



**XII Jornadas de  
Ciencias de la  
Computación**

# **Recuperación de Información de Gran Escala**

## **Conceptos y algoritmos detrás de los Motores de Búsqueda Web**

**Gabriel H. Tolosa**  
tolosoft@unlu.edu.ar



**48.000.000.000<sup>1</sup>**

<sup>1</sup><http://www.worldwidewebsize.com/>



**48.000.000.000<sup>1</sup>**

**0,25**

<sup>1</sup><http://www.worldwidewebsize.com/>



**48.000.000.000<sup>1</sup>**

**0,25**

Buscar en la Web

<sup>1</sup><http://www.worldwidewebsize.com/>

# Por ejemplo...

rosario



Web

Images

Videos

News

More ▾

Search tools

About 14,200,000 results (0.27 seconds)

## Sitio de la Municipalidad de Rosario

[www.rosario.gov.ar/](http://www.rosario.gov.ar/) ▾ [Translate this page](#)

**Rosario**, segunda ciudad de la República Argentina. Punto estratégico del Mercosur. Infórmese sobre su comercio exterior y posibles negocios. Conozca su ...

[Trámites](#) - [Registro Único de Postulantes](#) - [InfoMapa](#) - [Recorridos de colectivos](#)

## Rosario, Santa Fe - Wikipedia, the free encyclopedia

[en.wikipedia.org/wiki/Rosario,\\_Santa\\_Fe](http://en.wikipedia.org/wiki/Rosario,_Santa_Fe) ▾

**Rosario** is the largest city in the province of Santa Fe, in central Argentina. It is located 300 km (186 mi) northwest of Buenos Aires, on the western shore of the ...

[History](#) - [Government](#) - [Economy](#) - [Culture](#)

## Rosario (Argentina) - Wikipedia, la enciclopedia libre

[es.wikipedia.org/wiki/Rosario\\_\(Argentina\)](http://es.wikipedia.org/wiki/Rosario_(Argentina)) ▾ [Translate this page](#)

La ciudad de **Rosario** está ubicada en el centro-este argentino, en la provincia de Santa Fe. Es la tercera ciudad más poblada de Argentina después de Buenos ...

[Toponimia](#) - [Historia](#) - [Geografía](#) - [Estructura urbana](#)

## Rosario3.com | Noticias y entretenimiento en el diario digital ...

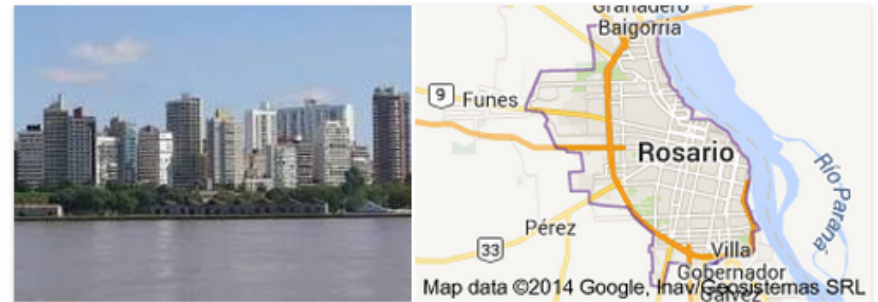
[www.rosario3.com/](http://www.rosario3.com/) ▾ [Translate this page](#)

El diario digital de **Rosario**. Todas las noticias, fotos, videos, foro, clima, cartelera de cine, TV y radio en vivo, y mucho mas.

## Rosario | Inicio

[www.rosarioturismo.com/](http://www.rosarioturismo.com/) ▾ [Translate this page](#)

RecreativaDescubrí **Rosario** en primavera. Esta época es ideal para combinar paseos a pie o en bicicleta con los atractivos más emblemáticos de la ciudad.



## Rosario

City in Argentina

Rosario is the largest city in the province of Santa Fe, in central Argentina. It is located 300 km northwest of Buenos Aires, on the western shore of the Paraná River. [Wikipedia](#)

**Area:** 178 km<sup>2</sup>

**Weather:** 19°C, Wind N at 6 km/h, 78% Humidity

**Local time:** Sunday 12:02 PM

**Population:** 907,718 (1991) UNdata

**Province:** [Santa Fe Province](#)

**Colleges and Universities:** [National University of Rosario](#), [more](#)

## Points of interest

[View 5+ more](#)



# Por ejemplo...

rosario



Web

Images

Videos

News

More ▾

Search tools

About 14,200,000 results (0.27 seconds)

## Sitio de la Municipalidad de Rosario

[www.rosario.gov.ar/](http://www.rosario.gov.ar/) ▾ [Translate this page](#)

**Rosario**, segunda ciudad de la República Argentina. Punto estratégico del Mercosur. Infórmese sobre su comercio exterior y posibles negocios. Conozca su ...

[Trámites](#) - [Registro Único de Postulantes](#) - [InfoMapa](#) - [Recorridos de colectivos](#)

## Rosario, Santa Fe - Wikipedia, the free encyclopedia

[en.wikipedia.org/wiki/Rosario,\\_Santa\\_Fe](http://en.wikipedia.org/wiki/Rosario,_Santa_Fe) ▾

**Rosario** is the largest city in the province of Santa Fe, in central Argentina. It is located 300 km (186 mi) northwest of Buenos Aires, on the western shore of the ...

[History](#) - [Government](#) - [Economy](#) - [Culture](#)

## Rosario (Argentina) - Wikipedia, la enciclopedia libre

[es.wikipedia.org/wiki/Rosario\\_\(Argentina\)](http://es.wikipedia.org/wiki/Rosario_(Argentina)) ▾ [Translate this page](#)

La ciudad de **Rosario** está ubicada en el centro-este argentino, en la provincia de Santa Fe. Es la tercera ciudad más poblada de Argentina después de Buenos ...

[Toponimia](#) - [Historia](#) - [Geografía](#) - [Estructura urbana](#)

## Rosario3.com | Noticias y entretenimiento en el diario digital ...

[www.rosario3.com/](http://www.rosario3.com/) ▾ [Translate this page](#)

El diario digital de **Rosario**. Todas las noticias, fotos, videos, foro, clima, cartelera de cine, TV y radio en vivo, y mucho mas.

## Rosario | Inicio

[www.rosarioturismo.com/](http://www.rosarioturismo.com/) ▾ [Translate this page](#)

RecreativaDescubrí **Rosario** en primavera. Esta época es ideal para combinar paseos a pie o en bicicleta con los atractivos más emblemáticos de la ciudad.



## Rosario

City in Argentina

Rosario is the largest city in the province of Santa Fe, in central Argentina. It is located 300 km northwest of Buenos Aires, on the western shore of the Paraná River. [Wikipedia](#)

**Area:** 178 km<sup>2</sup>

**Weather:** 19°C, Wind N at 6 km/h, 78% Humidity

**Local time:** Sunday 12:02 PM

**Population:** 907,718 (1991) UNdata

**Province:** [Santa Fe Province](#)

**Colleges and Universities:** [National University of Rosario](#), [more](#)

## Points of interest

[View 5+ more](#)





# Por ejemplo...

rosario



Web

Images

Videos

News

More ▾

Search tools

About 14,200,000 results (0.27 seconds)

## Sitio de la Municipalidad de Rosario

[www.rosario.gov.ar/](http://www.rosario.gov.ar/) ▾ Translate this page

**Rosario**, segunda ciudad de la República Argentina. Punto estratégico del Mercosur.

Infórmese sobre su comercio exterior y posibles negocios. Conozca su ...

[Trámites](#) - [Registro Único de Postulantes](#) - [InfoMapa](#) - [Recorridos de colectivos](#)

## Rosario, Santa Fe - Wikipedia, the free encyclopedia

[en.wikipedia.org/wiki/Rosario,\\_Santa\\_Fe](http://en.wikipedia.org/wiki/Rosario,_Santa_Fe) ▾

**Rosario** is the largest city in the province of Santa Fe, in central Argentina. It is located 300 km (186 mi) northwest of Buenos Aires, on the western shore of the ...

[History](#) - [Government](#) - [Economy](#) - [Culture](#)

## Rosario (Argentina) - Wikipedia, la enciclopedia libre

[es.wikipedia.org/wiki/Rosario\\_\(Argentina\)](http://es.wikipedia.org/wiki/Rosario_(Argentina)) ▾ Translate this page

La ciudad de **Rosario** está ubicada en el centro-este argentino, en la provincia de Santa Fe. Es la tercera ciudad más poblada de Argentina después de Buenos ...

[Toponimia](#) - [Historia](#) - [Geografía](#) - [Estructura urbana](#)

## Rosario3.com | Noticias y entretenimiento en el diario digital ...

[www.rosario3.com/](http://www.rosario3.com/) ▾ Translate this page

El diario digital de **Rosario**. Todas las noticias, fotos, videos, foro, clima, cartelera de cine, TV y radio en vivo, y mucho mas.

## Rosario | Inicio

[www.rosarioturismo.com/](http://www.rosarioturismo.com/) ▾ Translate this page

RecreativaDescubrí **Rosario** en primavera. Esta época es ideal para combinar paseos a pie o en bicicleta con los atractivos más emblemáticos de la ciudad.

Cómo  
obtenemos  
estos?



## Rosario

City in Argentina

Rosario is the largest city in the province of Santa Fe, in central Argentina. It is located 300 km northwest of Buenos Aires, on the western shore of the Paraná River. [Wikipedia](#)

**Area:** 178 km<sup>2</sup>

**Weather:** 19°C, Wind N at 6 km/h, 78% Humidity

**Local time:** Sunday 12:02 PM

**Population:** 907,718 (1991) UNdata

**Province:** [Santa Fe Province](#)

**Colleges and Universities:** [National University of Rosario](#), [more](#)

## Points of interest

[View 5+ more](#)



# Motores de búsqueda



- ¿Son importantes?
  - ~90% del tráfico a la mayoría de los sitios proviene de un motor de búsqueda
  - Son la **primera interface** entre los usuarios y la web
    - En el caso de sitios comerciales (productos) estar más allá de la posición 30 es ser “prácticamente” invisible.
  - Atraen la mayor **diversidad** de usuarios que cualquier sitio.
  - ~ 85% de las sesiones de usuario incluyen un MB
  - ~ 90% de los usuarios los usan para navegar la web





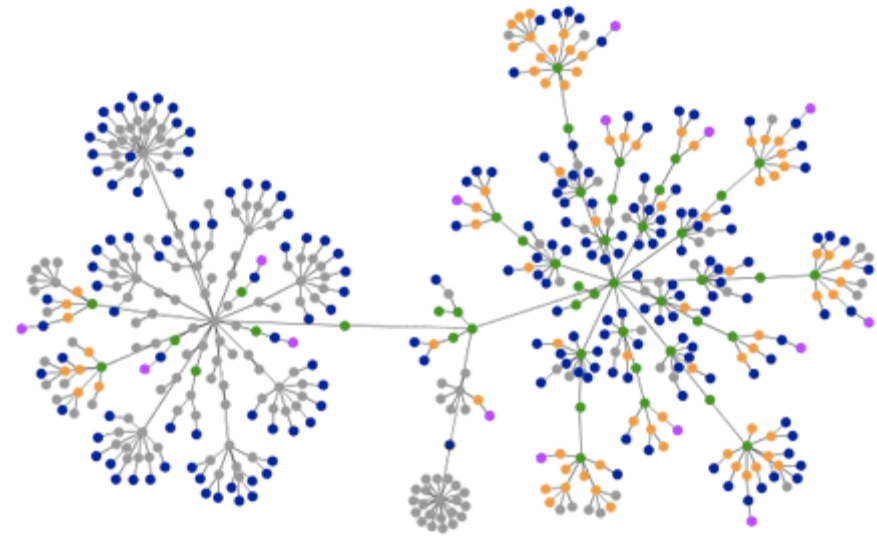
# Perspectivas

- **Usuarios**

- Respuestas relevantes y rápidas
- Necesidades? Queries?

- **Motor de Búsqueda**

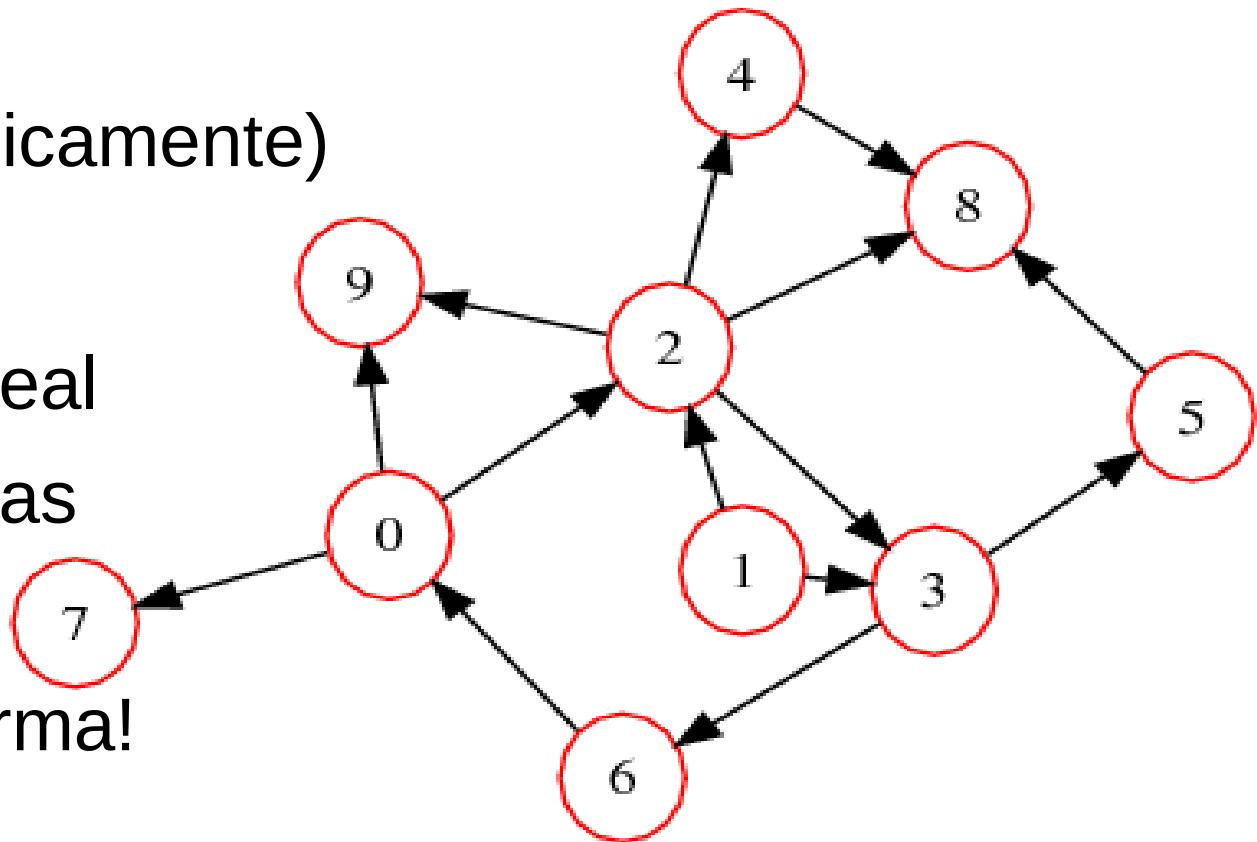
- Manejar la complejidad de la web
- Atraer más usuarios
- Reducir costos operacionales
- Incrementar ingresos (ads)



# Caracterizando el Problema

# La Web

- Repositorio distribuido
  - Grafo dirigido masivo
  - Complejo
- HTTP y HTML (básicamente)
- Hipertextual
  - Estructura no-lineal
  - Relaciones lógicas
  - No “tan” obvia
- Hoy es una plataforma!



# Qué es lo que dificulta la tarea de búsqueda?



**Tamaño**



**Diversidad**



**Dinamismo**



# Qué es lo que dificulta la tarea de búsqueda?



**Tamaño**



**Diversidad**



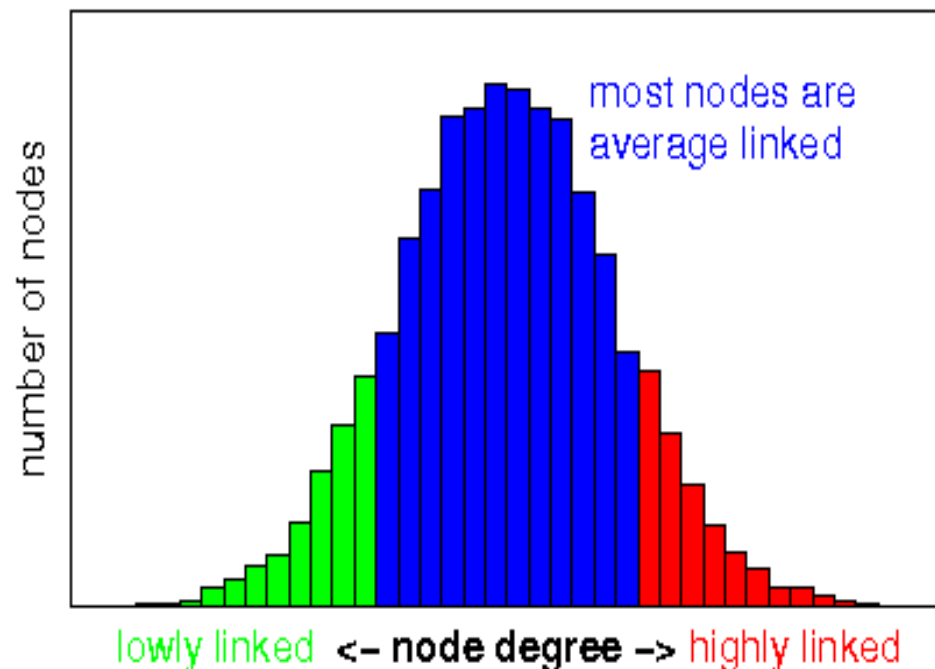
**Dinamismo**

Estas tres características también se observan en los **usuarios!!!!**

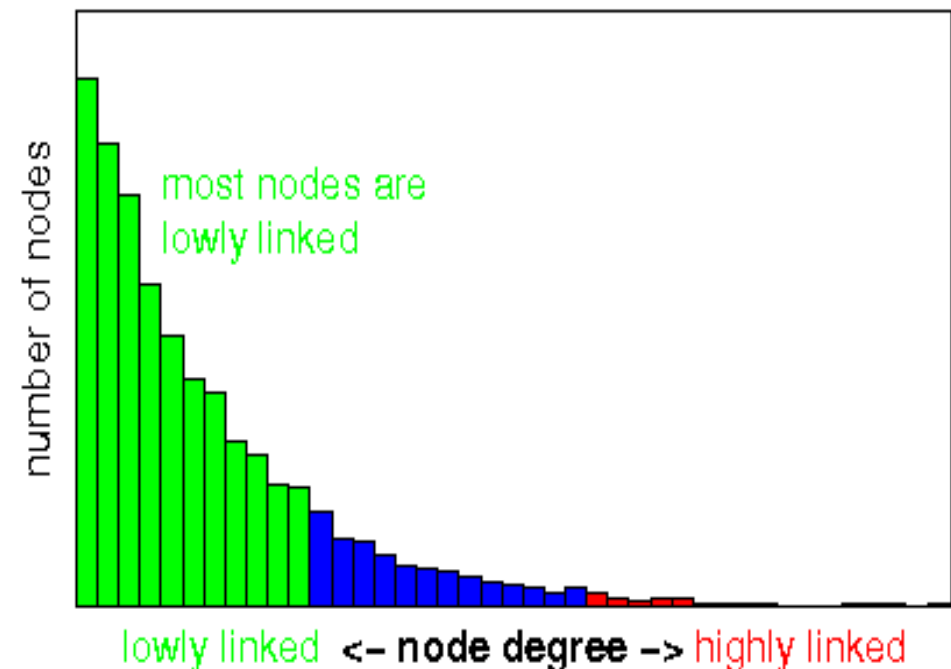
# Estructura de grafo

- “*Graph Structure on the Web*” [Broder, 1999]
- Grado entrante/saliente → Power-Law:  $y(x) = kx^n$

