encyclopedia

This article is about the retired Brazilian footballer. For other uses, see *Pele* (disambiguation).


"O Rei" redirects here. For the Portuguese footballer of the same nickname, see *Eusébio*.

This name uses Portuguese naming customs. The first or maternal family name is Arantes and the second or paternal family name is Nascimento.

Edson Arantes do Nascimento (Brazilian Portuguese: [ˈɛtsõ (w)ɐˈɾɐ̃tʃiz du nɐsiˈmẽtu]; born 23 October 1940), known as **Pelé** (Brazilian Portuguese: [pɛˈlɛ]), is a retired Brazilian professional footballer who played as a forward. He is widely regarded as the greatest football player of all time. In 1999, he was voted World Player of the Century by the International Federation of Football History & Statistics (IFFHS). That same year, Pelé was elected Athlete of the Century by the International Olympic Committee. According to the IFFHS, Pelé is the most successful league goal-scorer in the world, scoring 1281 goals in 1363 games, which included unofficial friendlies and tour games. During his playing days, Pelé was for a period the best-paid athlete in the world.

Pelé began playing for Santos at age 15 and the Brazil national football team at 16. During his international career, he won three FIFA World Cups: 1958, 1962 and 1970, being the only player ever to do so. Pelé is the all-time leading goalscorer for Brazil with 77 goals in 92 games. At club level he is also the record goalscorer for Santos, and led them to the 1962 and 1963 Copa Libertadores. Pelé's "electrifying play and penchant for spectacular goals" made him a star around the world, and his club team Santos toured internationally in order to take full advantage of his popularity. Since retiring in 1977, Pelé has been a worldwide ambassador for football and has made many acting and commercial ventures. In 2010, he was named the Honorary President of the New York Cosmos.

Pelé



Ponderación de términos

La frecuencia por sí sola no es suficiente

- Posible solución: eliminar palabras vacías.
- Pero sabemos que es una **mala** solución

Ponderación de términos

- Por otro lado, el hecho de que un término aparezca mucho puede ser poco significativo.
- Ejemplos:
 - `marca.com`: fútbol, real madrid, cristiano ronaldo, ...
 - `sport.es`: fútbol, barça, messi, ...
 - `abc.es`: españa, cataluña, ...
 - `boe.es`: decreto, disposición, ...

¿Qué ocurre con esos términos en sus respectivos sitios web?

Ponderación de términos

- El número de documentos de la colección en que aparece un término es muy importante (a la inversa) Karen Spärck-Jones (1972): inverse document frequency (idf)
 - A mayor número de documentos que contienen un término menor es la importancia del mismo.
 - A menor número de documentos que contienen el término mayor es su importancia.

$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Ponderación de términos

La forma más básica de ponderación de términos es $tf \cdot idf$

$weight(t, d) = tf(t, d) \times idf(t)$	
N	number of documents in collection
$n(t)$	number of documents in which term t occurs
$idf(t)$	inverse document frequency of term t
$occ(t, d)$	occurrence of term t in document d
t_{max}	term in document d with highest occurrence
$tf(t, d)$	term frequency of t in document d

