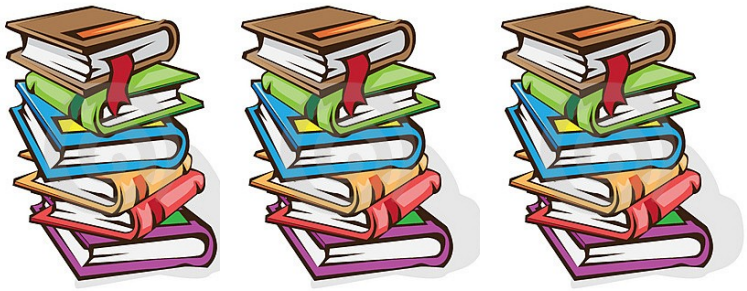
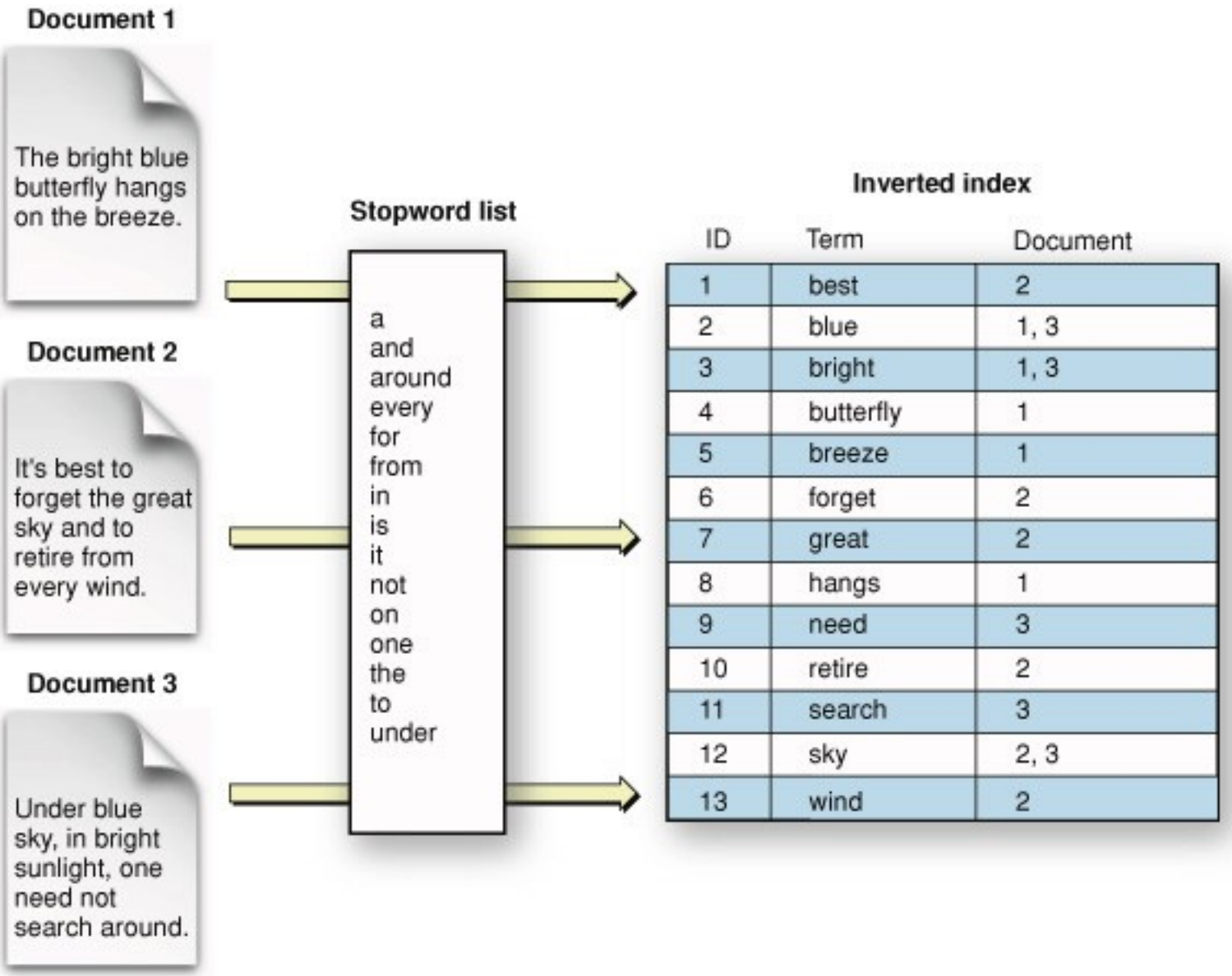
Construcción Índice Invertido Secuencial