random networks



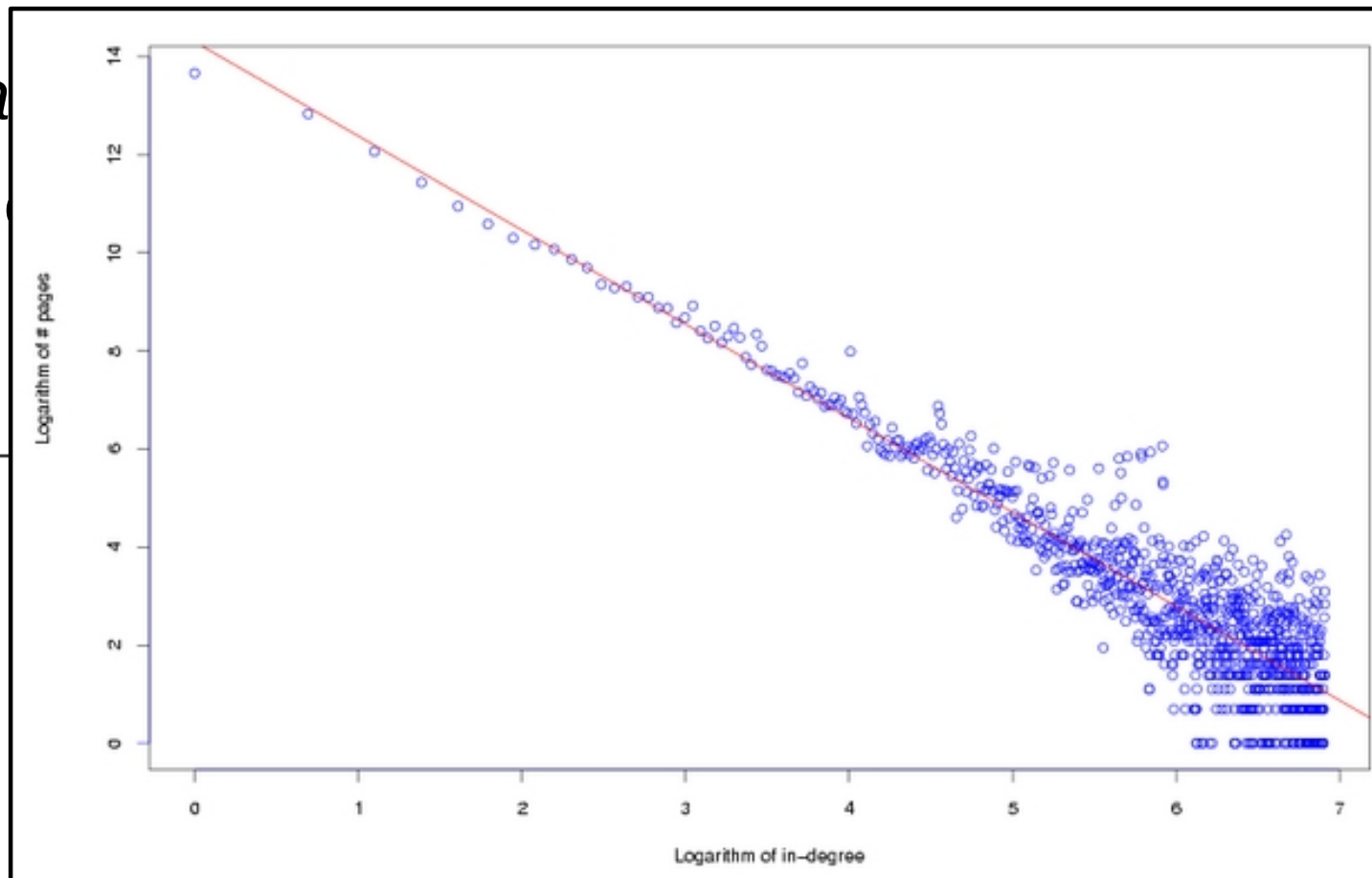
real networks (power-law, scale-free)





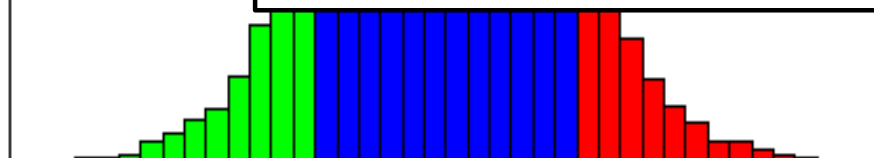
# Estructura de grafo

- “Gra
- Grad

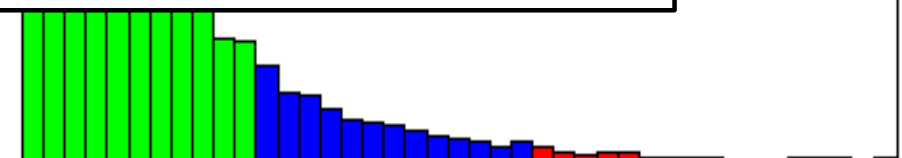


(le-free)

number of nodes



lowly linked <- node degree -> highly linked



lowly linked <- node degree -> highly linked

# Todo crece!!!



**2,988,263,442**

Internet Users in the world

[sources](#)

[more info](#)

[watch all](#)



**1,084,310,218**

Total number of Websites

[sources](#)

[more info](#)

[watch all](#)

<http://www.internetlivestats.com/>



**2,633,871,918**

Google searches [today](#)

[sources](#)

[more info](#)

[in 1 second](#)

## Y páginas web?

- En 2005, 11.500 millones de páginas [Gulli, et al., 2005]
- Hoy?

# Todo crece!!!



2,988,263,442

Internet Users in the world

[sources](#)

[more info](#)

[watch all](#)



1,084,310,218

Total number of Websites

[sources](#)

<http://www.internetlivestats.com/>

[more info](#)

[watch all](#)



2,633,871,918

Google searches [today](#)

[sources](#)

[more info](#)

[in 1 second](#)

## Y páginas web?

- En 2005, 11.500 millones de páginas [Gulli, et al., 2005]
- Hoy?

### •Pregunta abierta

- Nodos temporales
- Dinámica
- Duplicados
- Profunda: 95%?

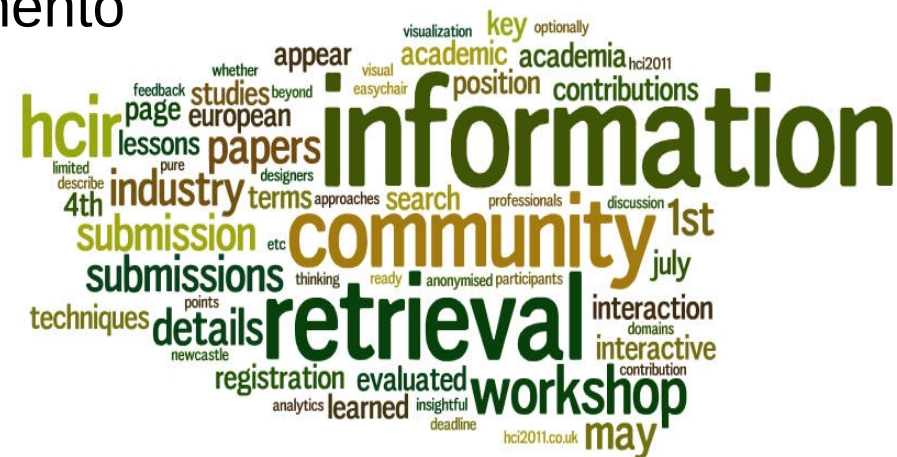
# Recuperación de Información

Involucra un conjunto de modelos/técnicas/algoritmos para obtener información **relevante** a una necesidad desde diferentes fuentes de información. En general, se basa en búsquedas "full-text"

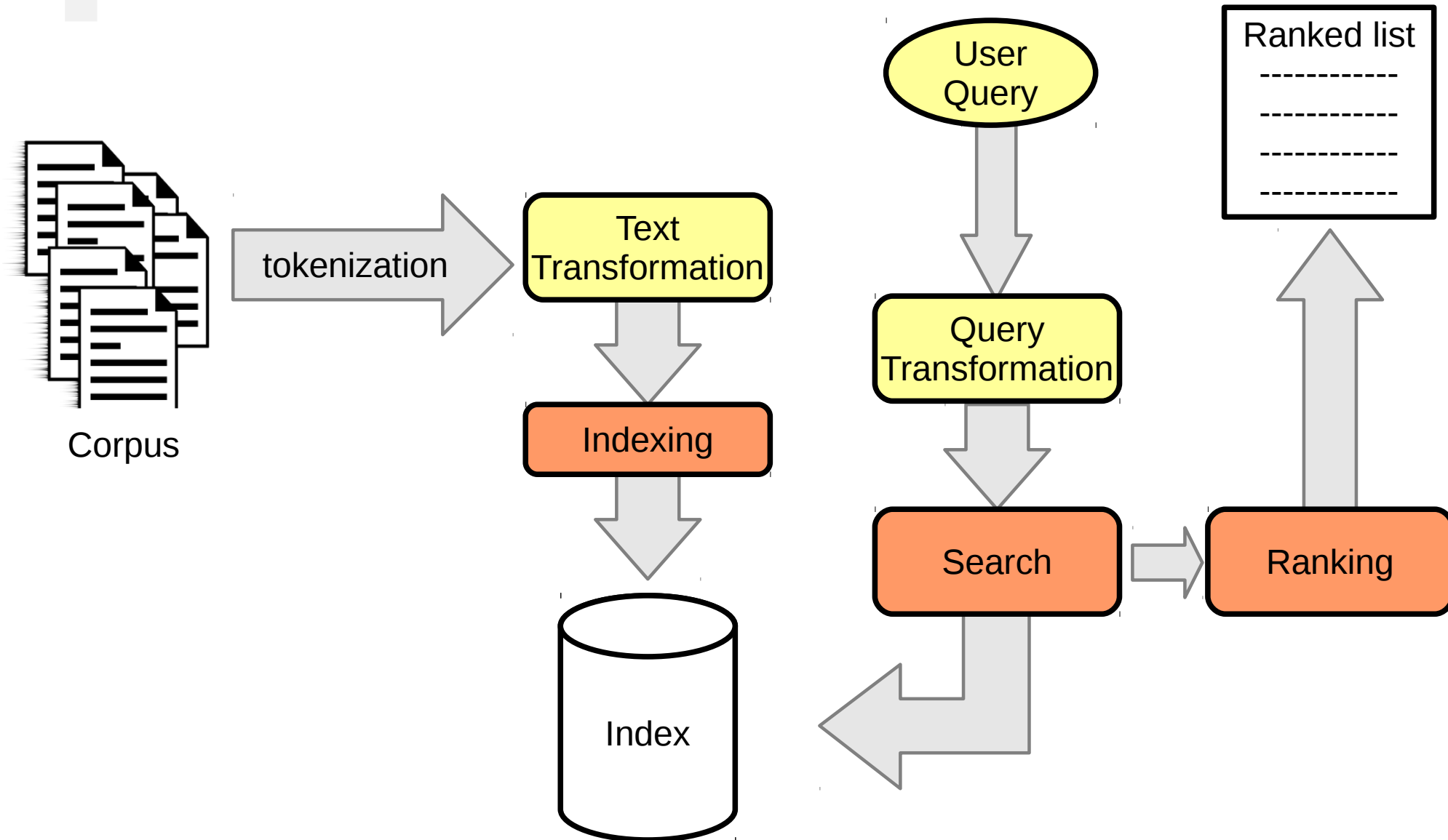
Intenta ayudar a abordar el problema conocido como "*information overload*".

Ejemplos,

- Acceso a libros, revistas, publicaciones, etc.
- En general, a cualquier tipo de documento
- Los motores de búsqueda son la aplicación más visible



# Arquitectura de un SRI



# Cómo se almacenan los documentos?

## Colección

doc 1= {casa, casa, gato, gato, gato, perro}

doc 2= {gato, gato, mate, mate, termo}

doc 3= {sopa, termo, termo}

doc 4= {auto, auto, perro, perro}

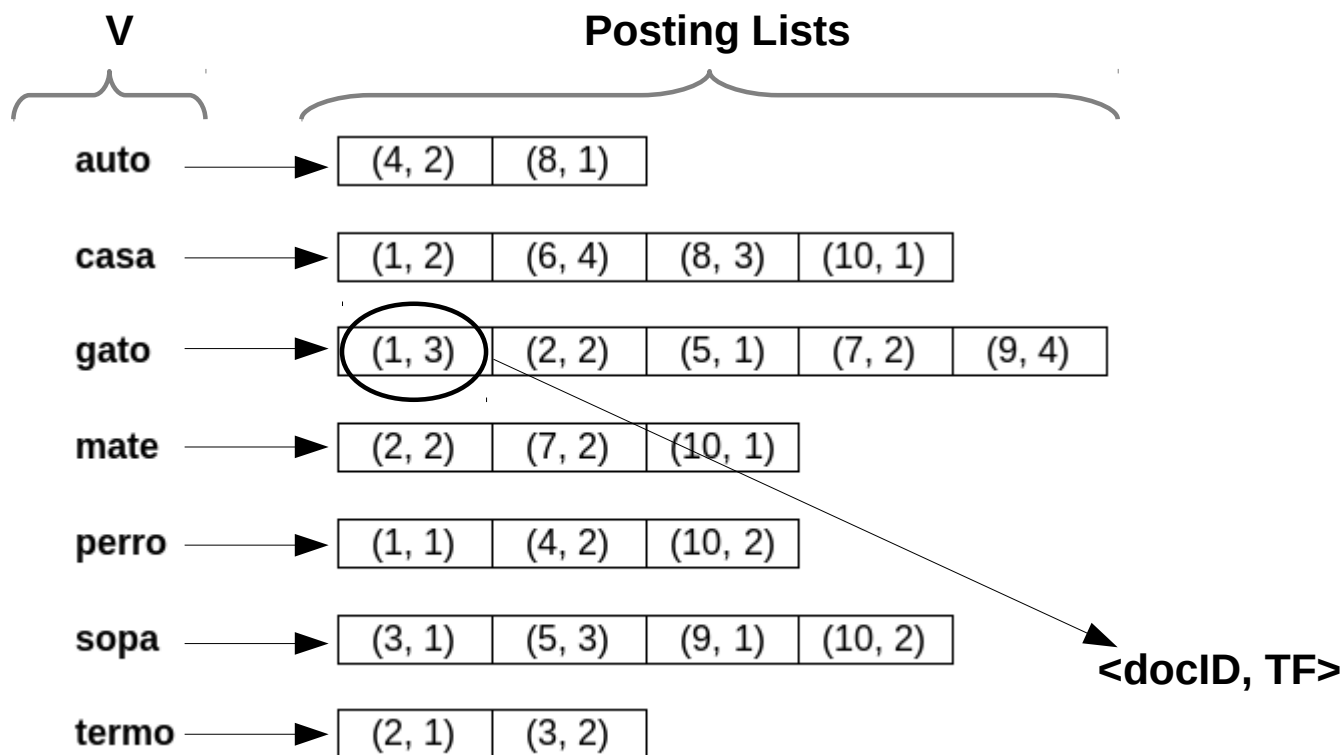
doc 5= {gato, sopa, sopa, sopa}

...

...

...

## Índice Invertido





# Cómo se comparan documentos y consultas?

- **Hasta ahora matemática/estadística (vs NLP)**
  - Representación de los docs/queries
  - Modelos de RI
    - $M[D, Q, F, R(d_i, q_j)]$
- **Cómo se ponderan los términos?**
  - TF\*IDF (básico)

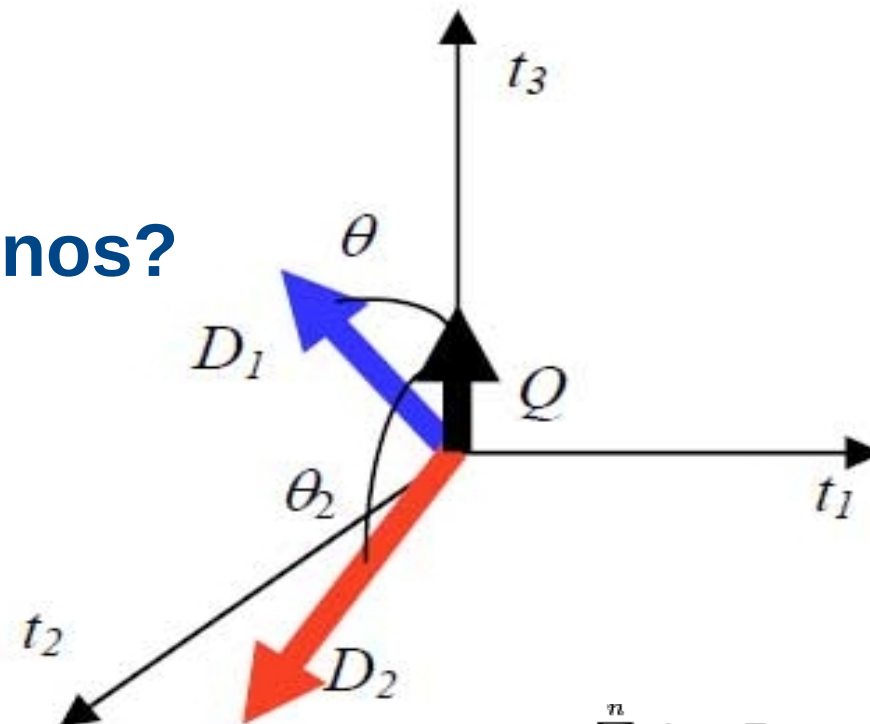


Diagram illustrating vector representation of documents and queries in a 3D space. A black vector labeled  $Q$  (query) points upwards along the  $t_3$  axis. Two other vectors,  $D_1$  (blue) and  $D_2$  (red), represent documents.  $D_1$  is in the  $t_1$ - $t_3$  plane, and  $D_2$  is in the  $t_2$ - $t_3$  plane. The angle between  $Q$  and  $D_1$  is labeled  $\theta$ . The axes are labeled  $t_1$ ,  $t_2$ , and  $t_3$ .