Ponderación de términos

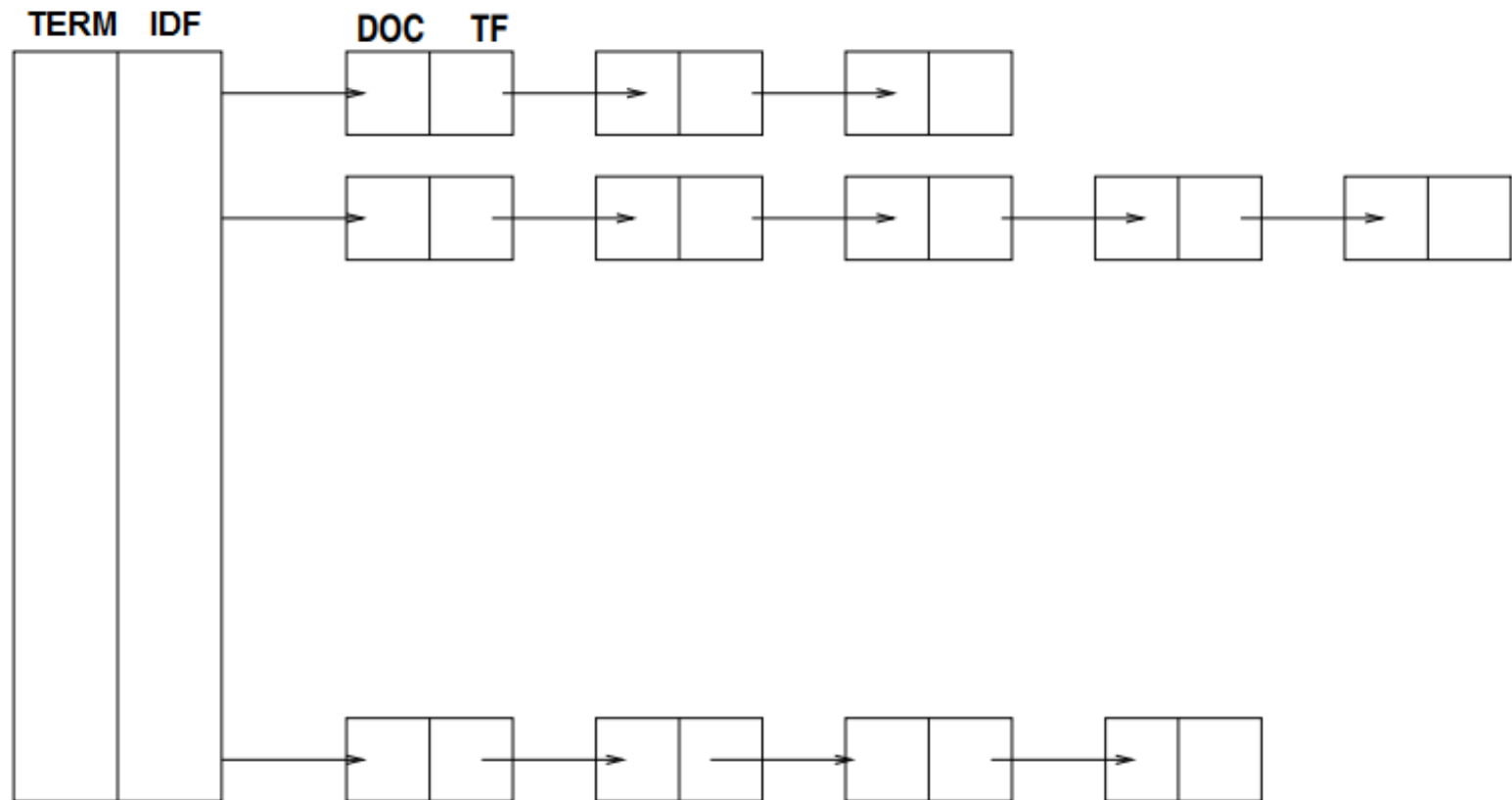
*¿Cuál es el peso de *pelé* y *the* en
<https://en.wikipedia.org/wiki/Pelé?>*

- N : 12.800.000 documentos
- $n(\text{pelé})$: 10.400 documentos
- $n(\text{the})$: 12.800.000 documentos
- $\text{idf}(\text{pelé})$: 3,09018
- $\text{idf}(\text{the})$: 0
- $\text{tf}(\text{pelé}, d)$: 337
- $\text{tf}(\text{the}, d)$: 705
- **$\text{tf-idf}(\text{pelé}, d)$: $337 * 3,09018 = 1041,3895$**
- **$\text{tf-idf}(\text{the}, d)$: $705 * 0 = 0$**

El fichero invertido

- Por razones obvias no es práctico comparar una consulta con **todos** los documentos de la colección
- Dada una consulta es preciso determinar:
 - qué documentos pueden satisfacerla potencialmente (**ocurrencias**)
 - determinar la **similitud** (p.ej. con tf-idf) de dichas ocurrencias con la consulta
 - para producir una **lista de resultados**

El fichero invertido



Documentos vs Fichero Invertido

	pelé	maradona	football
Pelé	337	12	41
Diego Maradona	15	381	37
Association football	0	0	143
FIFA World Cup	3	0	70

Documentos vs Fichero Invertido

	Pelé	Diego Maradona	Association football	FIFA World Cup
pelé	337	15	0	3
maradona	12	381	0	0
football	41	37	143	70

¿Qué información se almacena en el fichero invertido?

- Búsqueda booleana: el identificador del documento
- Búsqueda ordenada: el identificador del documento e información para calcular ponderación de términos (tf, idf, tf-idf, ...)
- Operadores de proximidad: además de lo anterior, los *offsets* de cada palabra dentro del documento

¿Cómo de grande es el fichero invertido?

- Más pequeño de lo que te imaginas
- Excepto si admite búsquedas por proximidad o frases, entonces es ¡enorme! Mayor incluso que la colección de documentos
- A cambio las búsquedas son muy rápidas

Evaluación

- ¿Qué evaluar?
- Colecciones de prueba
- Precisión y exhaustividad

¿Qué podemos evaluar?