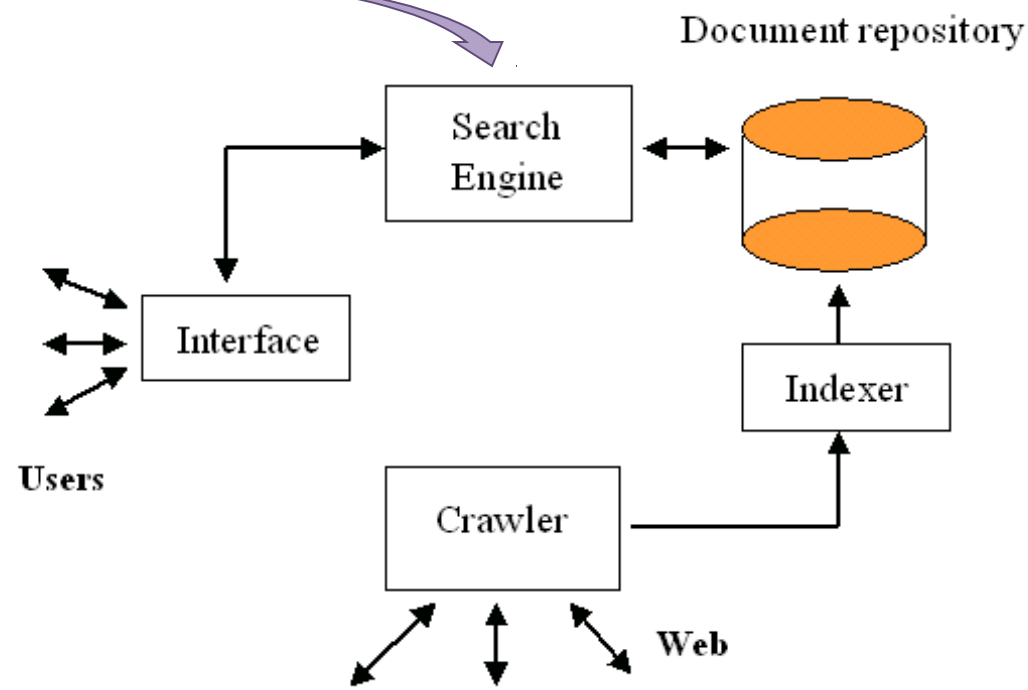
Colección de Texto

Eliminar stop-word, términos repetidos, tag, etc


Document 1
The bright blue butterfly hangs on the breeze.
Document 2
It's best to forget the great sky and to retire from every wind.
Document 3
Under blue sky, in bright sunlight, one need not search around.
Stopword list
a
and
around
every
for
from
in
is
it
not
on
one
the
to
under
Inverted index
ID
Term
Document
1 best 2
2 blue 1, 3
3 bright 1, 3
4 butterfly 1
5 breeze 1
6 forget 2
7 great 2
8 hangs 1
9 need 3
10 retire 2
11 search 3
12 sky 2, 3
13 wind 2

Resolver una transacción de lectura

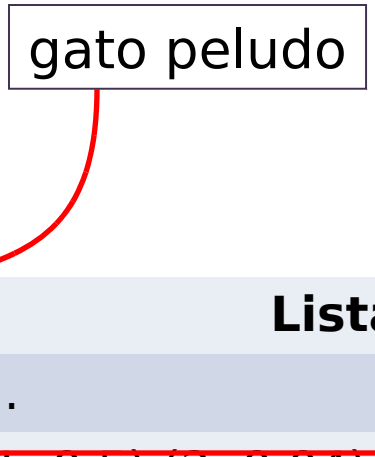
gato peludo
auto grande
terremoto chile
accidente fatal
....