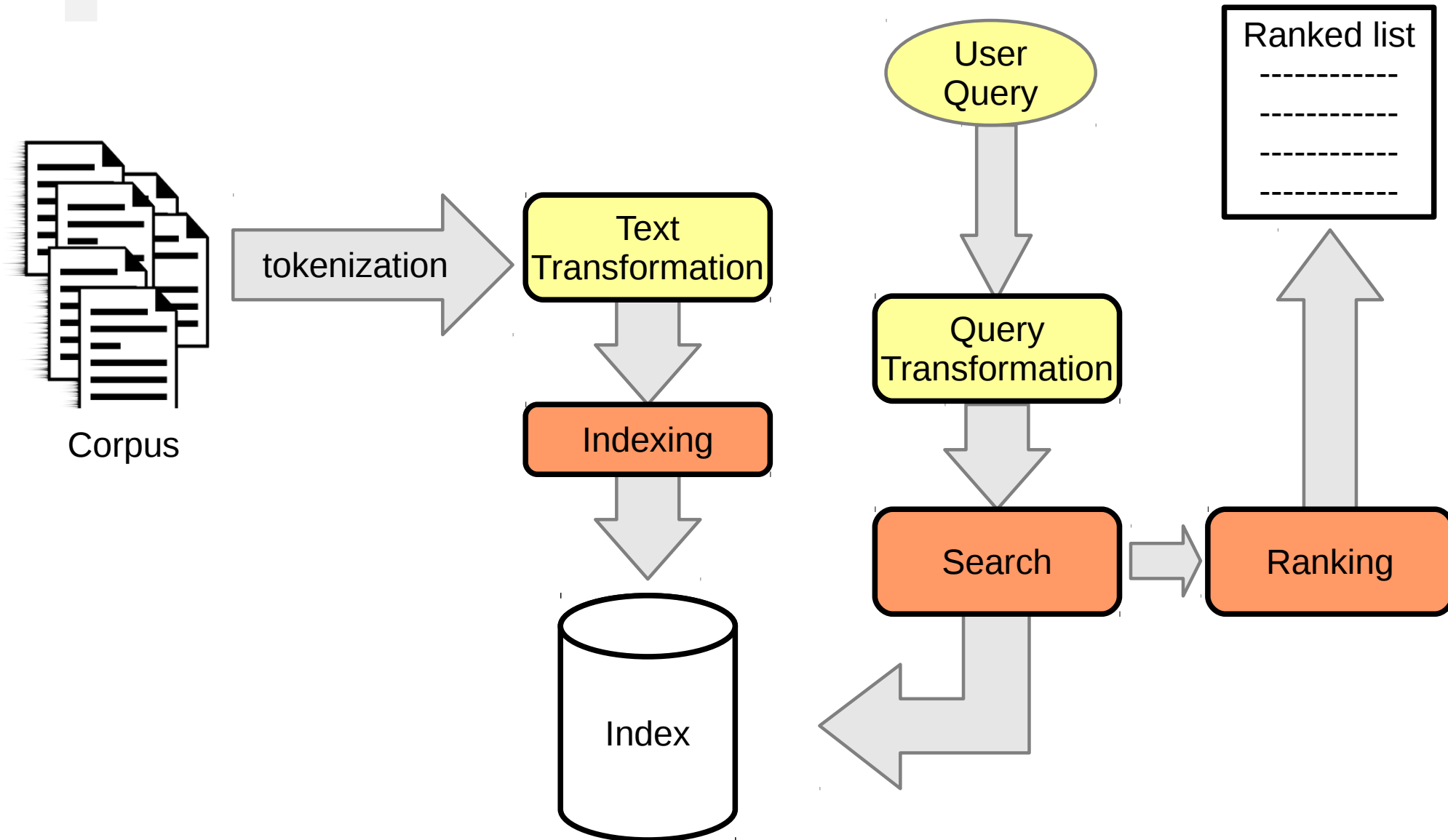
$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$



# Cómo se hace la búsqueda?

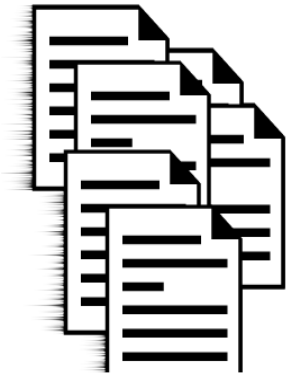
- **Básicamente, se recuperan las posting lists de los términos de las consultas:**
  - Intersección/unión para queries booleanos
  - Cálculo de la similitud para “ranked queries”
- **La similitud se calcula de acuerdo al M elegido**
  - Booleano, vectorial, probabilístico
- **Ranking**
  - En general, es una lista ordenada por  $\text{sim}(d_i, q_j)$

# Pero en los MB Web

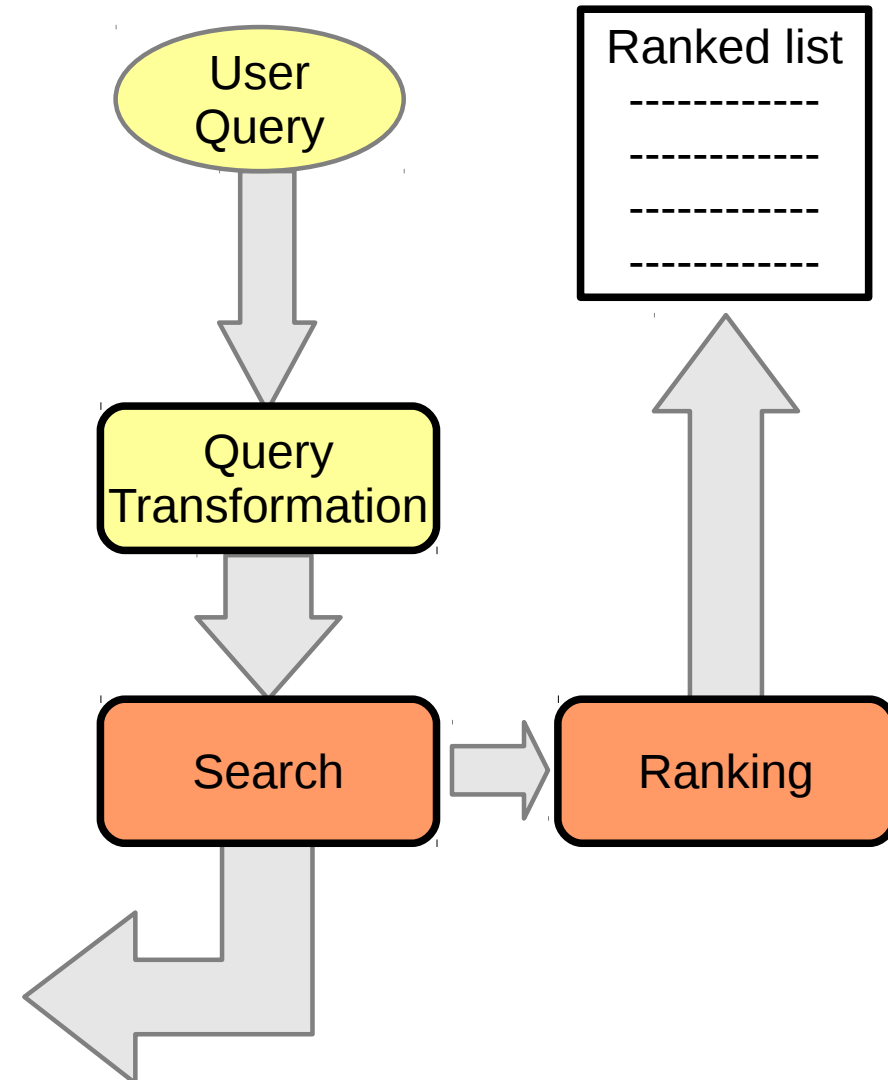
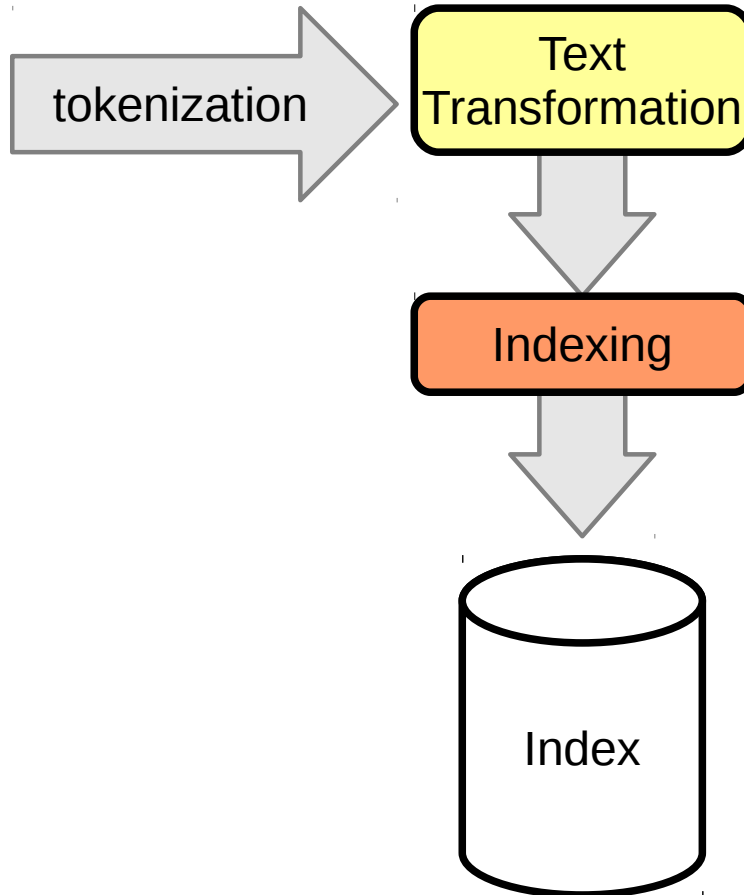


# Pero en los MB Web

No lo  
tenemos



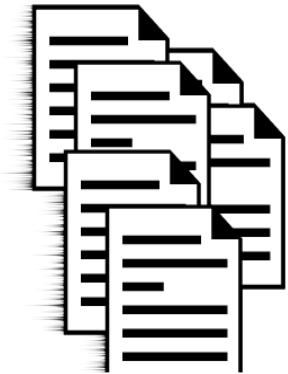
Corpus



# Pero en los MB Web

No lo  
tenemos

Múltiples  
formatos



Corpus

tokenization

Text  
Transformation

Indexing

Index

User  
Query

Query  
Transformation

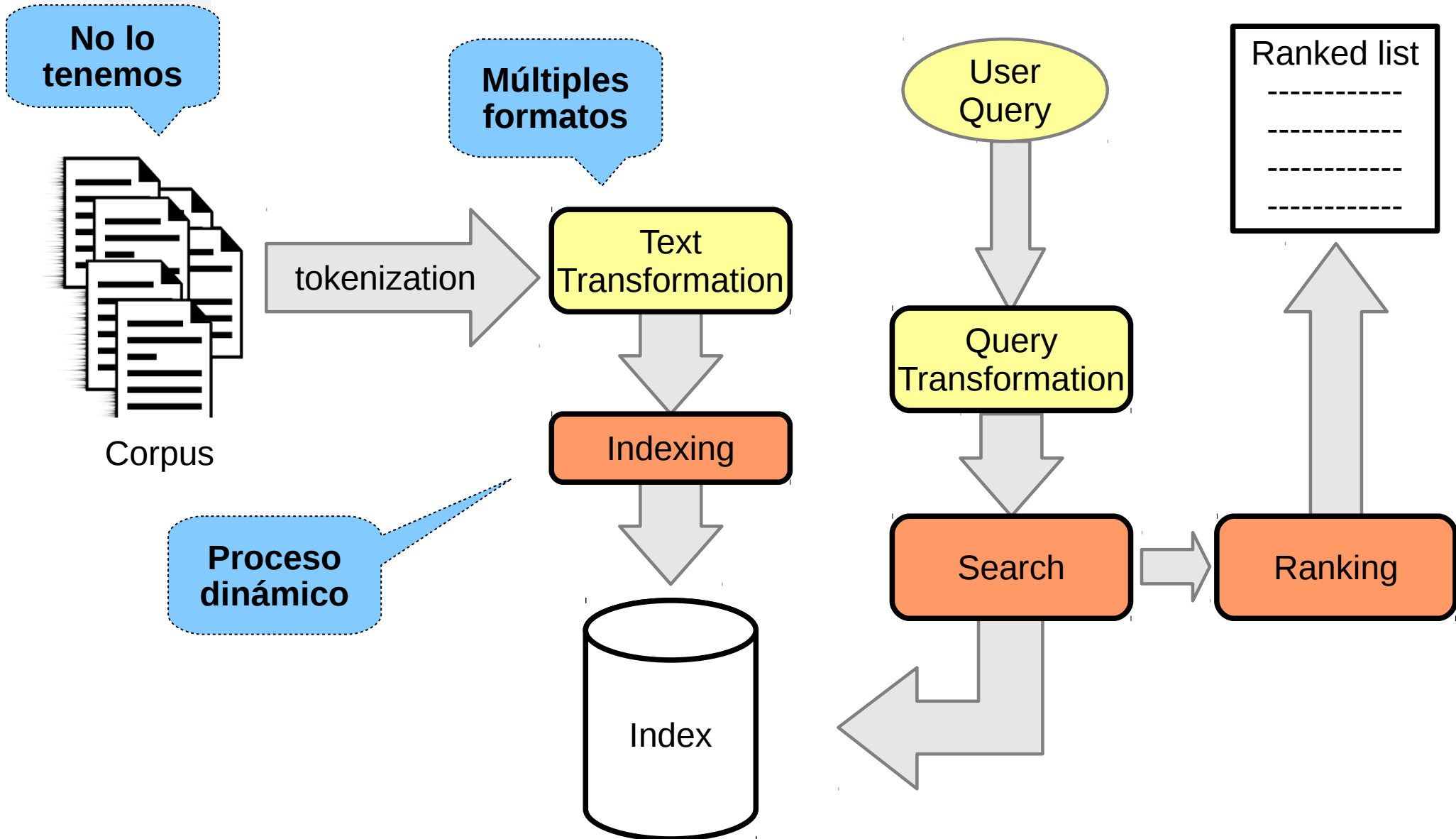
Search

Ranking

Ranked list

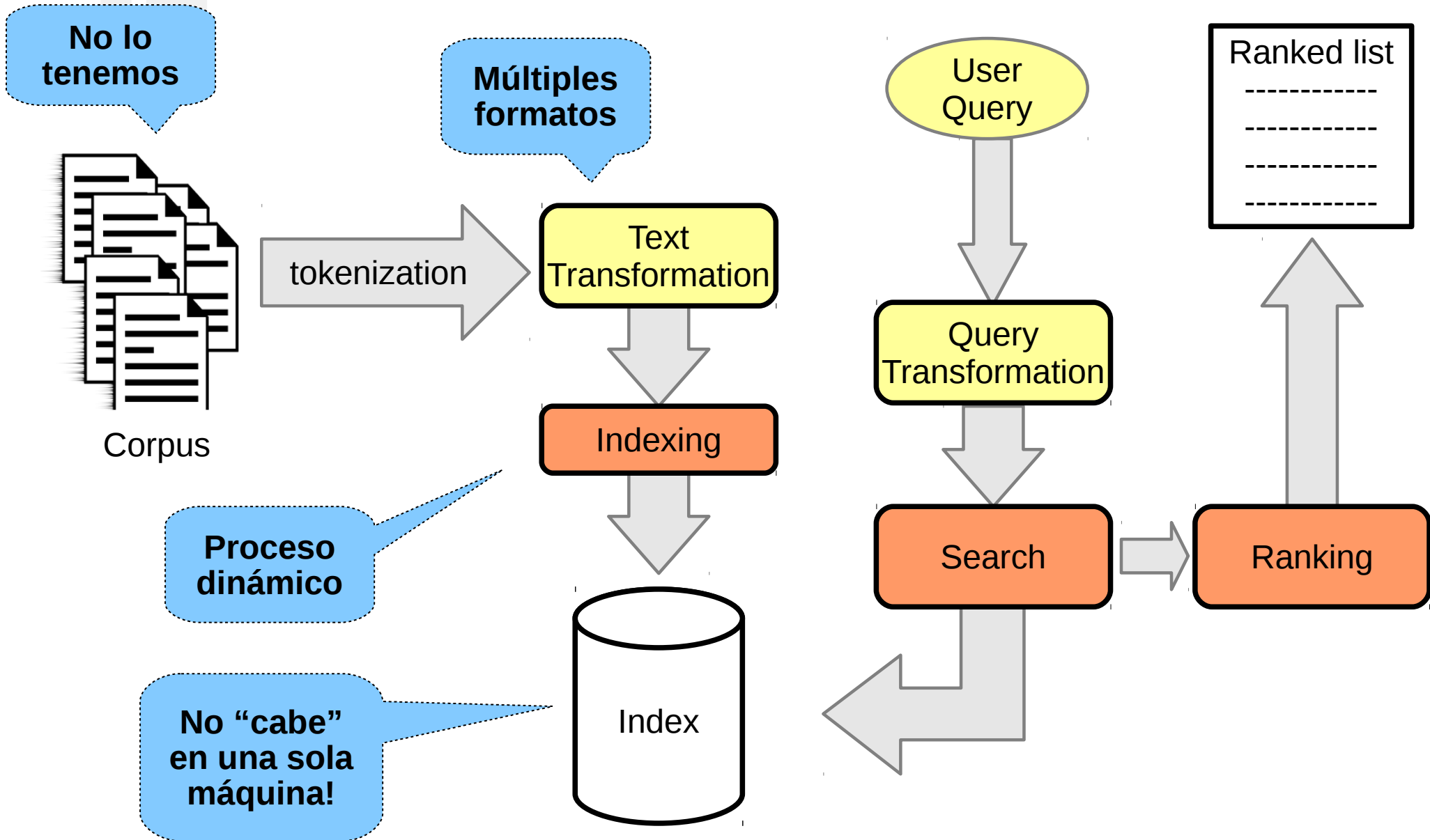
-----  
-----  
-----  
-----

# Pero en los MB Web

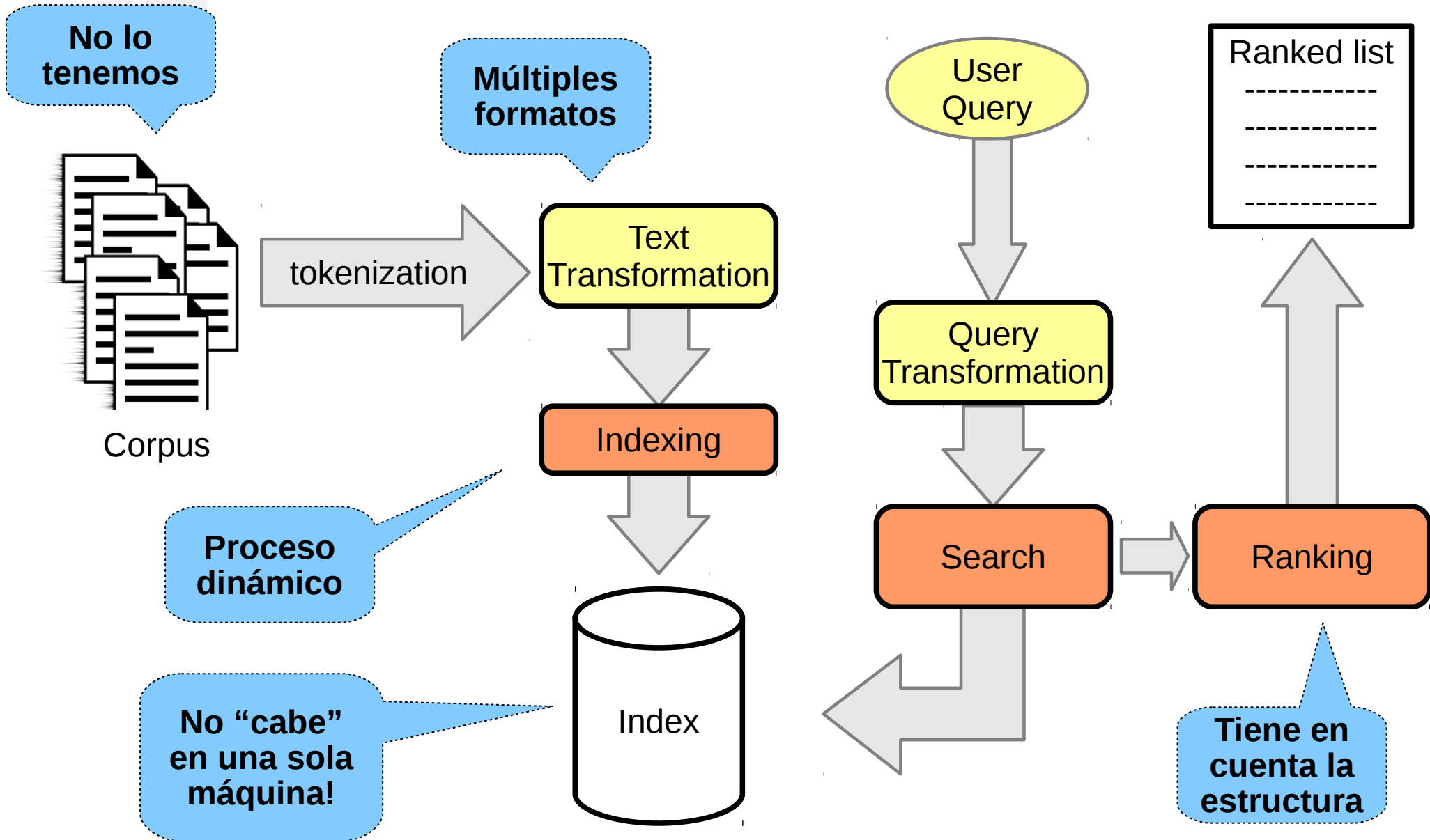




# Pero en los MB Web

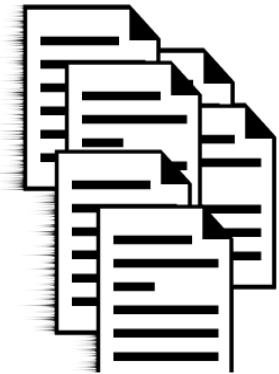


# Pero en los MB Web



# Pero en los MB Web

No lo tenemos



Corpus

tokenization

Múltiples formatos

Text Transformation

Indexing

Index

Proceso dinámico

No "cabe" en una sola máquina!