- Cobertura de la colección: muy importante históricamente en los buscadores Web, en la actualidad relevante al crear colecciones de medios sociales (p.ej. Twitter), también relevante en buscadores de artículos académicos.
- Eficiencia: tiempo de indexado, tiempo de respuesta, uso de memoria, uso de disco, etc.
- **Precisión:** porcentaje de resultados que son realmente relevantes.
- **Exhaustividad:** porcentaje de documentos relevantes que aparecen en los resultados.
- Aspectos de usabilidad y UX

Colecciones de prueba

- La recuperación de información es un campo empírico; es preciso experimentar para justificar la superioridad de una técnica sobre otra.
- Elementos necesarios para evaluar un sistema IR:
 - Una **colección** de documentos.
 - Una lista de **necesidades de información** expresables como consultas.
 - Un conjunto de **juicios de relevancia** para cada par (documento, necesidad de información).

Un documento de la colección

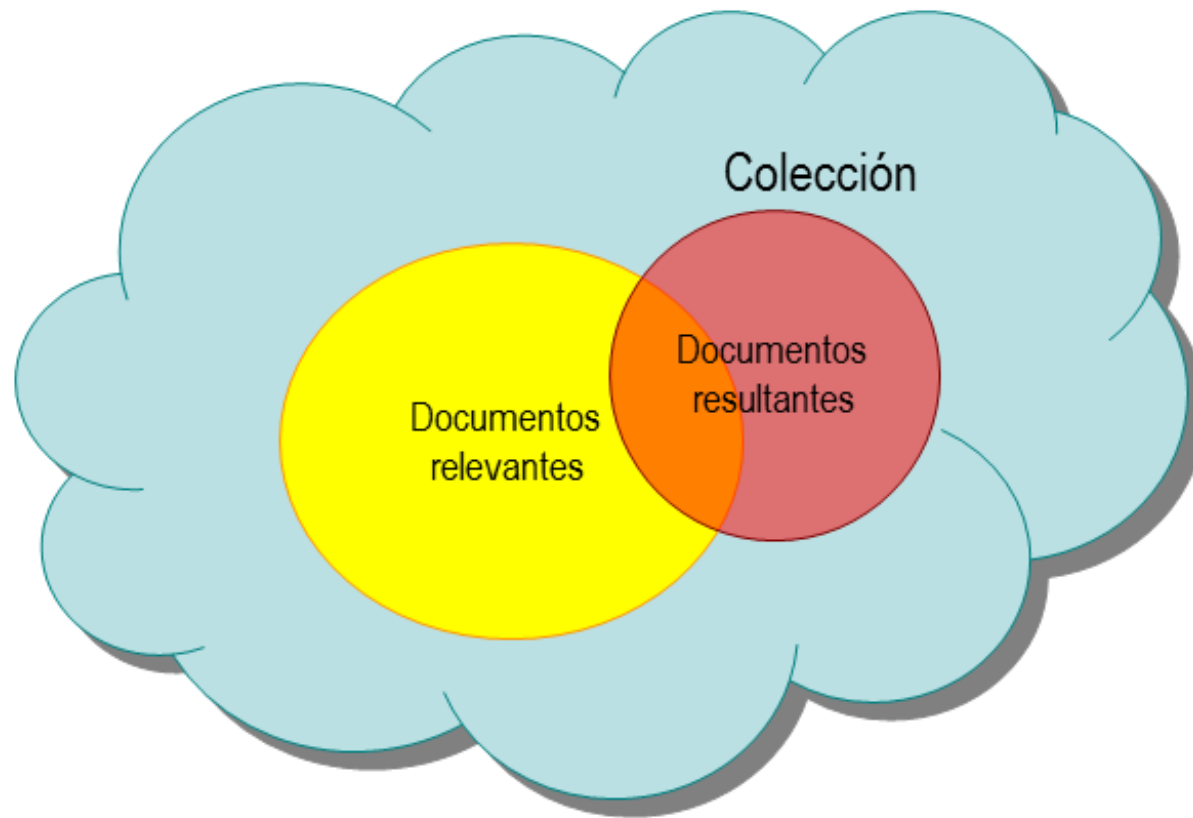
Reuters-21578

```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="5!  
  <DATE>26-FEB-1987 15:14:36.41</DATE>  
  <TOPICS><D>veg-oil</D> ... <D>wheat</D></TOPICS>  
  <PLACES><D>argentina</D></PLACES>  
  <PEOPLE></PEOPLE>  
  <ORGS></ORGS>  
  <EXCHANGES></EXCHANGES>  
  <COMPANIES></COMPANIES>  
  <UNKNOWN>  
    G f0754 reuter f BC-ARGENTINE-1986/87-GRA 02-26 0066  
  </UNKNOWN>  
  <TEXT>  
    <TITLE>  
      ARGENTINE 1986/87 GRAIN/OILSEED REGISTRATIONS  
    </TITLE>  
    <DATELINE>  
      BUENOS AIRES, Feb 26 -  
    </DATELINE>
```