Resolver una transacción de lectura: (1) Fetcher

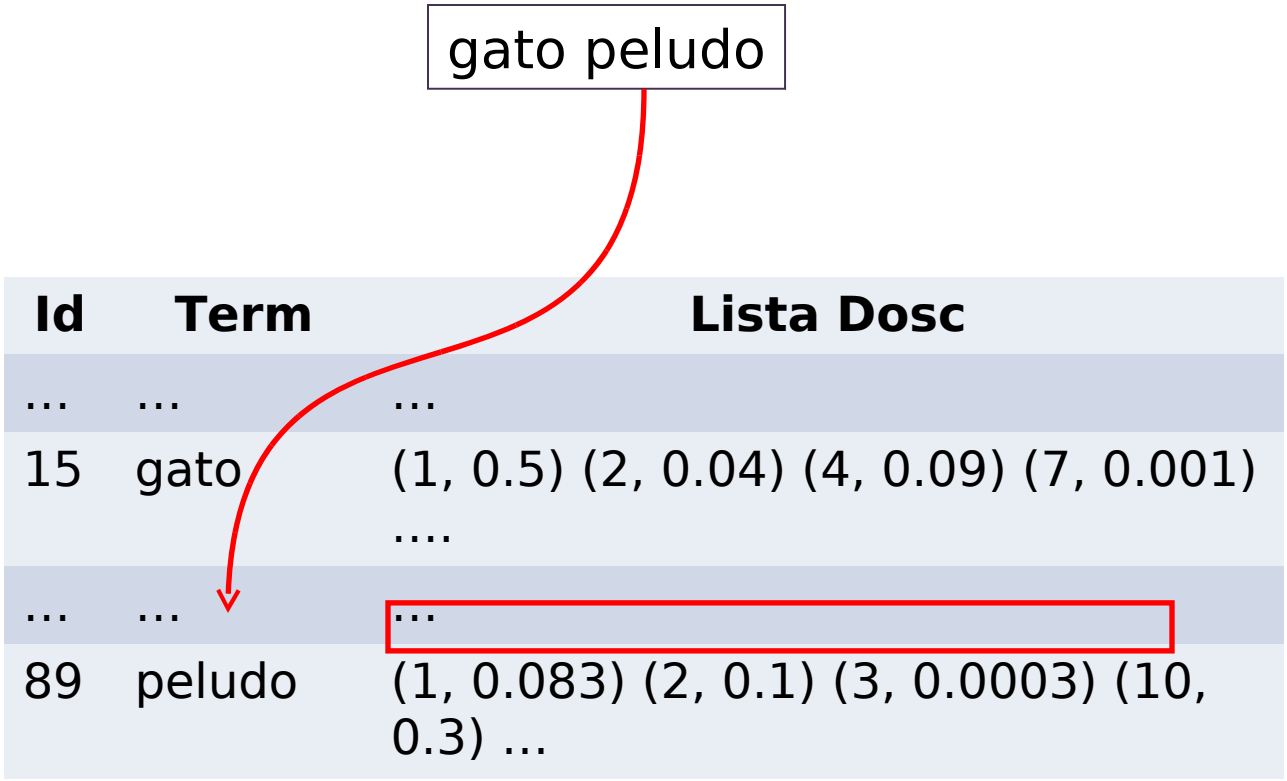
Por cada término de la transacción de lectura, se busca en el índice invertido la lista de documentos, que contiene el identificador del documento y la frecuencia de aparición, entre otros elementos.

Resolver una transacción de lectura: (1) Fetcher



Id	Term	Lista Dosc
...	...	...
15	gato	(1, 0.5) (2, 0.04) (4, 0.09) (7, 0.001) ....
...	...	...
89	peludo	(1, 0.083) (2, 0.1) (3, 0.0003) (10, 0.3) ...