User Query

Query Transformation

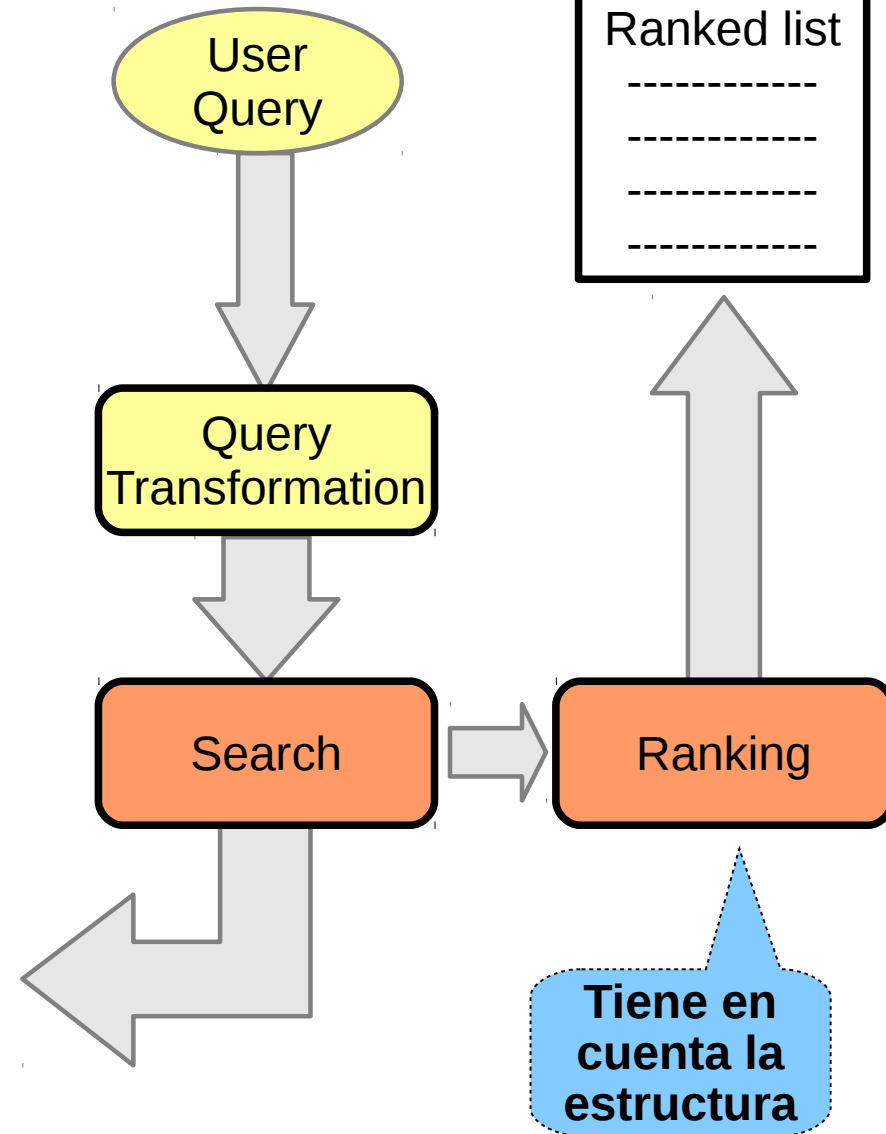
Search

Usuarios de diferentes contextos

Ranked list

Ranking

Tiene en cuenta la estructura



# Crawling → Obtener la colección

Es “básicamente” un problema de recorrido de un grafo





```
S := {páginas iniciales}

mientras no-vacía (S)
{
    tomar s desde S

    si s no fue recuperada antes:
        recuperar s

    parsear s

    para cada link l en s:
        agregar l a S
}
```



# Crawling → Cuestiones

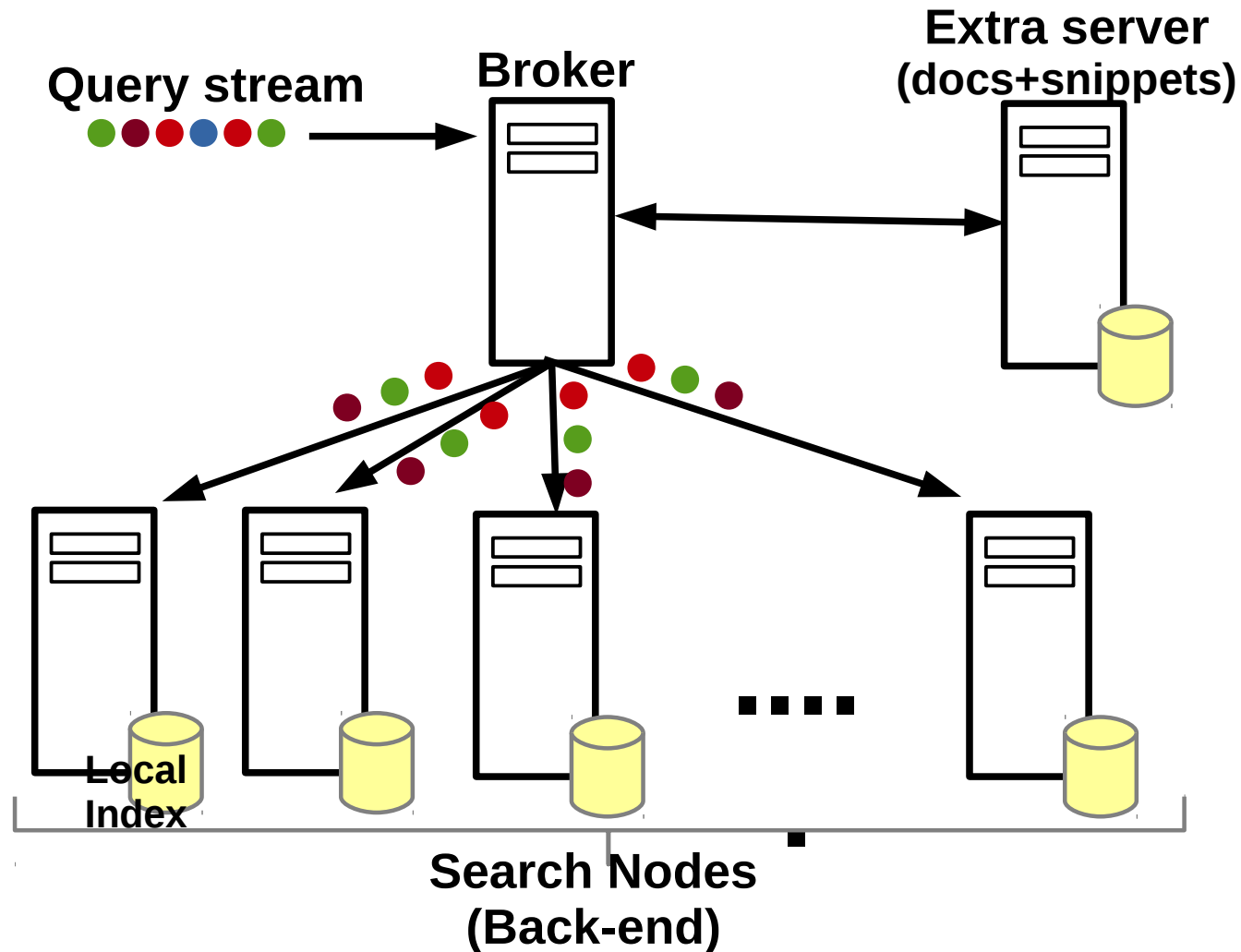
- **¿Cómo hacer el crawling?**
  - Calidad (las mejores páginas primero)
  - Eficiencia (evitar duplicados)
  - Cortesía (con los servidores)
- **¿Cuánto recolectar?**
  - Cobertura
  - Cobertura relativa
- **¿Con qué frecuencia?**
  - “Frescura”



# **Búsquedas a Gran Escala**



# Arquitectura de un MB





# Estructuras de Datos

- **Más sofisticadas, comprimidas y distribuidas!**

$L_i = \{5, 11, 17, 21, 26, 34, 36, 37, 45, 48, 51, 52, 57, 80, 89, 91, 94, 101, 104, 119\}$

- D-Gaps:  $\{5, 6, 6, 4, 5, 9, 2, 1, 8, 3, 3, 1, 5, 23, 9, 2, 3, 7, 3, 1\}$
- **Compresión** (V-Bye, Rice, Golomb, PforDelta)
  - Tradeoff entre tasa de compresión y tiempo de descompresión!

# Estructuras de Datos

- **Más sofisticadas, comprimidas y distribuidas!**

$L_i = \{5, 11, 17, 21, 26, 34, 36, 37, 45, 48, 51, 52, 57, 80, 89, 91, 94, 101, 104, 119\}$

- D-Gaps:  $\{5, 6, 6, 4, 5, 9, 2, 1, 8, 3, 3, 1, 5, 23, 9, 2, 3, 7, 3, 1\}$
- **Compresión** (V-Bye, Rice, Golomb, PforDelta)
  - Tradeoff entre tasa de compresión y tiempo de descompresión!
  - Pero:
    - También almacenamos la frecuencia  $\langle \text{docId}, \text{TF} \rangle$
    - Conviene “juntar”  $n$  docIDs y frecuencias en bloques

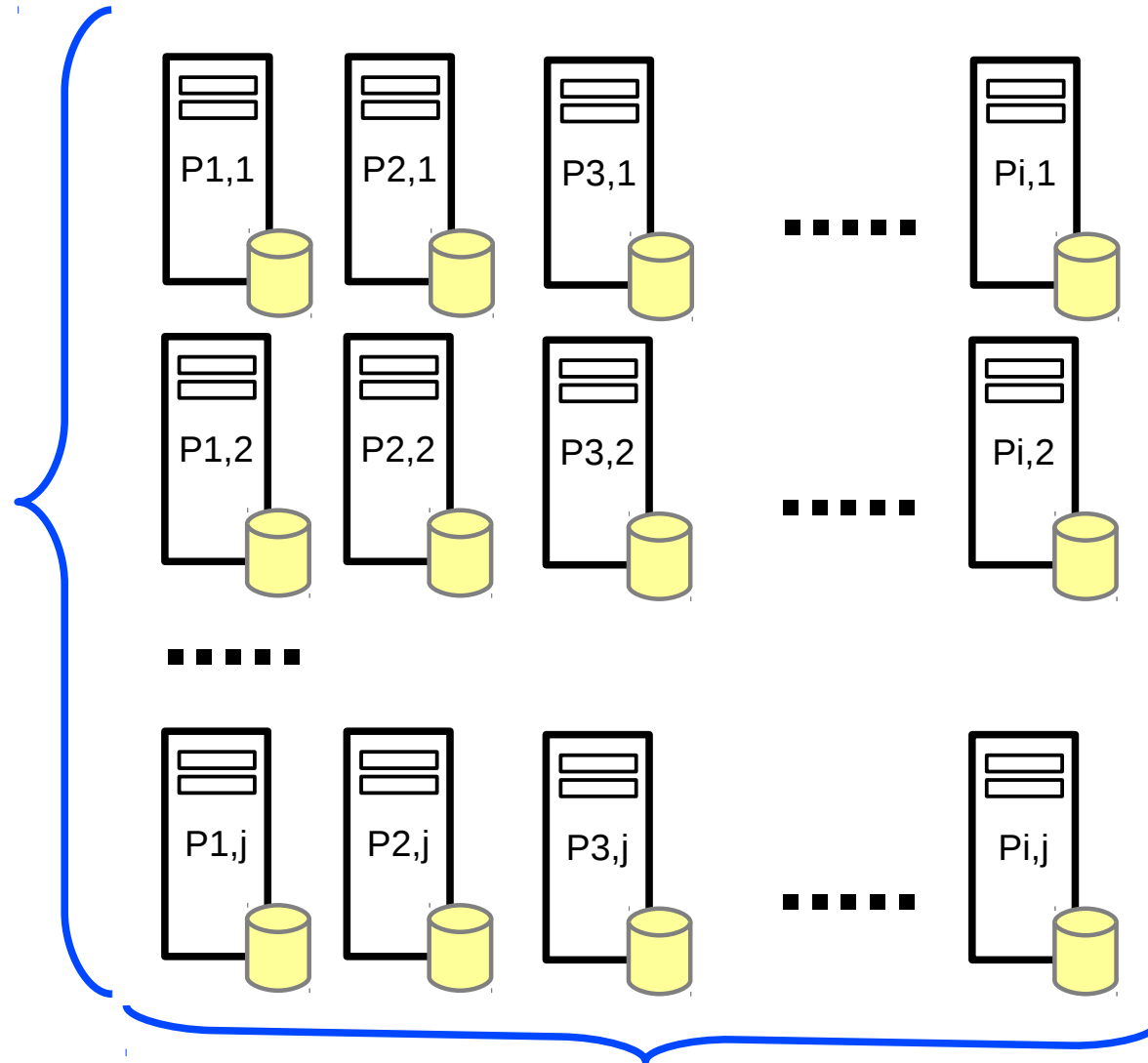
$\text{id}_1, \text{id}_2, \text{id}_3, \dots, \text{id}_n$

$\text{tf}_1, \text{tf}_2, \text{tf}_3, \dots, \text{tf}_n$

# Particionado y Replicación

## Replicación

+ tolerancia a fallos  
+ query throughput



## Particionado

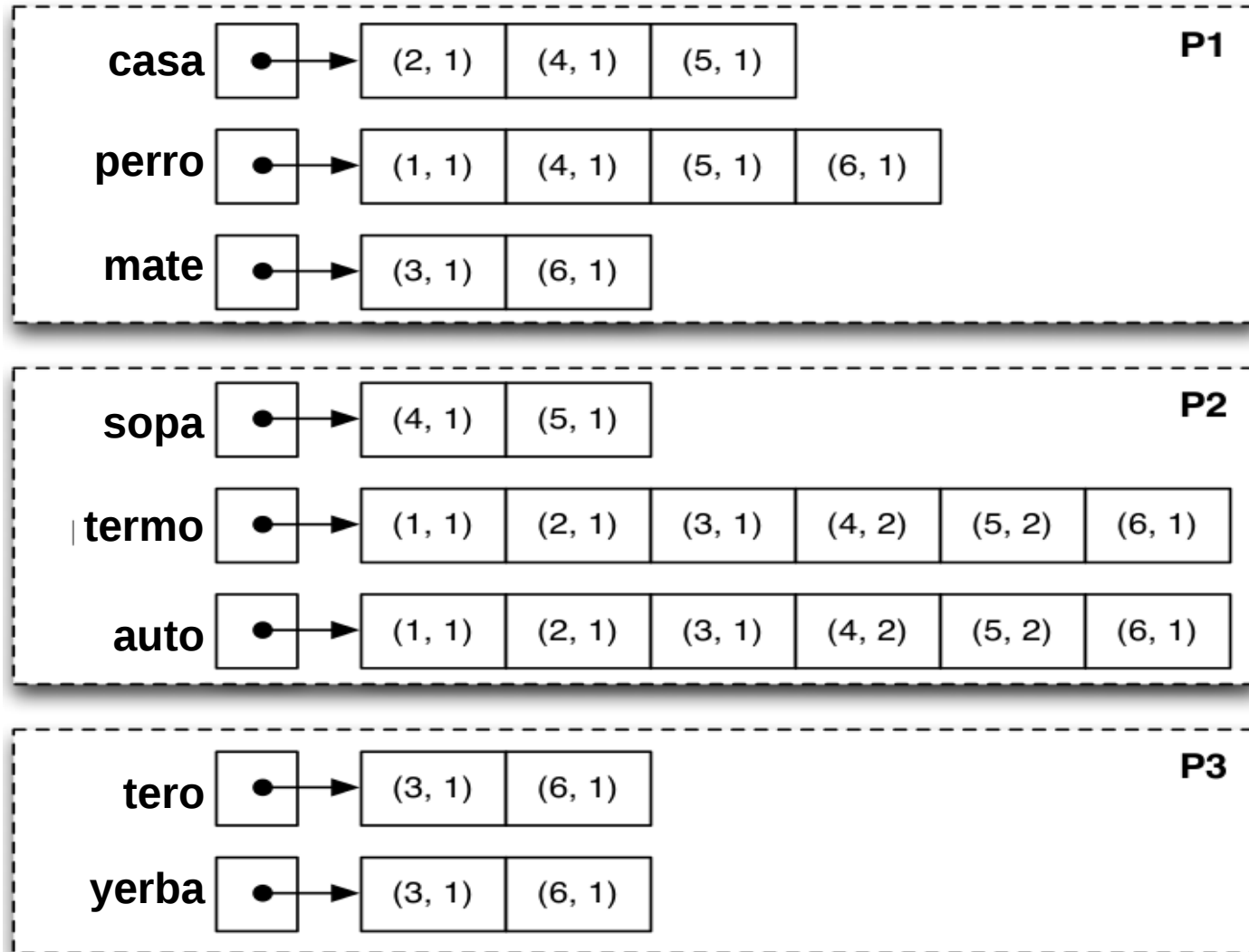
- tiempo promedio de procesamiento de cada queries