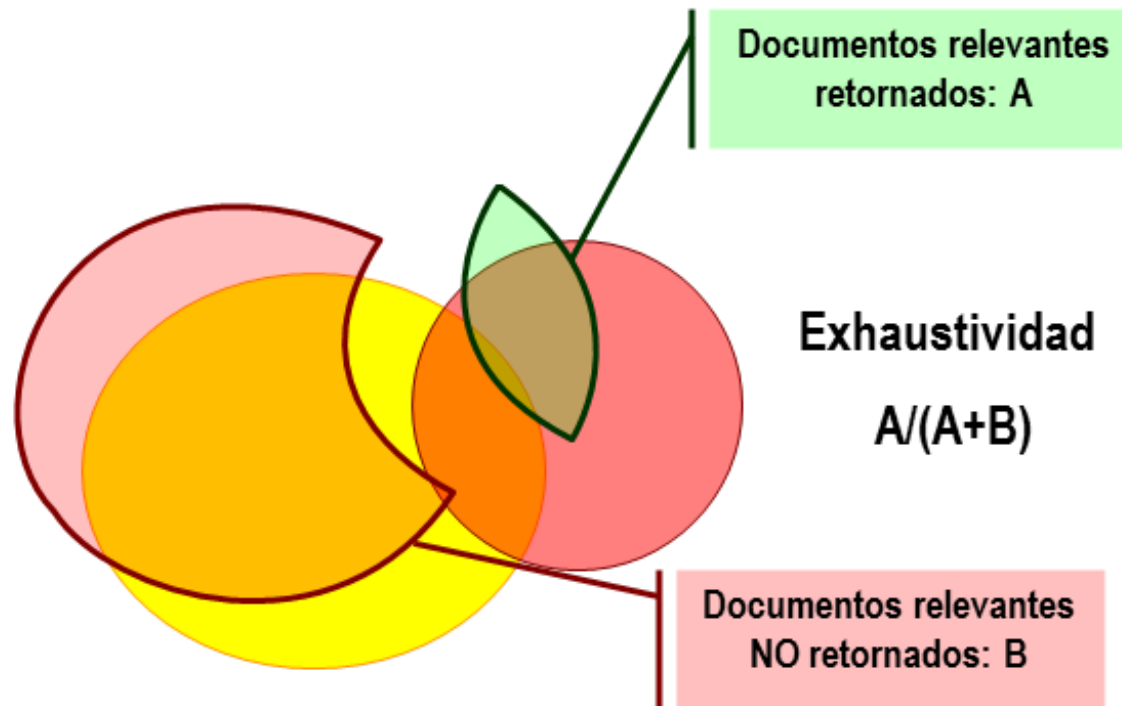
Un "tópico" del CLEF (que no una consulta)

```
<top>
  <num>
    C154
  </num>
  <ES-title>
    Libertad de Expresión en Internet
  </ES-title>
  <ES-desc>
    Encontrar documentos en los que se hable sobre la censura y la libertad
  </ES-desc>
  <ES-narr>
    Los documentos en los que se discutan asuntos como la pornografía o e
  </ES-narr>
</top>
```

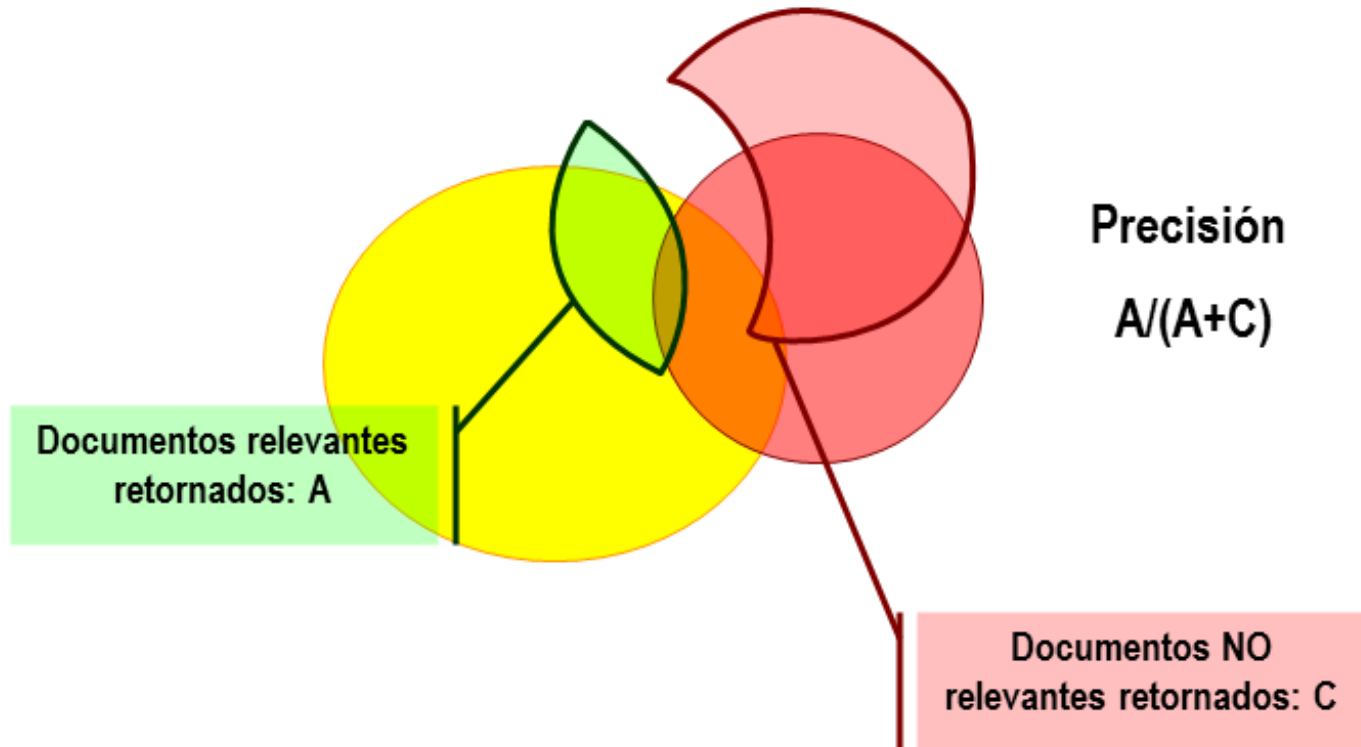
Precisión (precision) y exhaustividad (recall)