Resolver una transacción de lectura: (1) Fetcher



Resolver una transacción de lectura: (2) Ranking

gato -> (1, 0.5) (2, 0.04) (4, 0.09) (7, 0.001)
peludo -> (1, 0.083) (2, 0.1) (3, 0.0003) (10, 0.3)

(1, 0.083) (2, 0.04) (4, 0.09) (7, 0.001) (10, 0.3)

Resolver una transacción de lectura tipo OR

. Fetcher de cada término

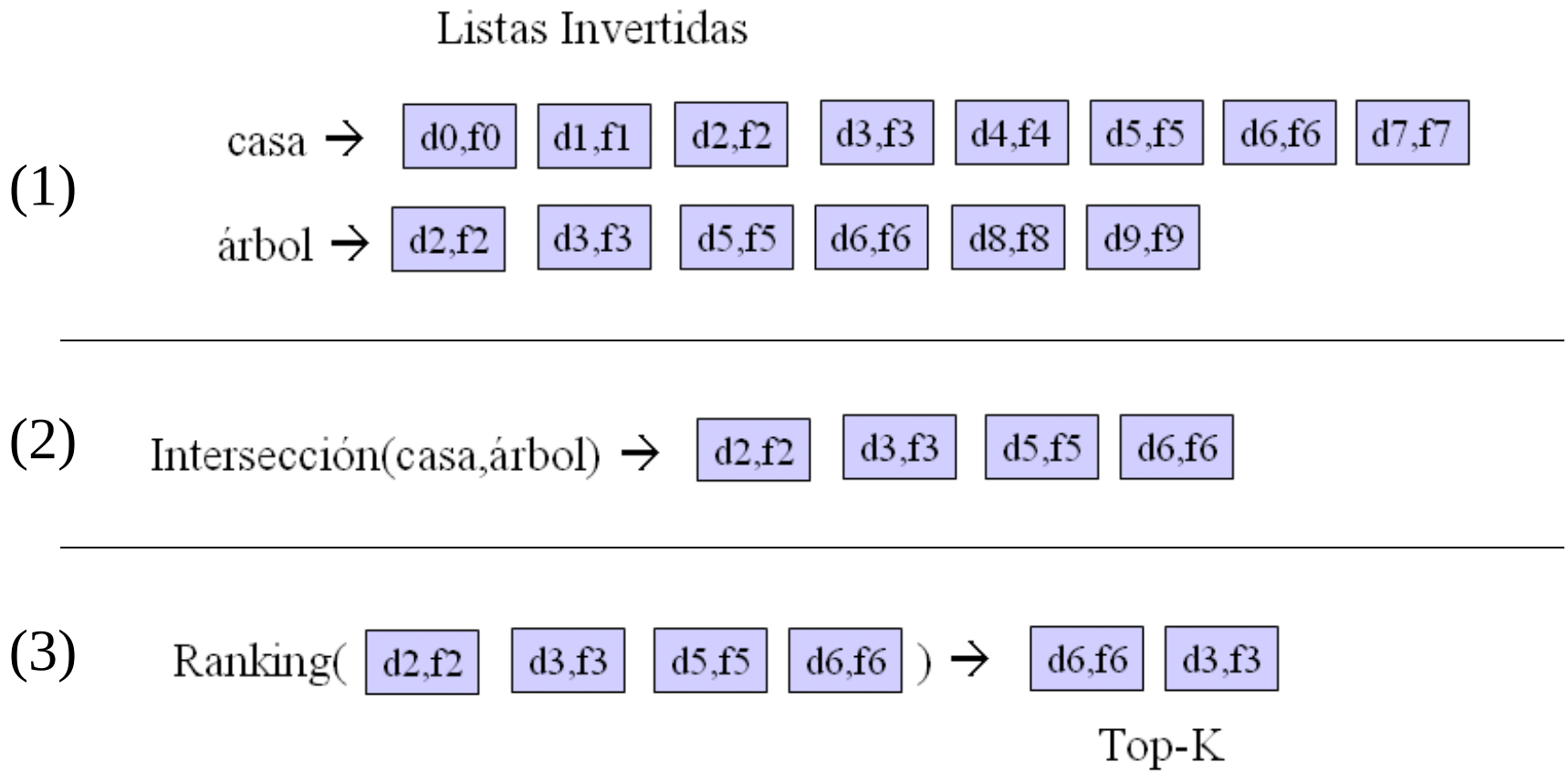
casa + k primeros elementos de su lista invertida
peludo + k primeros elementos de su lista invertida