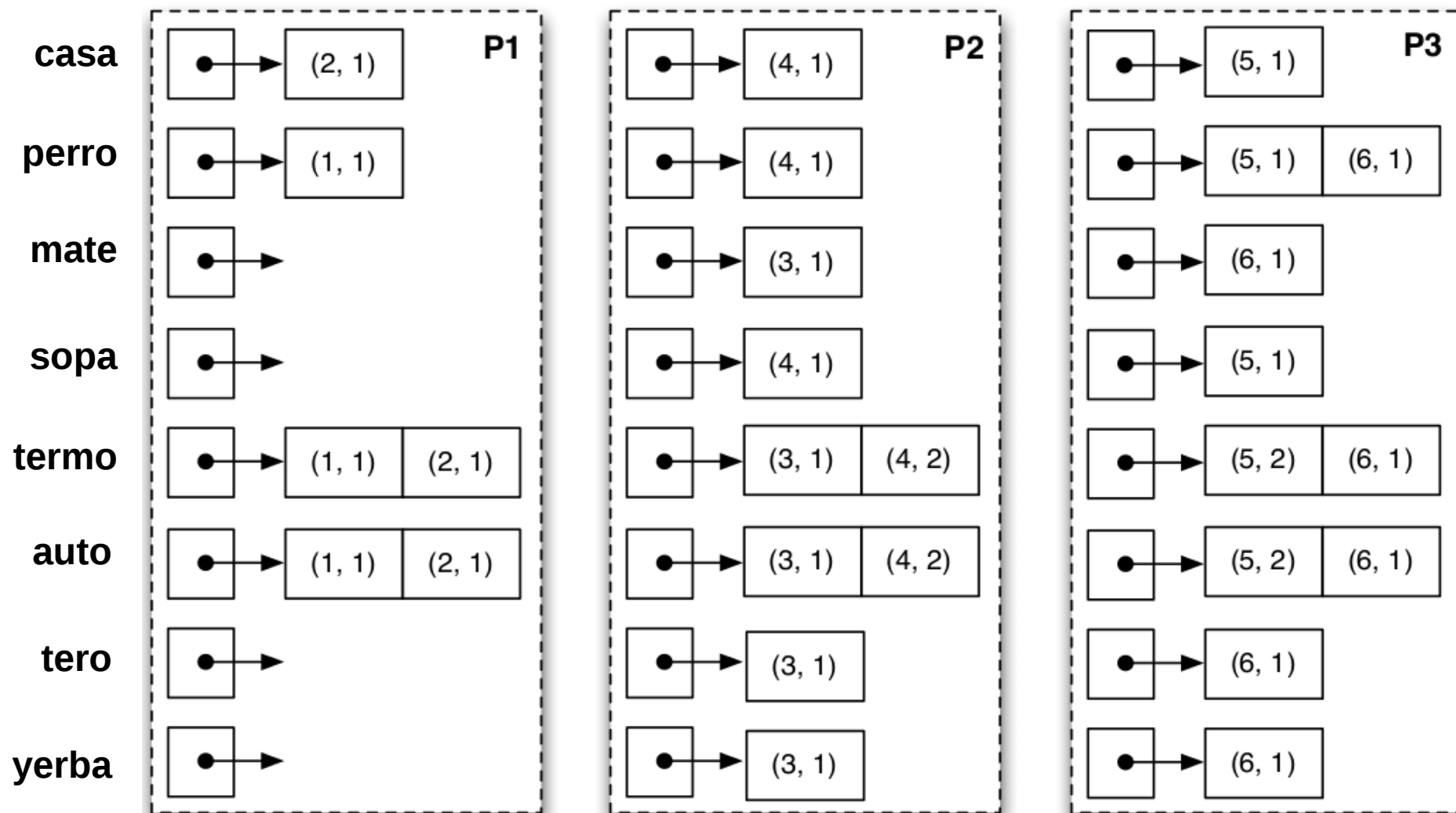
# Partición del Índice

- **Dos enfoques clásicos**
  - Partición por Términos
  - Partición por Documentos
- **Dos enfoques híbridos**
  - Índice 2D
  - Índice 3D

# Partición por Términos

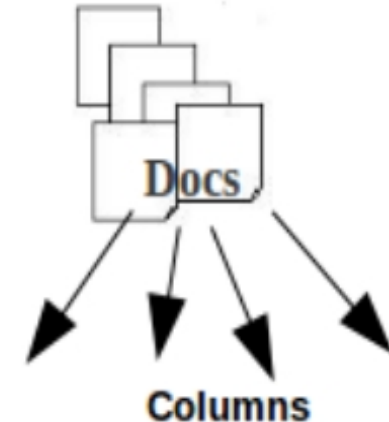
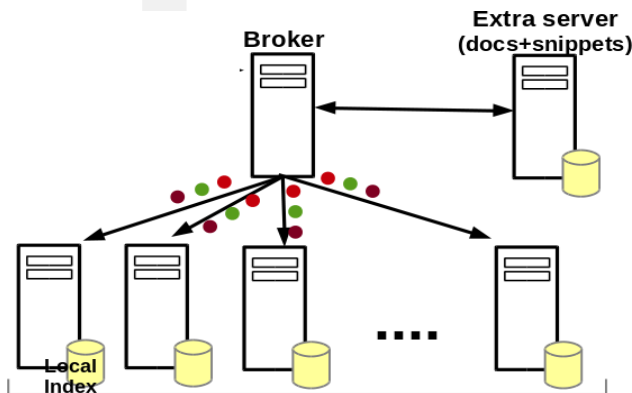


# Partición por Documentos





# Índice 2D [Feuerstein, SPIRE 2009]



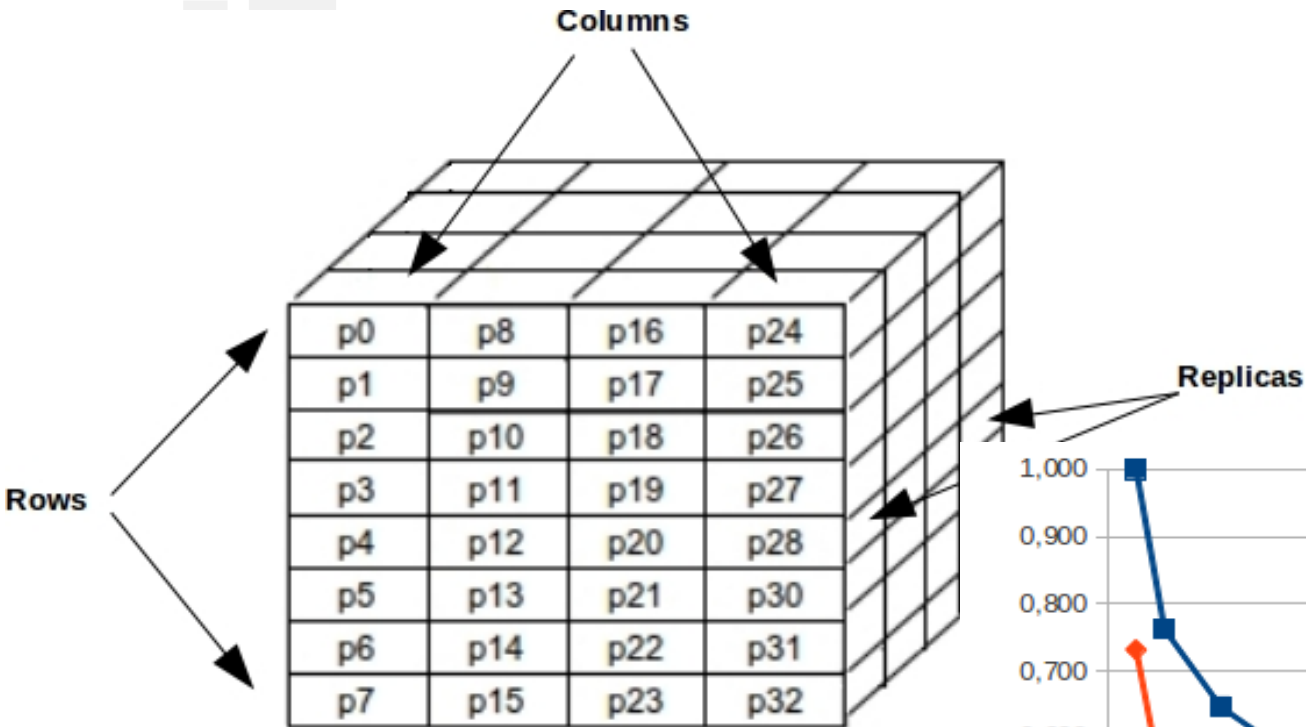
Terms

Rows

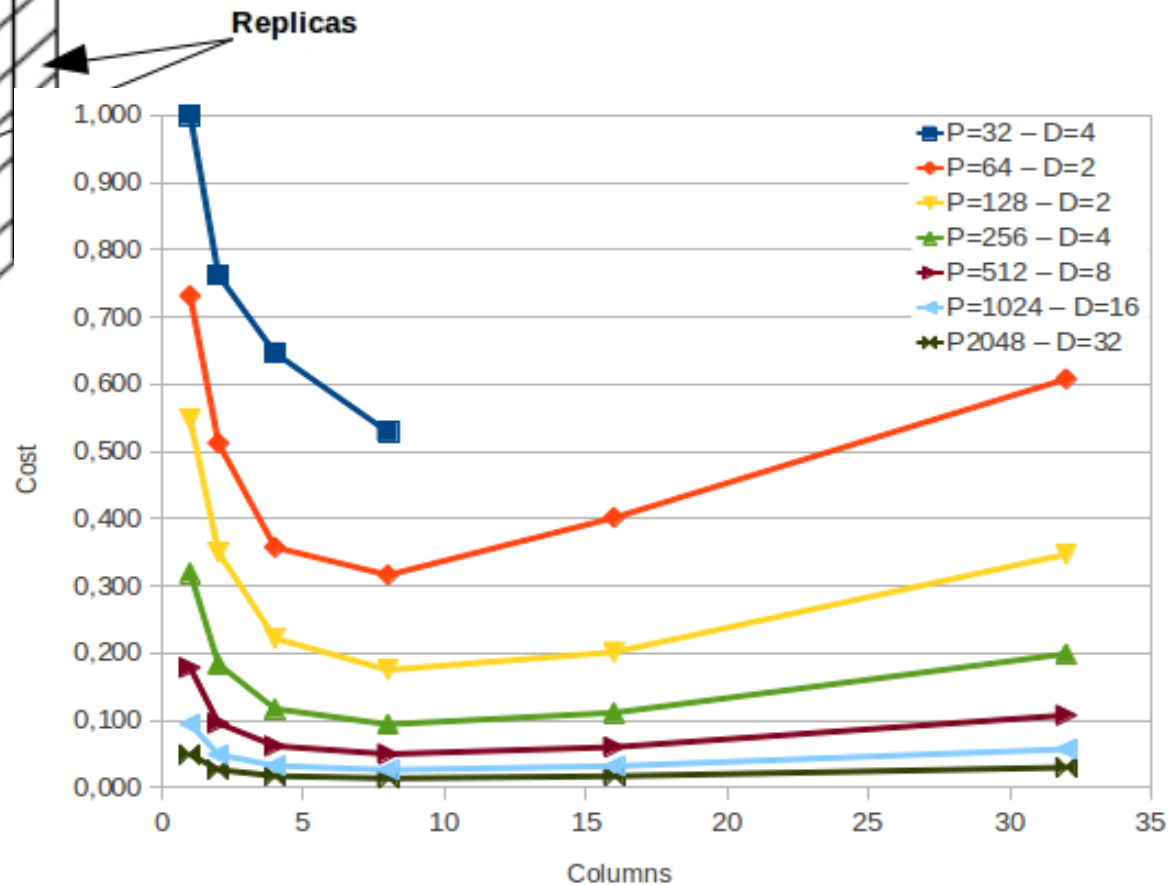
	0	1	2	3
0	p0	p8	p16	p24
1	p1	p9	p17	p25
2	p2	p10	p18	p26
3	p3	p11	p19	p27
4	p4	p12	p20	p28
5	p5	p13	p21	p30
6	p6	p14	p22	p31
7	p7	p15	p23	p32

- Para cada número  $P$  de nodos existe una configuración de  $R \times C$  que  $\min(\text{cost}(Q))$
- Tradeoff entre el overhead de comunicación y el tiempo de cómputo

# Índice 3D [Feuerstein, Europar 2012]

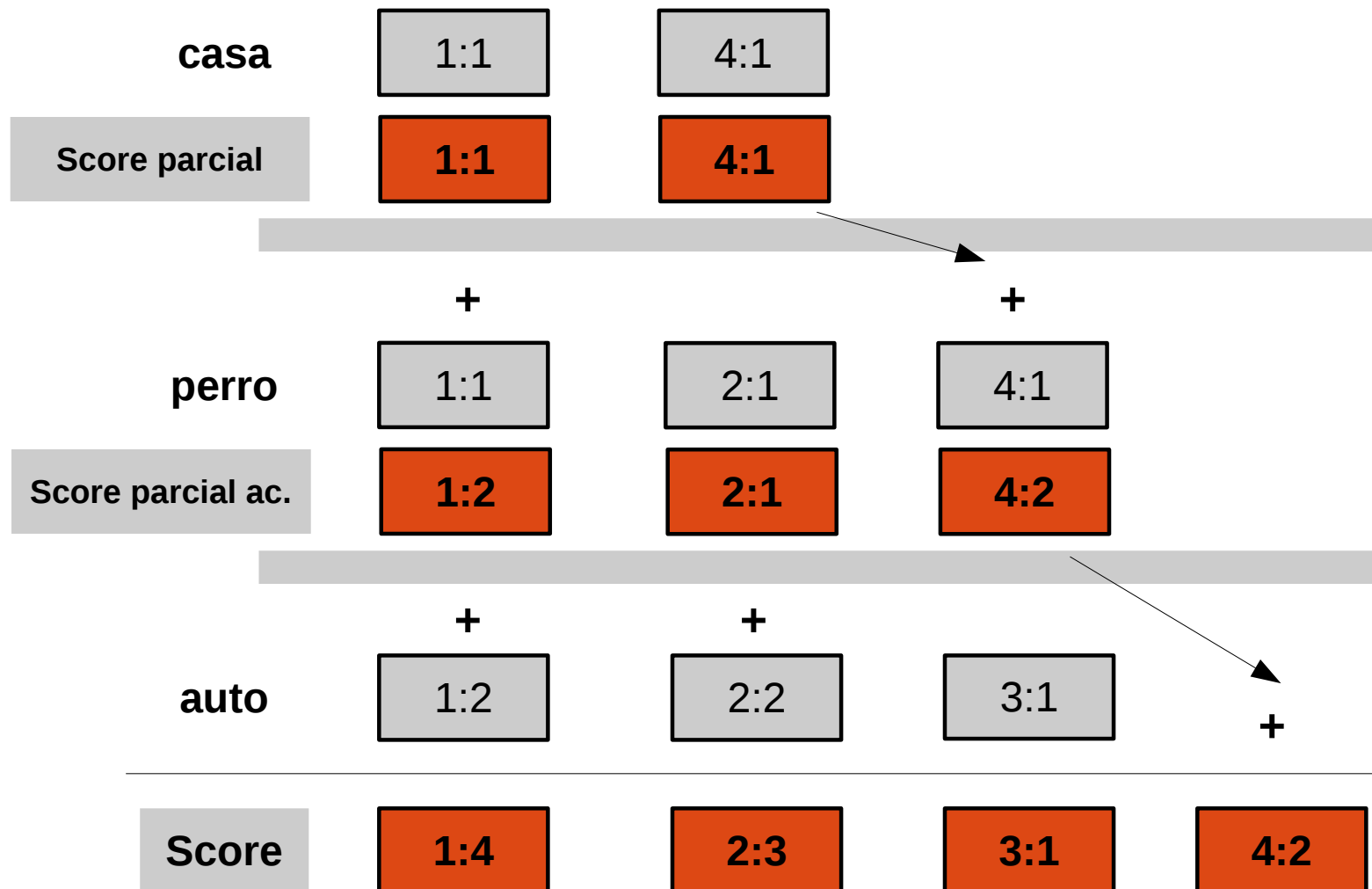


- Simulación con  $P = [32-2048]$
- Colección web de Yahoo! (UK)



# Estrategias de Búsqueda

- Term-at-a-Time (TAAT)



# Estrategias de Búsqueda

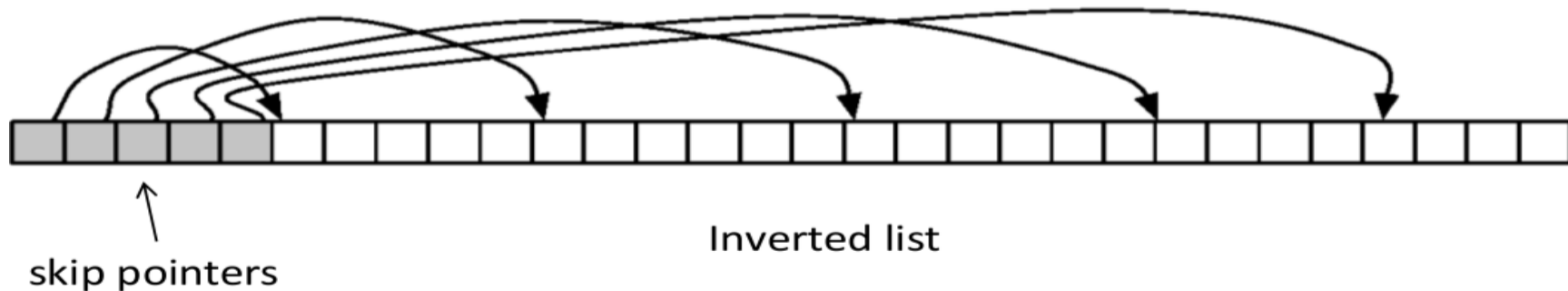
- Document-at-a-Time (DAAT)

casa	1:1				4:1
perro	1:1	2:1			4:1
auto	1:2	2:2	3:1		
Score	1:4	2:3	3:1		4:2

# Estrategias de Búsqueda

- **Más sofisticadas:**
  - SAAT, WAND [Broder, 2003]
- **Otras técnicas de optimización**
  - **Leer menos datos**
    - Poda de listas
    - Early-Termination
    - Skipping

{ (17,3) (34,6) (45,9) (52,12) (89,15) (101,18) }



# Ranking

- No alcanza solo con  $REL(q, d_i)$  para “aproximar” la intención del usuario.
- Algunas “señales”
  - Relevancia
  - Autoridad
  - Frescura
  - Preferencia
- Dependientes/independientes de la consulta  $q$





# Ranking

- **Dependientes del query**
  - Texto: documento, “anchors”, URLs
  - Historia de Clicks (Q2P, query to picks)
  - ...
- **Independientes del query**
  - Popularidad (links) → Análisis de Enlaces (PageRank)
  - Análisis de Spam
  - Propiedades de la página o el sitio
  - ...



# Ranking

- **Dependientes del query**

- Texto: document
- Historia de Clicks
- ...

- **Independientes**

- Popularidad
- Análisis de Sp
- Propiedades de n
- ...