Precisión (precision) y exhaustividad (recall)

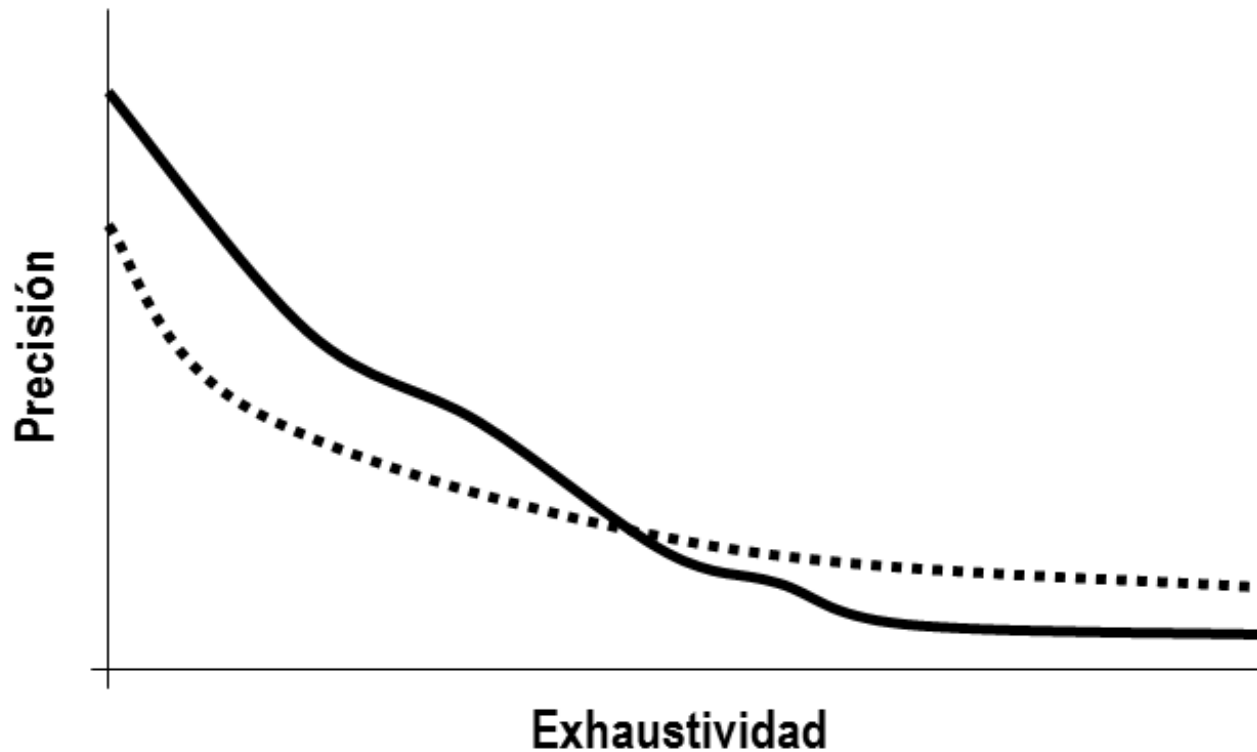


Precisión (precision) y exhaustividad (recall)



¡Atención! Pregunta

¿Qué sistema es mejor? ¿Por qué?