. Ranker de la lista de cada término

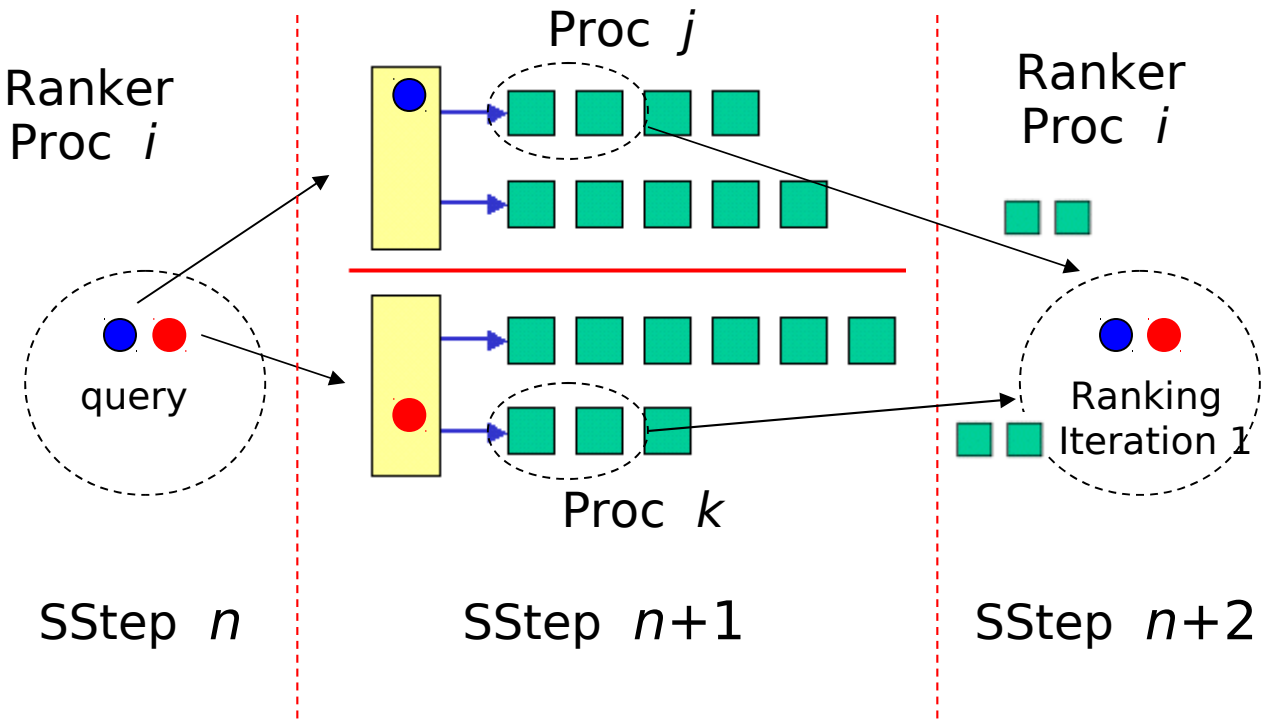
. Merge del resultado anterior

os índices pueden estar ordenados por id de documento o por frecuencia

Resolver una transacción de lectura tipo AND

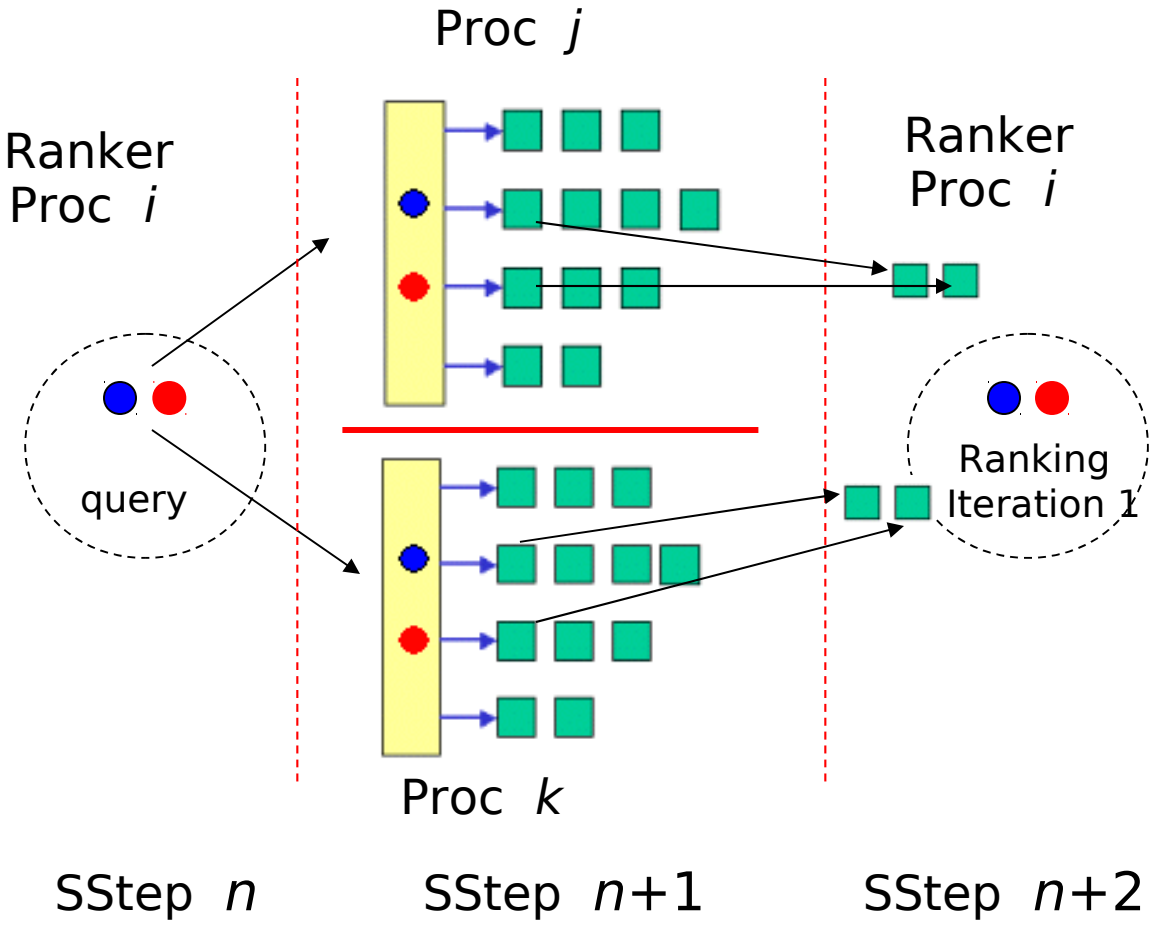


Indice distribuido por términos



Tipo And. El ranking de la consulta se ejecuta en el procesador que contiene las listas de los términos de la consulta.

Indice distribuido por documento



Tipo OR. Un procesador fetcher busca las listas involucradas que envia a un procesador ranker, el cual hace un merge de las listas y calcula el ranking de esa consulta. Es más eficiente para métodos q obligan a terminar antes la consulta y utilizan en promedio menos recursos de HW para resolver una consulta.

Métodos de Ranking: Método Vectorial

- Las consultas y documentos tienen asignado un peso para cada uno de los términos (palabras) de la base de texto (documentos).

- Los pesos se usan para calcular el grado de similitud entre cada documento almacenado en el sistema y las consultas que puedan hacer los usuarios.

- El grado de similitud calculado, se usa para ordenar de forma **decreciente** los documentos que el sistema devuelve al usuario, en forma de clasificación (ranking).

Métodos de Ranking: Método Vectorial

- Se define un vector que representa cada documento y consulta:
  - El vector dj está formado por los pesos asociados de cada uno de los términos en el documento dj.
  - El vector q está compuesto por los pesos de cada uno de los términos en la consulta q.

  Ambos vectores estarán formados por tantos pesos como términos se
  hayan determinado en la colección.

  El modelo vectorial evalúa el grado de similitud entre el documento dj y
  la consulta q, utilizando una relación entre los vectores dj y q.

Métodos de Ranking: Método Vectorial

Más parecidos, más cercanos
a 0 será el ángulo que formen, entonces
el coseno del ángulo se aproximará más
a 1.