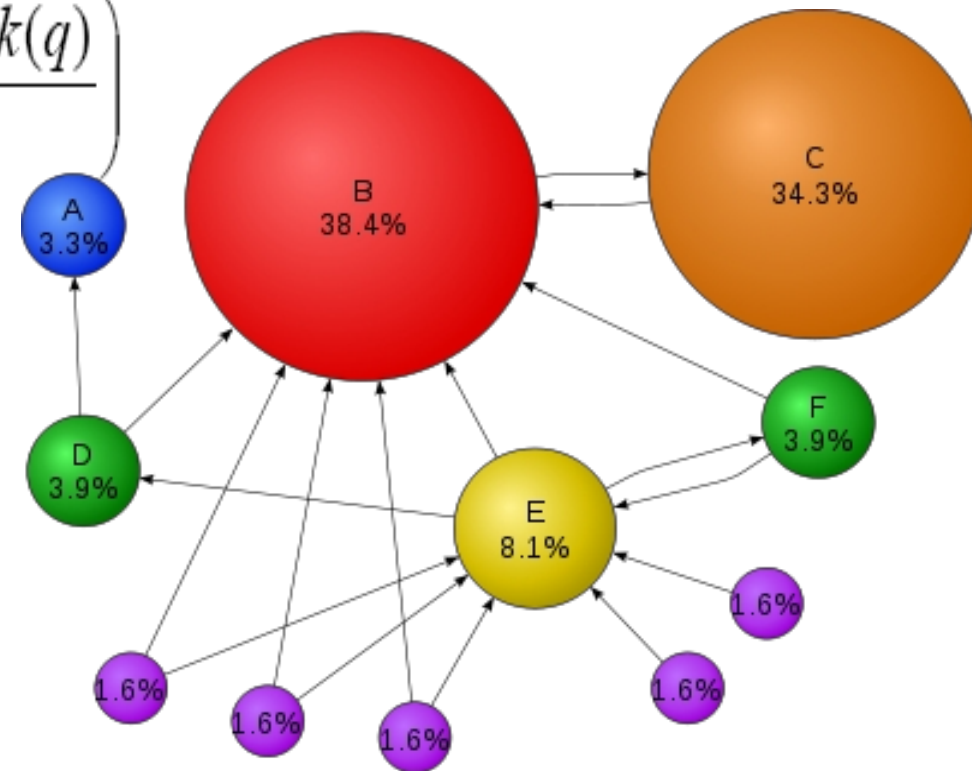
**De acuerdo a Matt Cutts  
(ingeniero de Google)  
Se utilizan más de  
200 variables!!!**

(PageRank)

# Pagerank [Brin & Page, 1998]

- “Una página es importante si otras páginas importantes apuntan a ésta”
- Cada link entrante es un voto
- Random Walk sobre el grafo web

$$PageRank(p) = (1-d) + d \times \sum_{\text{all } q \text{ linking to } p} \left( \frac{PageRank(q)}{c(q)} \right)$$





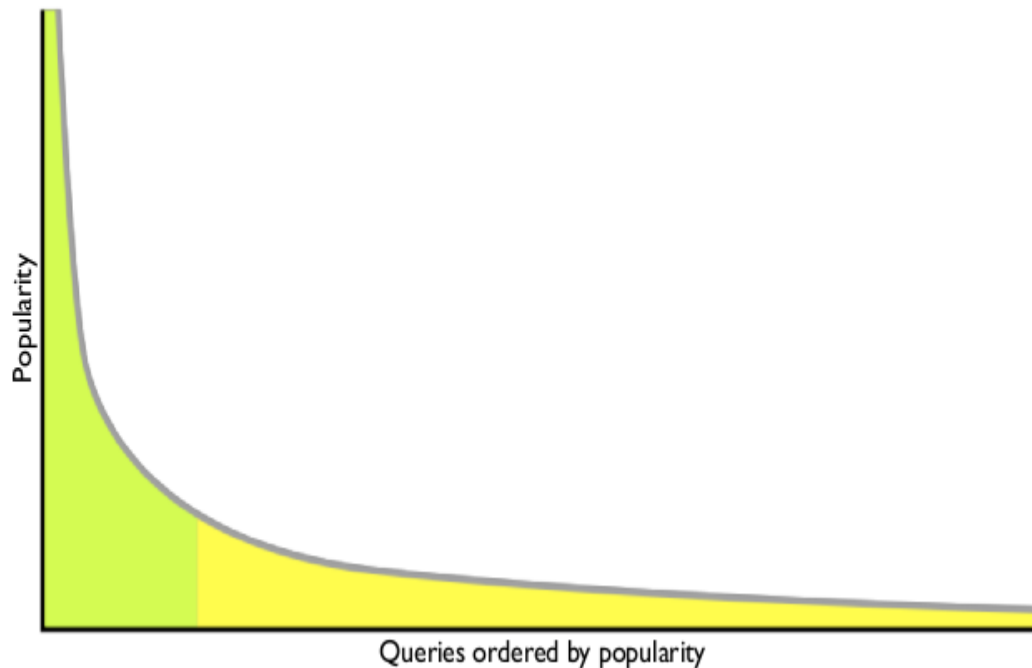
# Más allá de Pagerank

- **Personalización**
  - Ubicación geográfica
  - Historial de búsqueda
  - Conexiones Sociales?
- **AIR**
  - Panda (2014)
  - Penguin (2013)



# Caching en Motores de Búsqueda

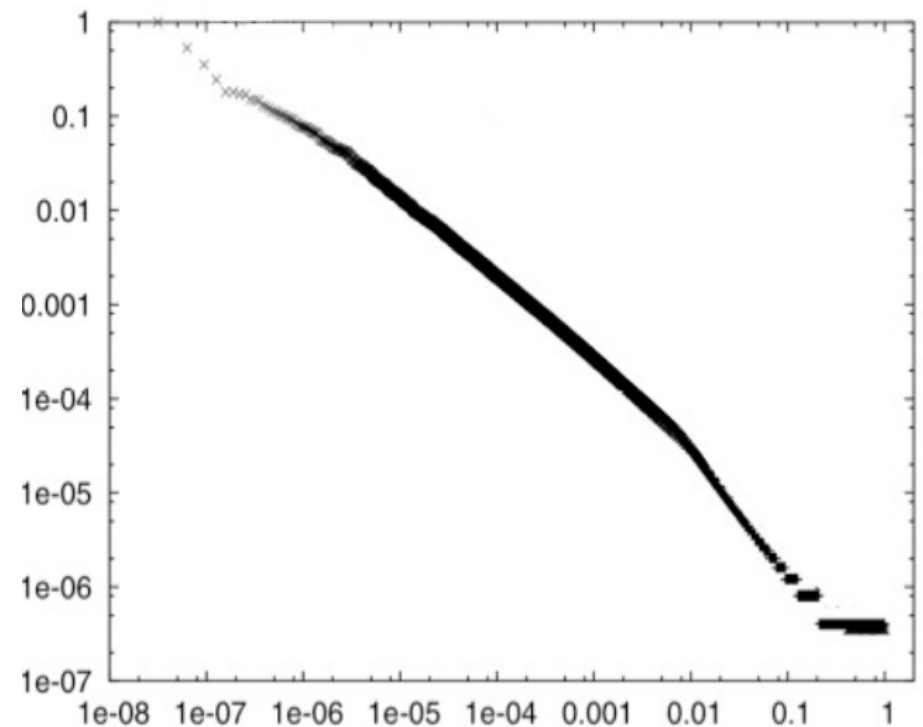
# Caching



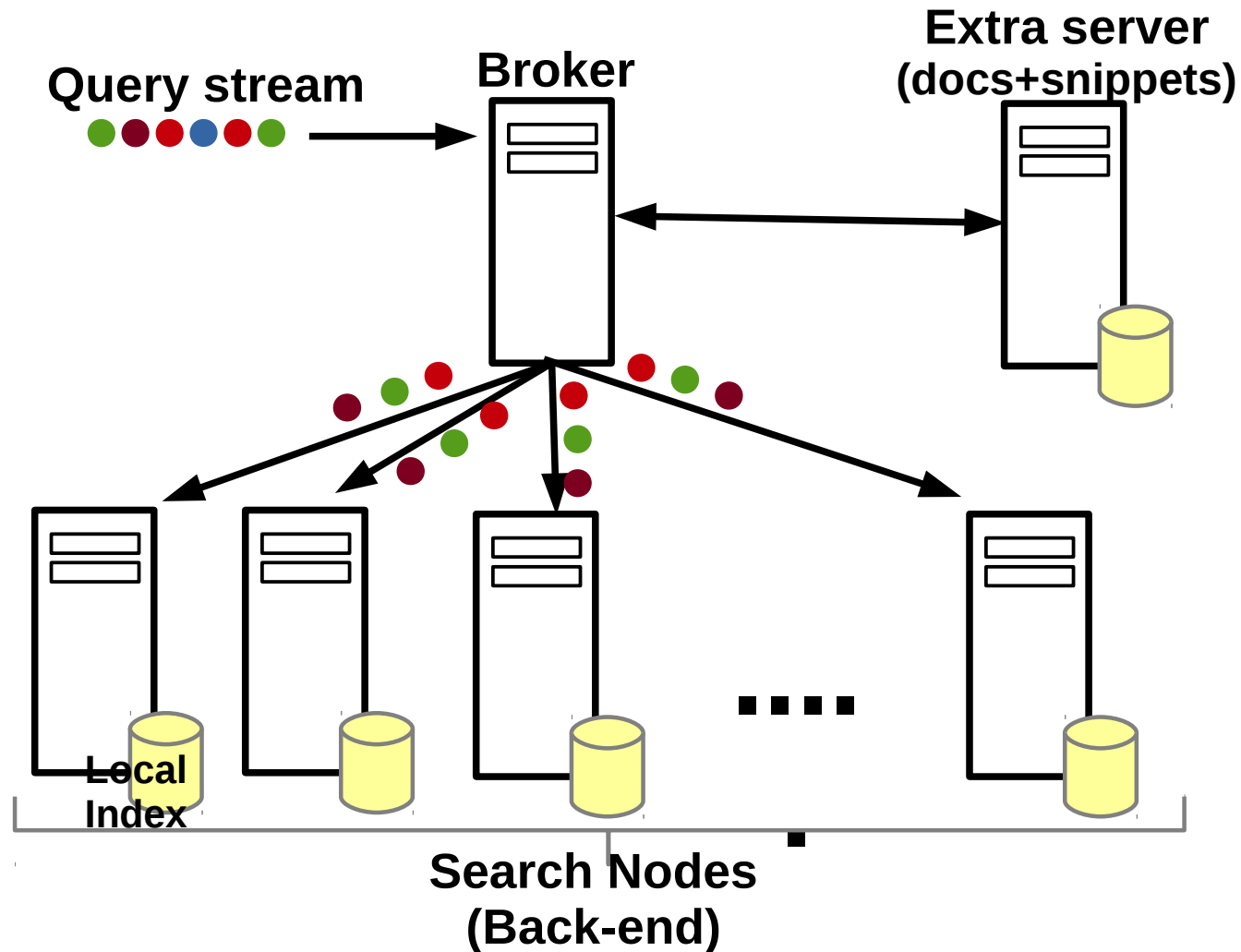
- Explotar localidad de los datos
- La frecuencia de aparición de los queries sigue una power-law

- Ejemplo:

Yahoo! Log File  
[Baeza-Yates, 2008]

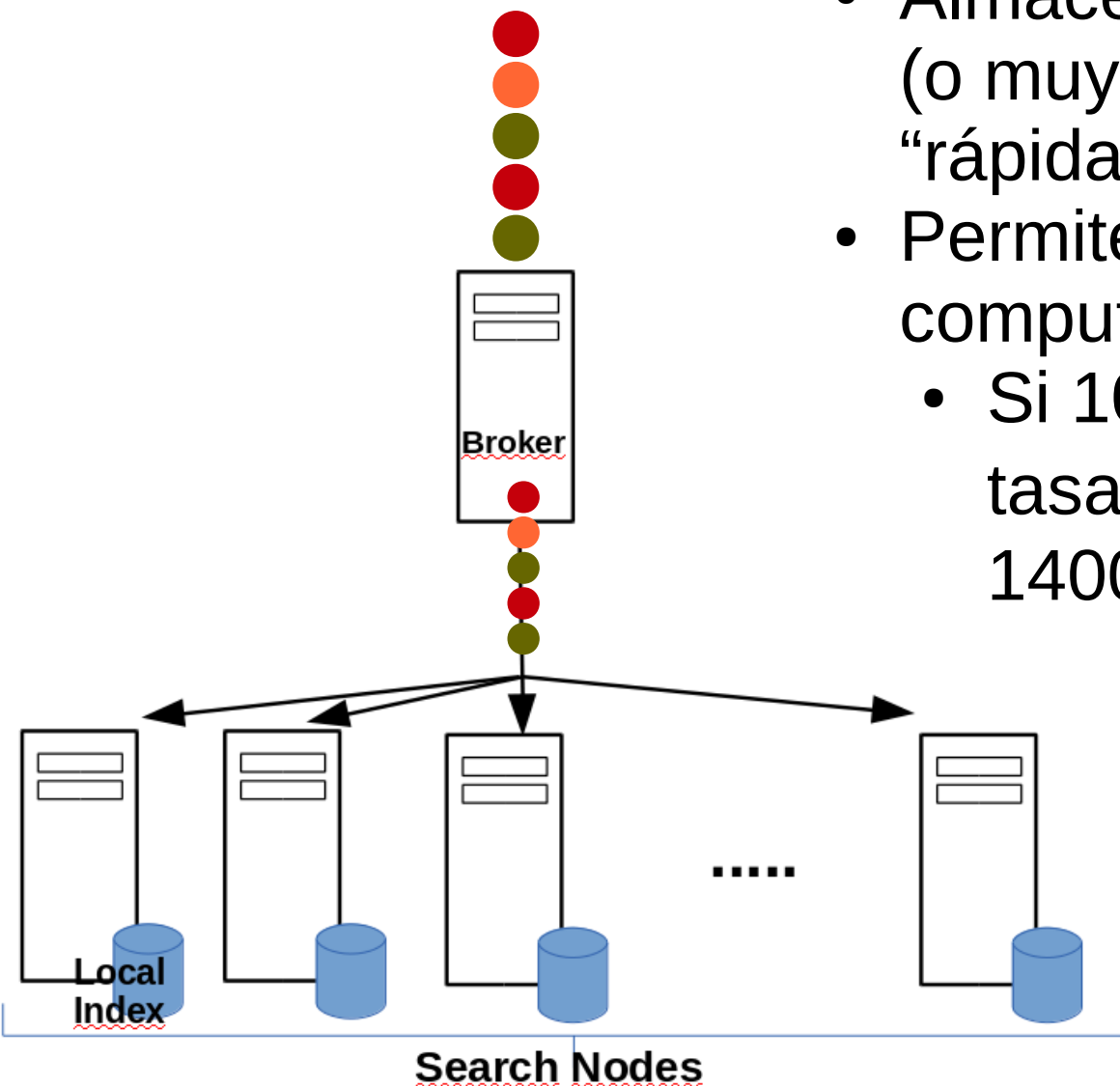


# Arquitectura de un MB (recap)



# Caching

- Almacenar items muy frecuentes (o muy costosos) en memorias “rápidas”
- Permite ahorrar recursos computacionales significativos
  - Si 1000 queries/seg. y una tasa de acierto del 30% → 1400 q/s



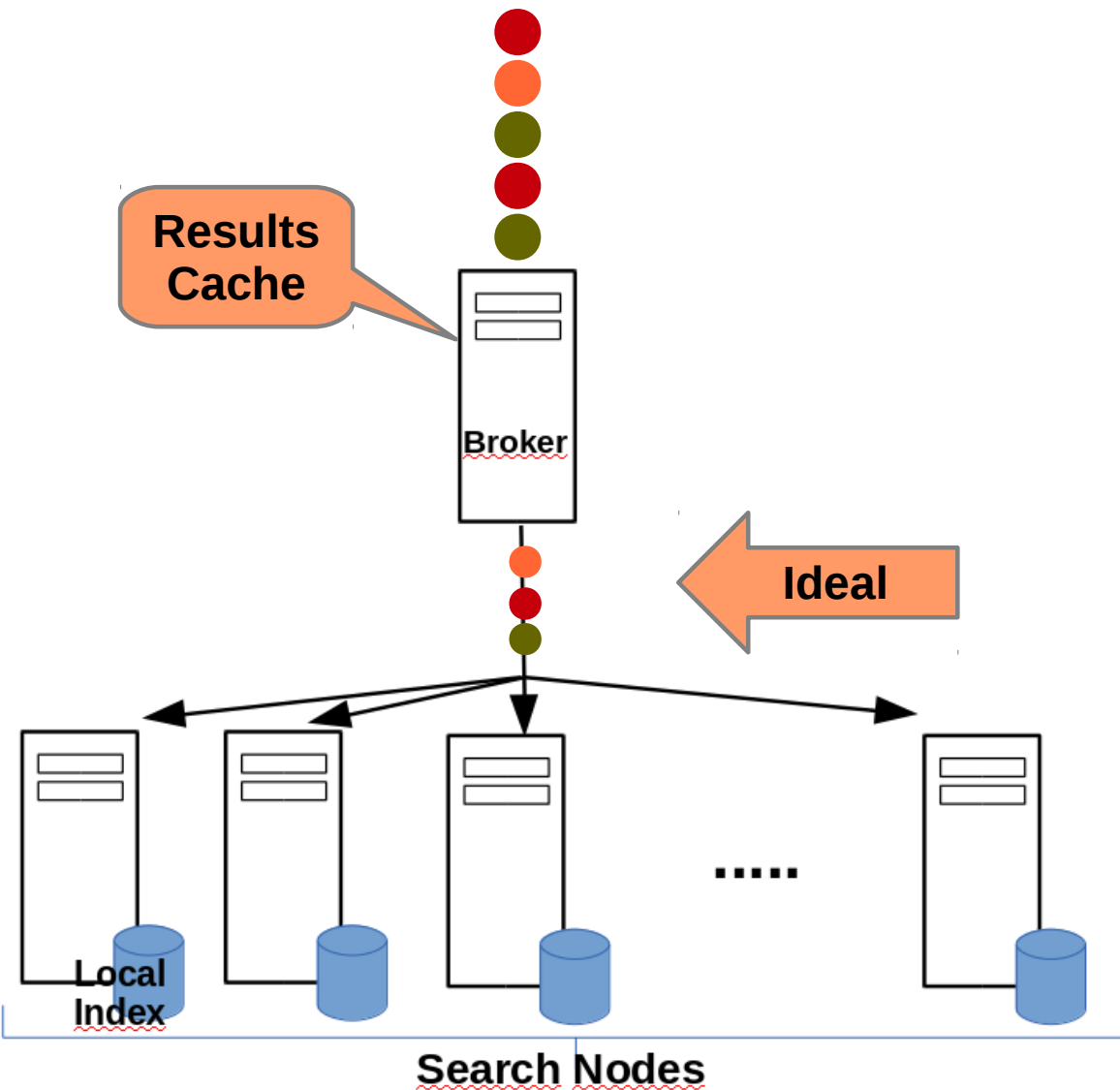
- **Métricas:**
  - Hit Ratio
  - Benefit (ahorro)