Modelos de recuperación

- Componentes de un modelo RI
- Principales modelos:
 - Booleano
 - Basados en conjuntos
 - Vectorial
 - Probabilístico (p.ej. BM25)
 - PageRank

Componentes de un modelo RI

- D es el conjunto de representaciones de documentos (a partir de ahora "**documentos**")
- Q es el conjunto de representaciones de necesidades de información (a partir de ahora "**consultas**")
- $R(d,q)$ es una función de ranking que:
 - asocia un número real (normalmente entre 0 y 1) con un documento d del conjunto D y una consulta q de Q
 - puede usarse para determinar una ordenación de los documentos de D con respecto a la consulta q
 - dicho ordenamiento se supone que refleja la relevancia

Componentes de un modelo RI

- Para cada modelo de recuperación de información es preciso especificar los siguientes componentes:
 1. La representación del documento d
 2. La representación de la consulta q
 3. La función de ranking $R(d,q)$

Modelo booleano

- Recupera documentos que hacen cierta la consulta
- La función de ranking sólo admite dos valores:
 - 1 si el documento satisface la consulta
 - 0 en caso contrario
- Las consultas son expresiones lógicas que combinan términos.
P.ej. (web AND semántica) OR websem OR ontología
- Los documentos también podrían ser expresiones lógicas pero generalmente se supone que son una combinación AND de los términos que los componen *modelo bag-of-words*, merece una explicación...
- La evaluación de la consulta se hace directamente sobre el fichero invertido.
- Características principales:
 - **No hay ranking** un documento aparece o no, no hay información sobre su relevancia
 - Muy difícil obtener un balance entre precisión y exhaustividad: o se obtienen muchísimos resultados (la mayor parte irrelevantes y sin orden) o muy pocos
- Ejemplo: [Buscador de la Biblioteca de UniOvi](#)

Modelos basados en conjuntos

- El modelo booleano "puro" es francamente malo, muy pronto se determinó que era preferible **representar documentos y consultas como conjuntos de términos**. Maron, M.E. y Kuhns, K.L. 1960, "On relevance, probabilistic indexing and information retrieval", Journal of the ACM, vol. 7, no. 3, pp. 216-244.
- La función de ranking se basa en la comparación del conjunto de términos en d con el conjunto de términos en q . Pueden usarse distintas funciones:

Coeficiente de Dice	$2 \frac{ X \cap Y }{ X + Y }$
---------------------	----------------------------------

Coeficiente de Jaccard	$\frac{ X \cap Y }{ X \cup Y }$
------------------------	---------------------------------

Coseno	$\frac{ X \cap Y }{ X \cdot Y }$
--------	------------------------------------

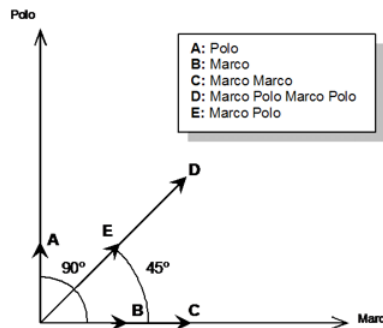
Coeficiente de solapamiento	$\frac{ X \cap Y }{\min(X , Y)}$
-----------------------------	-------------------------------------

- Nótese que los modelos basados en conjuntos no ponderan los términos (cosa que sabemos debería hacerse)

Modelo Vectorial

- Extensión de los modelos basados en conjuntos debida a Salton (1960s a 1980s)
- Tanto d como q son conjuntos de términos pero los términos están ponderados, al menos en d
- La ponderación puede hacerse de muy diversas formas: más habitual **tf-idf**
- Puede usarse cualquier función de ranking de las anteriores **pero** suele usarse la **similitud del coseno**

$$\frac{\sum_{i=1}^n q_i \cdot d_i}{\sqrt{\sum_{i=1}^n q_i^2} \cdot \sqrt{\sum_{i=1}^n d_i^2}}$$




Modelo probabilístico

- Debido a Robertson y Spärck-Jones. Robertson, S.E. y Spärck-Jones, K. 1976, “Relevance weighting of search terms”, Journal of the ASIS, vol. 27, no. 3, pp. 129-146.
- Trata de responder la siguiente pregunta:
Dada una consulta de usuario q y un documento d , ¿cuál es la probabilidad de que el usuario encuentre d relevante?
- En última instancia se basa en el **teorema de Bayes de probabilidad condicionada**
- Existen múltiples aproximaciones pero la más popular se denomina **Okapi BM25** (y la implementa Elasticsearch)

PageRank

Antes de hablar de PageRank...

- En 1990s la Recuperación de Información era un campo maduro pero la Web lo cambió todo.
- Entre 1990 y 1994 **no había buscadores** en la Web
- Entre 1994 y 1998 había buscadores pero eran 

Ver más abajo cómo funcionaban...

- Posibles razones:
 - La mayor colección de evaluación IR en 1998 tenía 7.5 millones de documentos
 - La cota inferior para el tamaño de la Web era de 320 millones de documentos
 - Las consultas de los usuarios eran muy cortas (3 términos o menos)
 - El spam de keywords en la Web era brutal

El mejor buscador Web en 1997

- Empleaba robots para explorar la Web en busca de documentos
- Almacenaba el texto completo de las páginas web además del texto de los enlaces entrantes
- No tenía en cuenta las palabras vacías en documentos ni en consultas
- Los términos podían ponderarse mediante $tf \cdot idf$
- Retornaba resultados ordenados por relevancia decreciente
- La relevancia se calculaba *ad hoc* teniendo en cuenta no sólo el peso de los términos según el modelo vectorial sino relativos a la proximidad entre los términos o aspectos de “formateo” (título, cabeceras, etc.)
- **Y no funcionaba bien...**

Hasta aquí hemos llegado...

- Brin, S. y Page, L. 1998, “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Computer Networks and ISDN Systems, vol. 30, no. 1-7, pp. 107-117.
- *"as of November 1997, only one of the top four commercial search engines finds itself."*
- *"[...]we have seen a major search engine return a page containing only "Bill Clinton Sucks" and picture from a "Bill Clinton" query. [...] If a user issues a query like "Bill Clinton" they should get reasonable results since there is a enormous amount of high quality information available on this topic. Given examples like these, we believe that the standard information retrieval work needs to be extended to deal effectively with the web."*

ANBURG'S BILL CLINTON AND OTHER PO

A major search engine result
for the query "BILL CLINTON"
(The Web cca. 1997)



Bill Clinton Sucks



Web IR no es lo mismo que IR

- La Web no es una colección de documentos
- La Web es un grafo
- Idea loca, ignoremos completamente el texto, centremos únicamente en el **grafo**. Marchiori, M. 1997 “The Quest for Correct Information on the Web: Hyper Search Engines”. The Sixth International WWW Conference (WWW 97).
 - *A great problem with search engines' scoring mechanisms is that they tend to score text more than hypertext.*
 - *[...] focusing separately on the "textual" and "hyper" components.*
 - *The presence of links in a Web object clearly augments the informative content with the information contained in the pointed Web objects. Recursively, links present in the pointed Web objects further contribute, and so on. Thus, in principle, the analysis of the informative content of a Web object A should involve all the Web objects that are reachable from it [...]*
 - *This is clearly unfeasible in practice, so, for practical reasons, we have to stop the analysis at a certain depth [...]*

PageRank

- Google comienza a operar en 1998

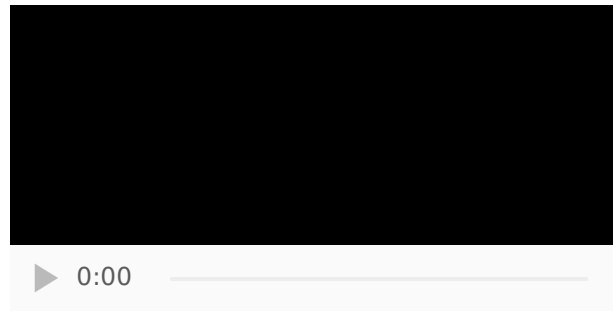
Brin, S. y Page, L. 1998, “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Computer Networks and ISDN Systems, vol. 30, no. 1-7, pp. 107-117.

- El núcleo de su sistema de ponderación es el algoritmo PageRank

Page, L., Brin, S., Motwani, R. y Winograd, T. 1998, The PageRank Citation Ranking: Bringing Order to the Web