# Caching

## Queries

- 45-50% aparecen solo una vez
- Caché infinito: 50% de hit rate

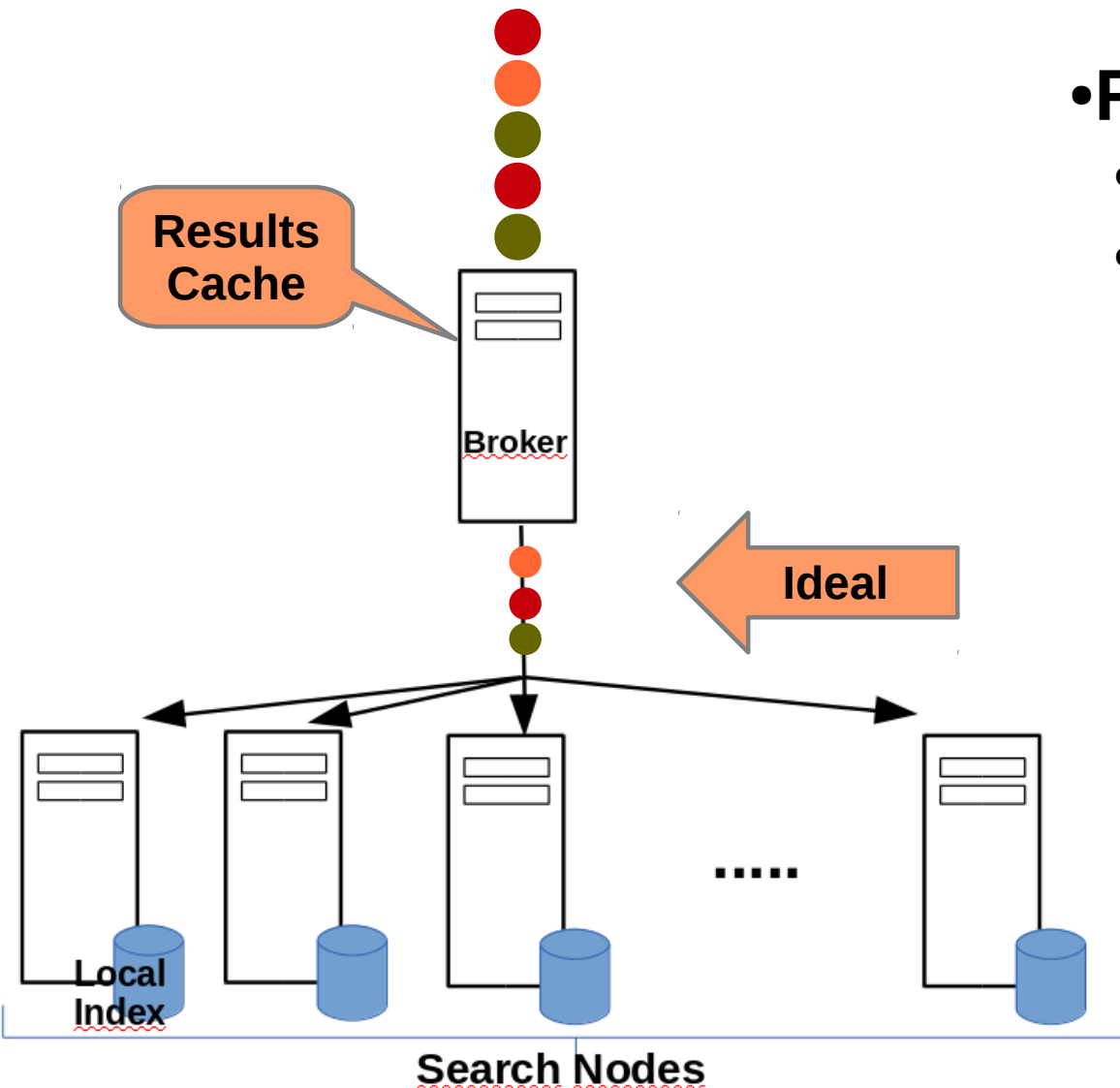


# Caching

## Queries

- 45-50% aparecen solo una vez
- Caché infinito: 50% de hit rate

- **RCache** [Markatos, 2001]
  - Análisis de un query log
  - LRU

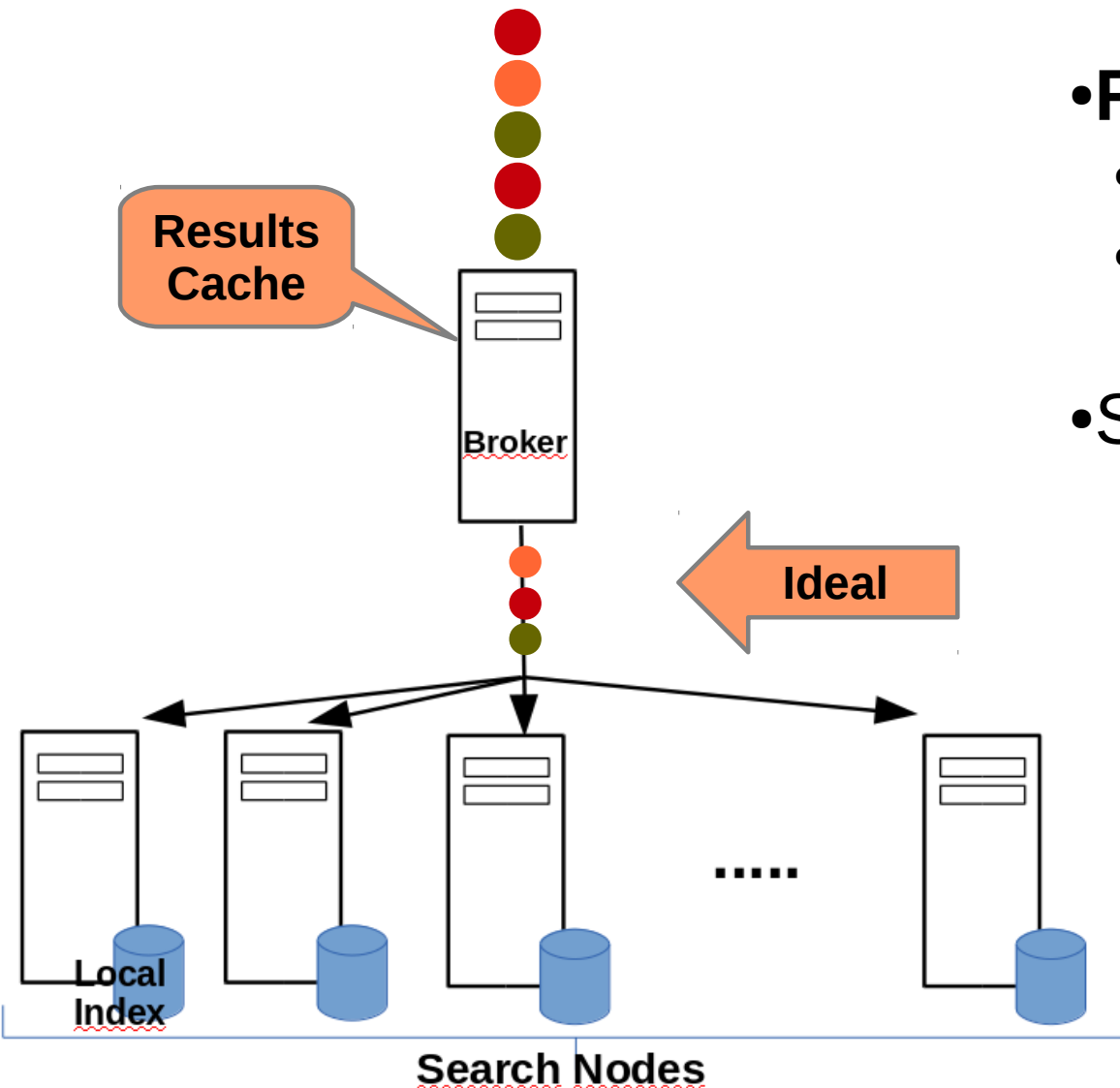


# Caching

## Queries

- 45-50% aparecen solo una vez
- Caché infinito: 50% de hit rate

- **RCache** [Markatos, 2001]
  - Análisis de un query log
  - LRU
- **SDC** [Fagni, 2006]

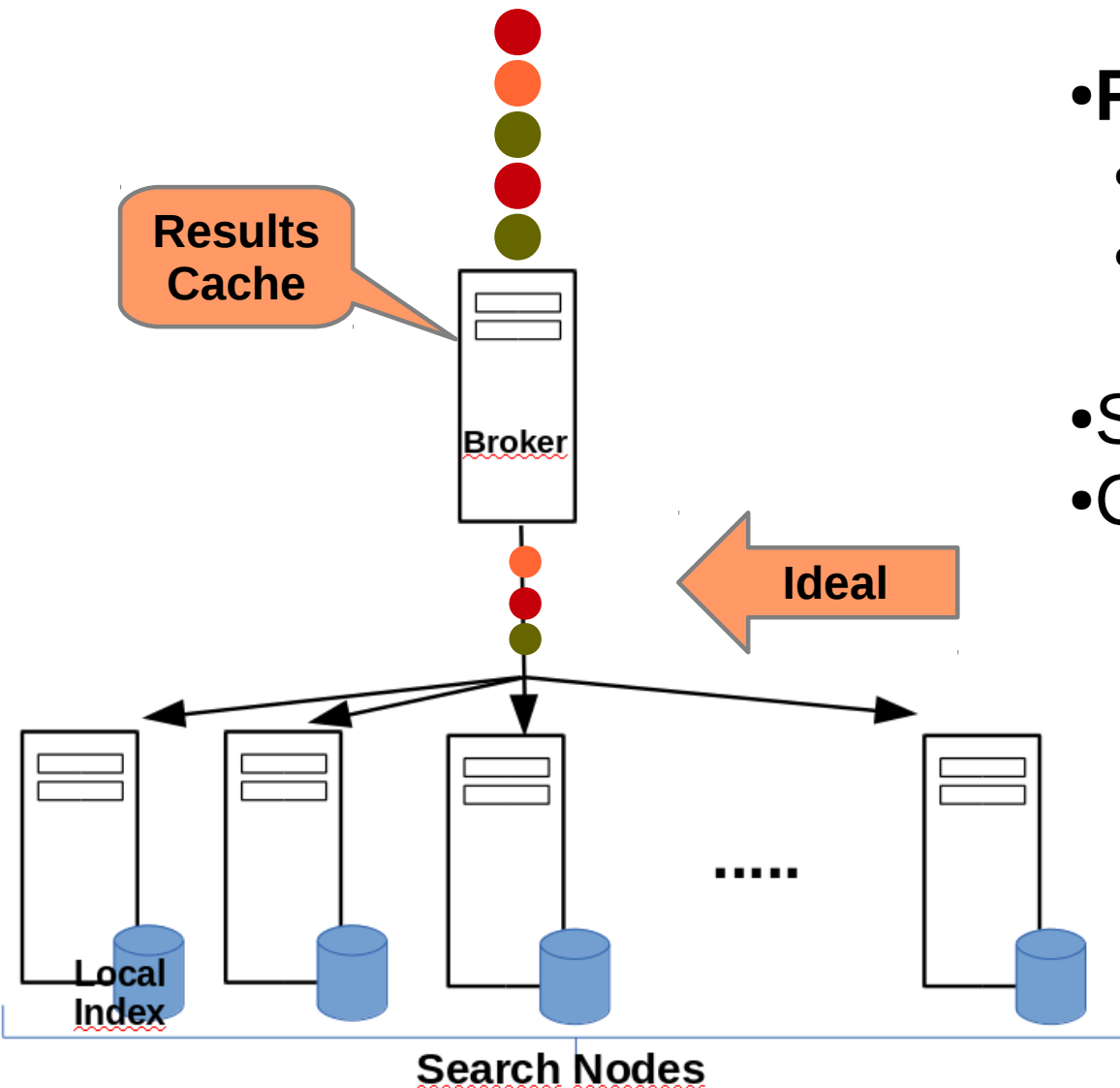


# Caching

## Queries

- 45-50% aparecen solo una vez
- Caché infinito: 50% de hit rate

- **RCache** [Markatos, 2001]
  - Análisis de un query log
  - LRU
- SDC [Fagni, 2006]
- Cost-Aware [Ozcan, 2011]



# Caching

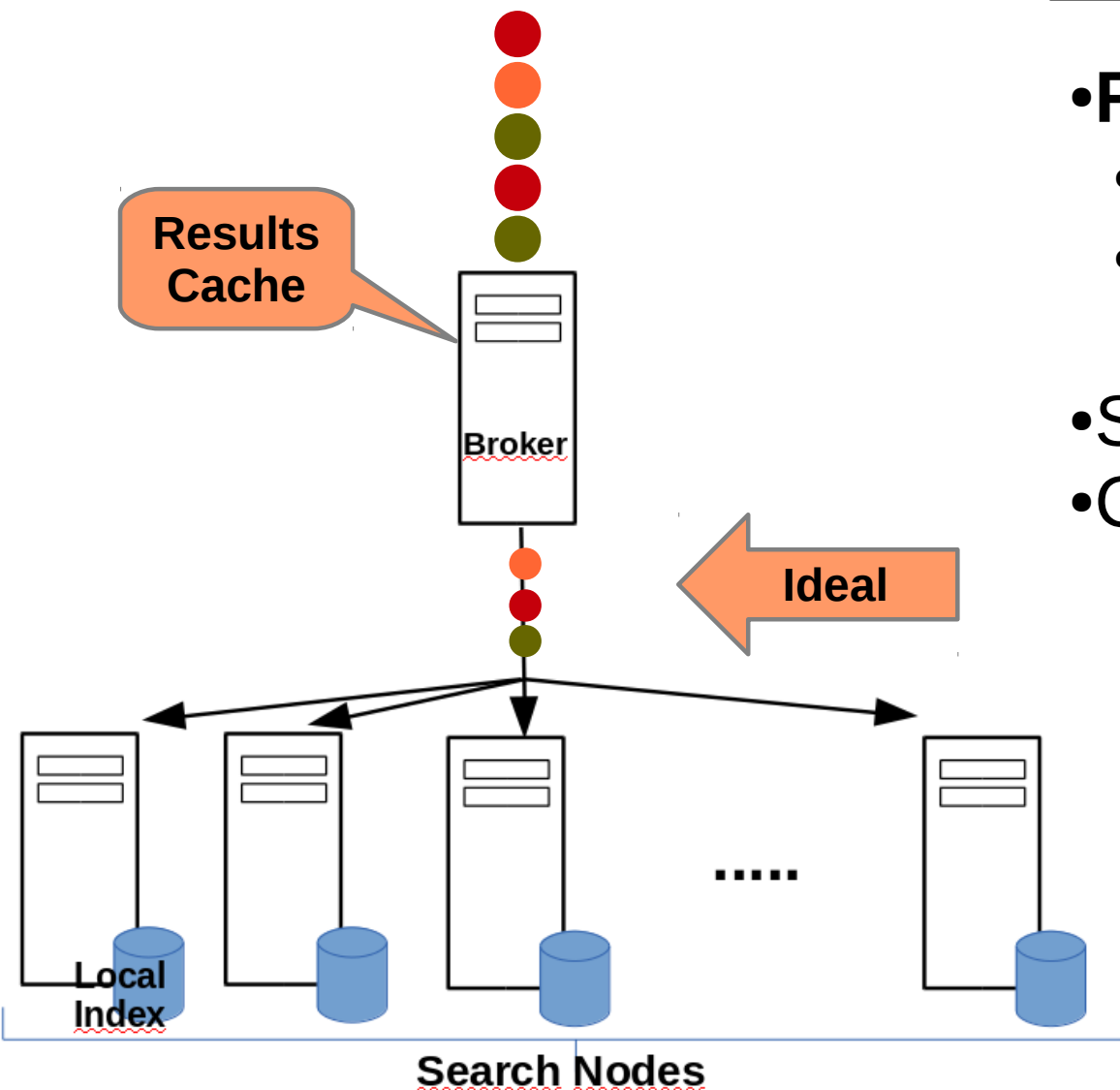
## Queries

- 45-50% aparecen solo una vez
- Caché infinito: 50% de hit rate

- **RCache** [Markatos, 2001]
  - Análisis de un query log
  - LRU
- SDC [Fagni, 2006]
- Cost-Aware [Ozcan, 2011]