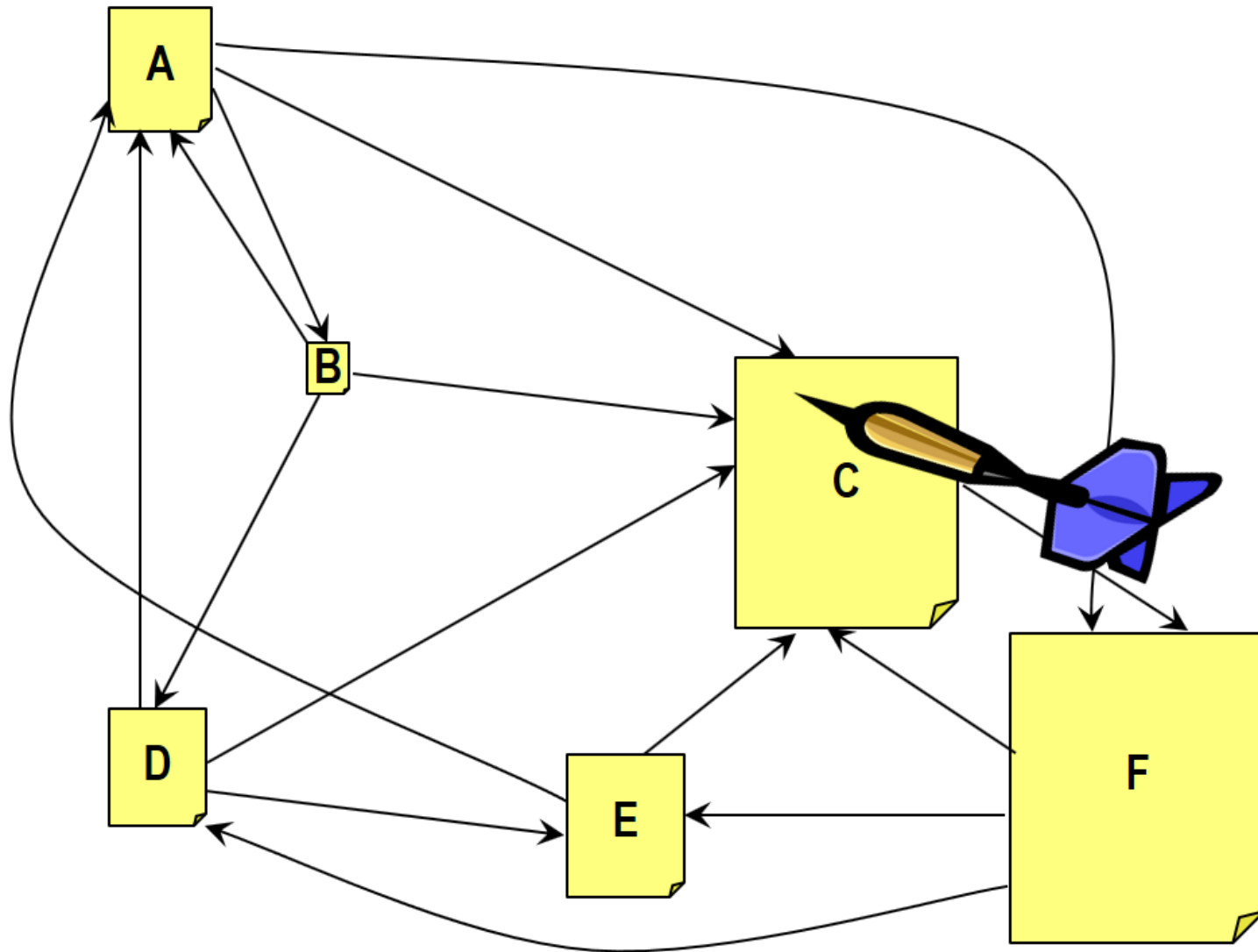
- El algoritmo asocia a cada documento un valor (tb. PageRank) de este modo:

- Un documento transmite a todos los documentos que enlaza su valor PageRank dividido por el número de enlaces salientes
- Un documento muy enlazado tendrá un PageRank elevado
- Un documento enlazado desde documentos prestigiosos tendrá un PageRank elevado



Algunas características interesantes de PageRank

- Los valores de PageRank calculados para los nodos se “**estabilizan**” con **rapidez** (p.ej. 52 iteraciones son suficientes para obtener valores razonables para 322 millones de enlaces)
- Es **relativamente insensible a los valores de “partida”**, afectaría al número de iteraciones necesarias y a los valores finales (obviamente) pero no al ranking obtenido
- El PageRank total en la Web es constante
- Si el valor inicial asignado a cada documento es $1/N$ (número de documentos) el valor de PageRank equivale a la probabilidad de que un usuario llegue a dicho documento siguiendo enlaces al azar (*random surfer model*)



Búsquedas en la Web con PageRank

- **Google no solo usa PageRank** pero PageRank supuso una ventaja competitiva enorme de Google sobre el resto de buscadores en 1998
- Recordemos lo que dijo Marchiori
[...] focusing separately on the "textual" and "hyper" components.
- PageRank no tiene en cuenta el contenido de los textos para determinar el prestigio/autoridad/relevancia de un nodo, sólo los enlaces
- **¿Cómo se realizan las búsquedas entonces?** (Versión simplificada)
 - Se extraen los términos (palabras) de la consulta
 - Se localizan documentos que contengan todos los términos
 - Se ordenan los documentos obtenidos por PageRank decreciente
- Es decir, **Google proporciona a los usuarios aquellos documentos que satisfacen la consulta y tienen más prestigio en la Web**

Software de interés para construir un buscador Web

- [Nutch](#)
- [Heritrix](#)
- [Xapian](#)

Y por supuesto Elasticsearch...

Conclusiones / Recapitulación

- Qué es recuperación de información
- TF-IDF
- Precisión y exhaustividad
- Modelos de recuperación de información: booleano, conjuntos, vectorial, probabilístico (BM25), PageRank
- Recuperación de información y recuperación de información en la Web son muy diferentes