## • Problema!

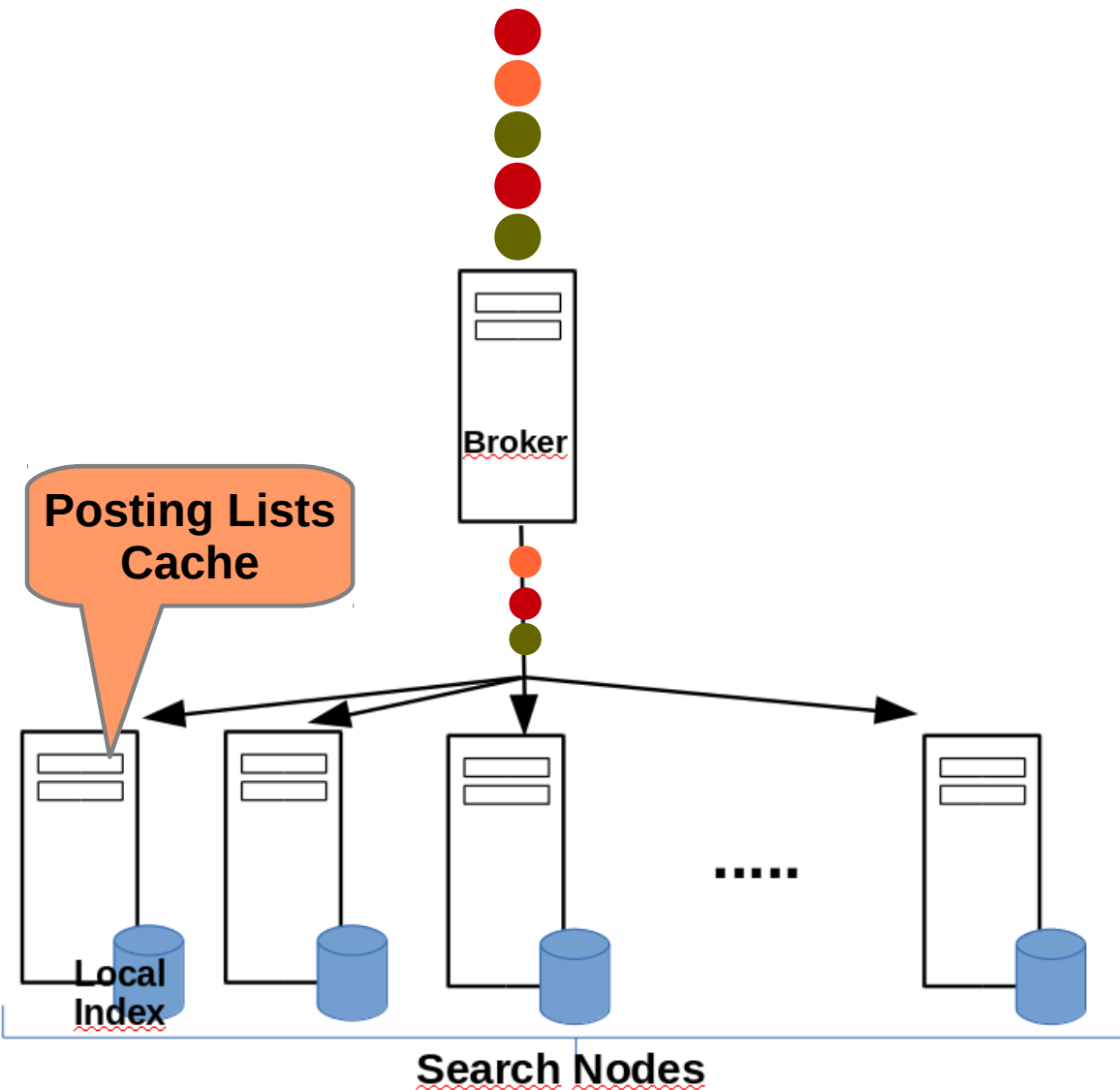
- Cache invalidation



# Caching

## Términos en Queries

- 5-10% aparecen solo una vez
- Caché infinito: 95% de hit rate

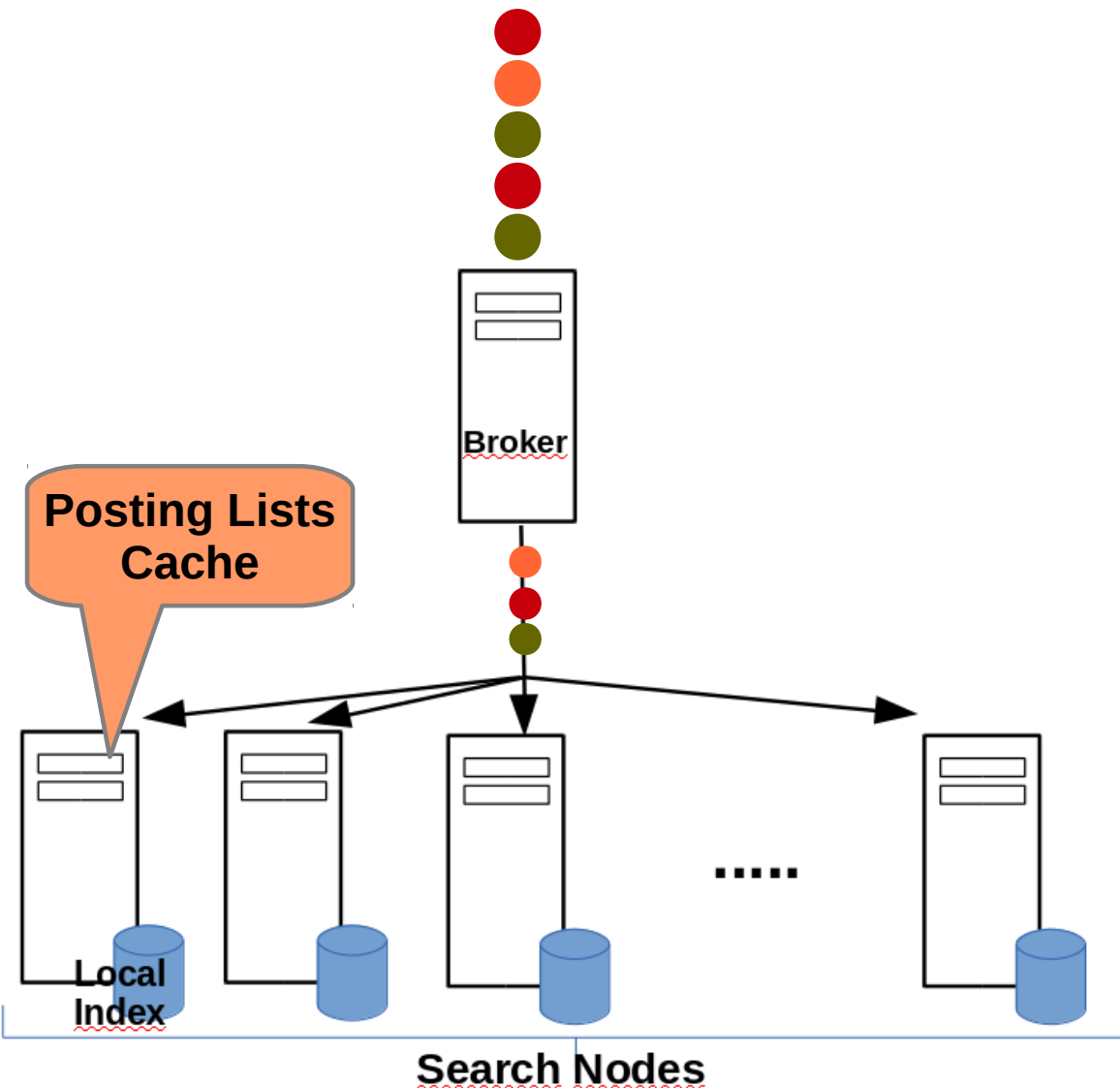


# Caching

## Términos en Queries

- 5-10% aparecen solo una vez
- Caché infinito: 95% de hit rate

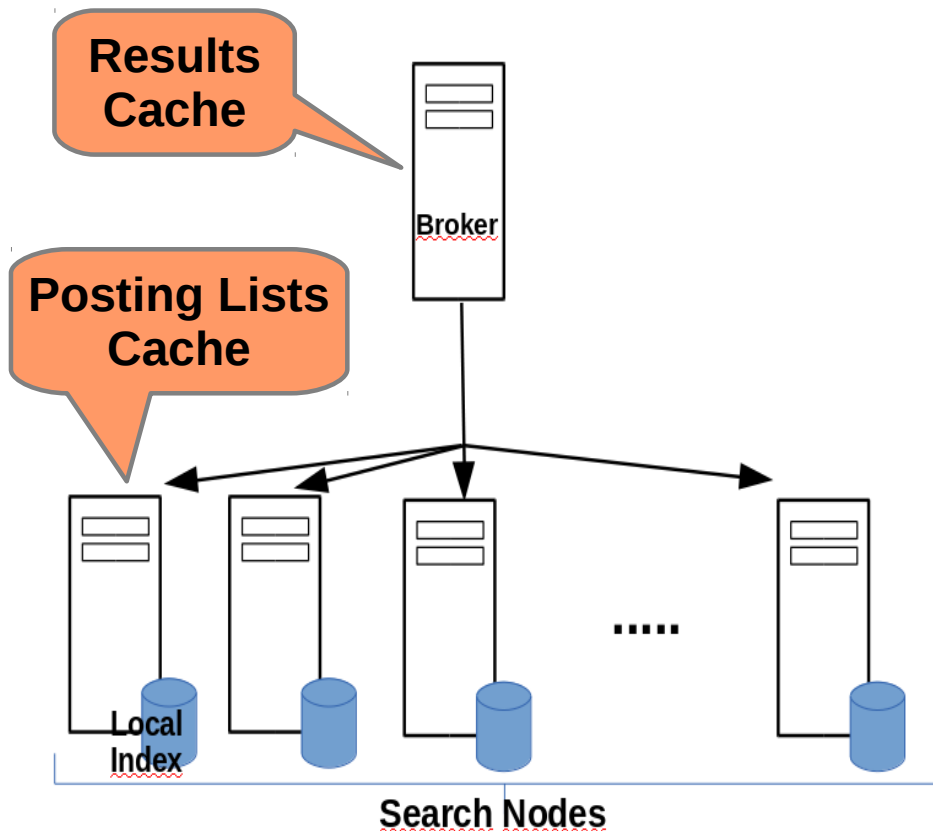
- Qué listas mantener en cache?
  - $Q_{TF}D_F$  [Baeza, 2003]
  - Tradeoff:  $f_q(t)$  y  $f_d(t)$ 
    - Términos con alta frecuencia son buenos
    - Listas muy largas ocupan mucho espacio





# Arquitecturas combinadas

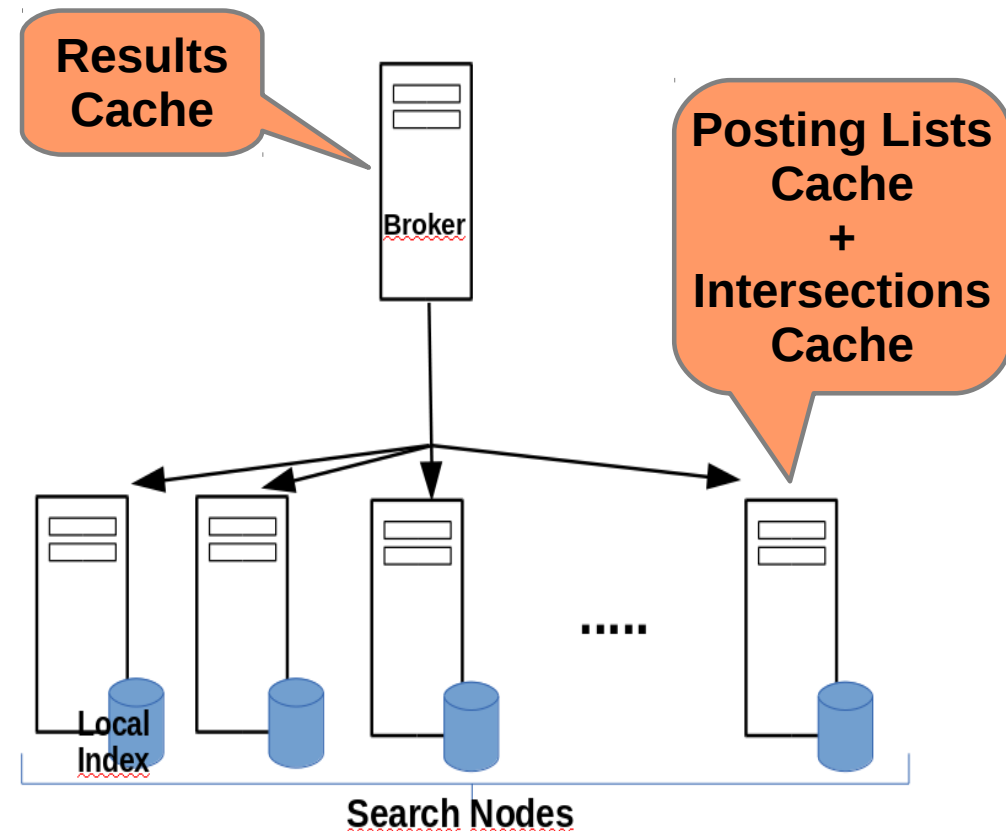
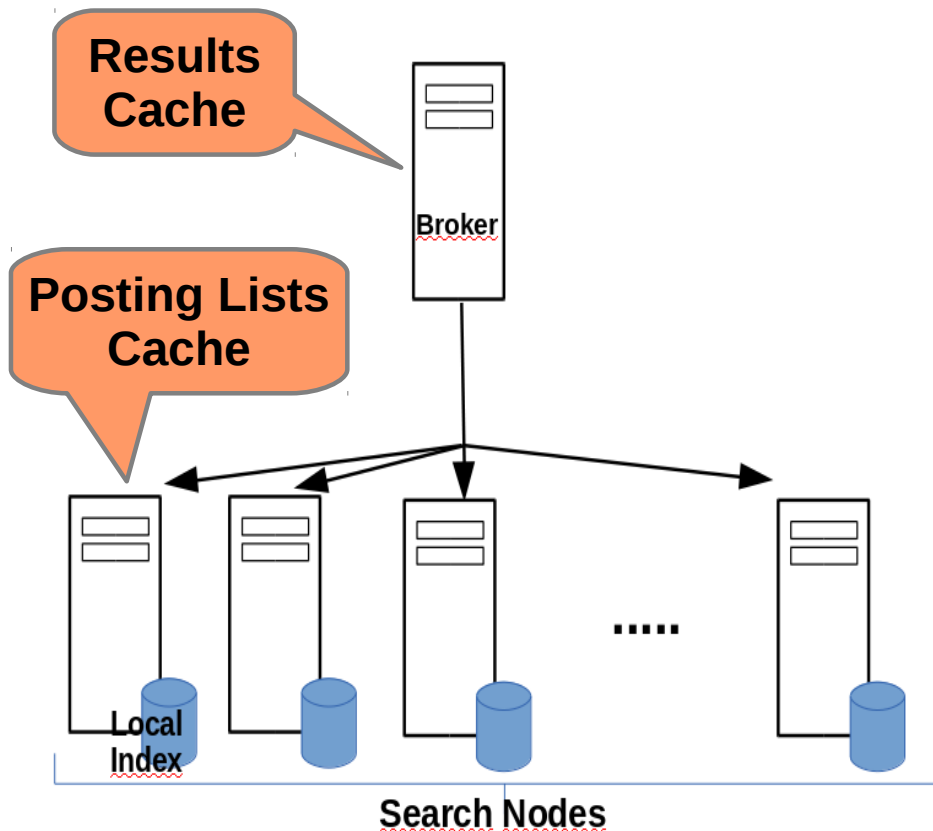
**2 Niveles** [Saraiva, 2001]



# Arquitecturas combinadas

**2 Niveles** [Saraiva, 2001]

**3 Niveles** [Long, 2005]





# Intersections Cache

- **Consultas parciales (partes)**
  - Para  $q = \{\text{sistemas operativos modernos}\}$  solo almaceno “**sistemas operativos**”
  - Qué ahorro?
- **Cuestiones**
  - Frecuencia (Hit rate) vs costo (Benefit)
  - Tamaños variables



# Cost-Aware Intersections Cache

- **Con el índice en disco** [Feuerstein, 2013]
- **3 Estrategias**
  - $Q = \text{"casa perro sopa tero"}$ 
    - $S1: (((\text{casa} \cap \text{perro}) \cap \text{sopa}) \cap \text{tero})$
    - $S2: (\text{casa} \cap \text{perro}) \cap (\text{sopa} \cap \text{tero})$
    - $S3: (\text{casa} \cap \text{perro}) \cap (\text{perro} \cap \text{sopa}) \cap (\text{sopa} \cap \text{tero})$
- **Diferentes políticas**
  - Estáticas
  - Dinámicas



# Cost-Aware Intersections Cache

- **Con el índice en disco** [Feuerstein, 2013]
  - **3 Estrategias**
    - $Q = \text{"casa perro sopa tero"}$ 
      - $S1: (((\text{casa} \cap \text{perro}) \cap \text{sopa}) \cap \text{tero})$
      - $S2: (\text{casa} \cap \text{perro}) \cap (\text{sopa} \cap \text{tero})$
      - $S3: (\text{casa} \cap \text{perro}) \cap (\text{perro} \cap \text{sopa}) \cap (\text{sopa} \cap \text{tero})$
  - **Diferentes políticas**
    - Estáticas
    - Dinámicas
- **S3 mejor (cachés + grandes)**
  - **Estática mejor que dinámica**
  - **"Cost-aware" mejor que "Cost-Oblivious"**



# Cost-Aware Intersections Cache

- **Con el índice en memoria** [Feuerstein, 2014]
- **4 estrategias**
  - 3 anteriores + todas las combinaciones (chequeo)
    - S4: (casa  $\cap$  perro), (casa  $\cap$  sopa), (casa  $\cap$  tero), (perro  $\cap$  sopa), (perro  $\cap$  tero), (sopa  $\cap$  tero)
- **Diferentes políticas**
  - Estáticas
  - Dinámicas
  - Híbridas



# Cost-Aware Intersections Cache

- **Con el índice en memoria** [Feuerstein, 2014]
- **4 estrategias**
  - 3 anteriores + todas las combinaciones (chequeo)
    - S4: (casa  $\cap$  perro), (casa  $\cap$  sopa), (casa  $\cap$  tero), (perro  $\cap$  sopa), (perro  $\cap$  tero), (sopa  $\cap$  tero)
- **Diferentes políticas**
  - Estáticas
  - Dinámicas
  - Híbridas

**- S4 mejor estrategia**  
**- Dinámicas mejor que híbridas**

# Integrated Cache [Tolosa, 2014]

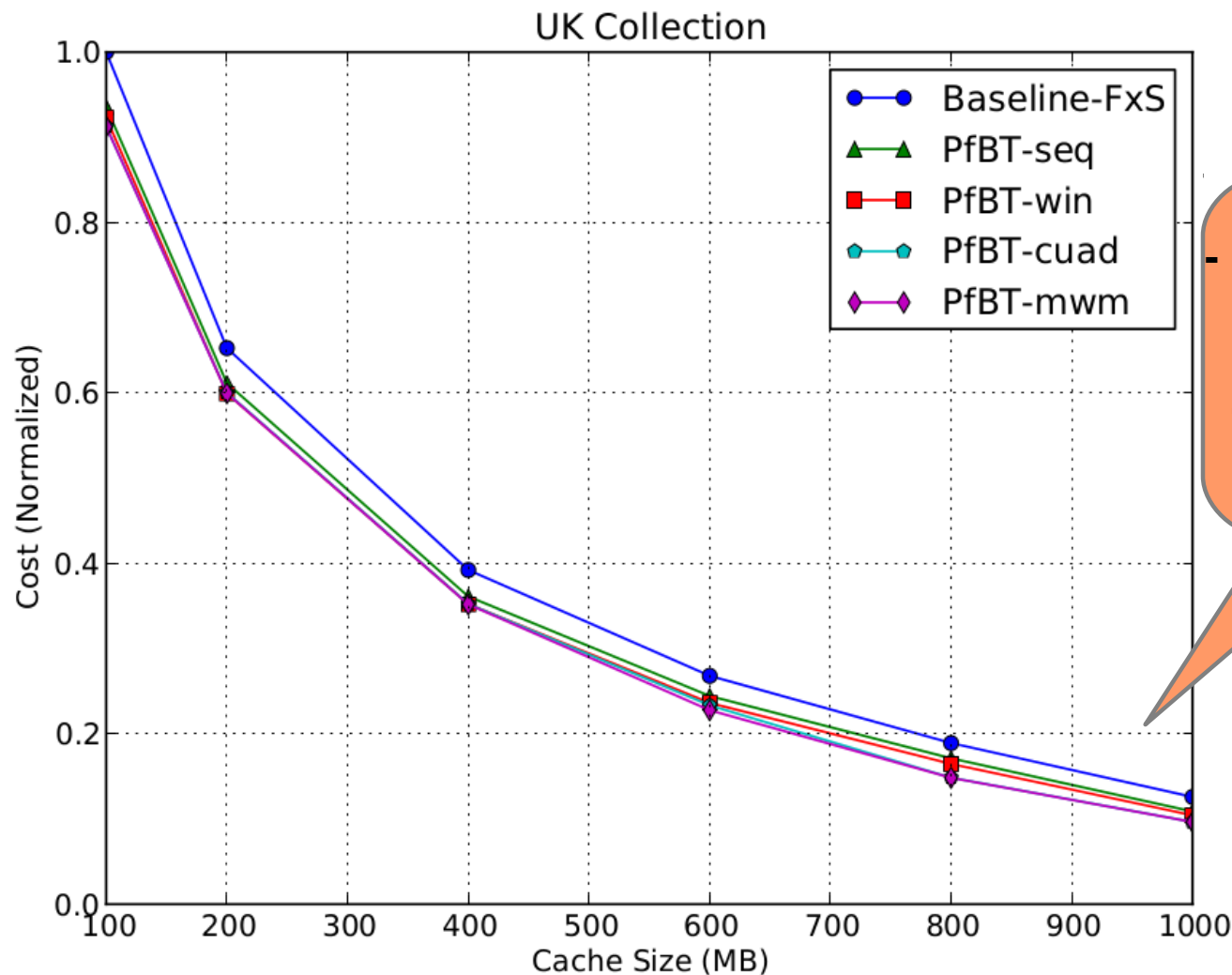
- List-caching + Intersection-Caching
- Cómo seleccionar los pares?
  - Greedy
  - MWM

Keys (pairs of terms)		Integrated Lists			
1	$t_1, t_2$	$\rightarrow$	$\ell_1 - (\ell_1 \cap \ell_2)$	$\ell_2 - (\ell_1 \cap \ell_2)$	$(\ell_1 \cap \ell_2)$
2	$t_3, t_4$	$\rightarrow$	$\ell_3 - (\ell_3 \cap \ell_4)$	$\ell_4 - (\ell_3 \cap \ell_4)$	$(\ell_3 \cap \ell_4)$
3	$t_1, t_5$	$\rightarrow$	$\emptyset$	$\ell_5 - (\ell_1 \cap \ell_5)$	$(\ell_1 \cap \ell_5)$
4	$t_3, t_5$	$\rightarrow$	$\emptyset$	$\emptyset$	$(\ell_3 \cap \ell_5)$



# Integrated Cache [Tolosa, 2014]

- vs List-Caching (cost-aware)



- Mejora en todos los casos

- Hasta un 30% (MWM)



**Finalizando...**



# Investigación en MB

- **Mix interesante entre ciencia e ingeniería**
  - Muchos problemas abiertos
    - Muchos **nuevos** problemas
- **Involucra muchas áreas de las CC**
  - Arquitecturas, Sistemas Distribuidos, Algoritmos, Paralelismo, Machine Learning, Minería de Datos (web), Datacenters, Interfaces...
- **Diversas Aplicaciones**
  - Búsquedas Web/Empresariales/Verticales
  - Redes Sociales (Cómo busca Twitter?)



**Preguntas?**





**Muchas Gracias!!!**

Gabriel H. Tolosa  
tolosoft@unlu.edu.ar