Inverted indexes: Types and techniques

Inverted indexes

Los documentos se almacenan normalmente como listas de palabras, pero los índices invertidos invierten esto almacenando para cada palabra la lista de documentos en los que aparece la palabra, de ahí el nombre "*Inverted indexes*". Hay varias variaciones en los índices invertidos.

Almacenar la frecuencia total de cada palabra puede ser útil para optimizar los planes de ejecución de consultas.

Otra posible variación es si el léxico se almacena por separado o no. El léxico almacena todos los tokens indexados para toda la colección. Por lo general, también almacena información estadística para cada token, como el número de documentos en los que aparece.

Compression

Almacenar listas invertidas totalmente sin comprimir desperdicia enormes cantidades de espacio.

Hay muchas maneras de almacenar las listas en una forma más compacta. Se pueden dividir en dos categorías dependiendo de si el número de bits que utilizan para codificar un solo valor es siempre un múltiplo de 8 o no.

En lugar de utilizar 32 bits para almacenar cada valor, podemos usar enteros de longitud variable que solo usan tantos bytes como sea necesario.

Construction

Construir un índice invertido es fácil. Hacerlo sin usar cantidades obscenas de memoria, espacio en disco o tiempo de CPU es una tarea mucho más difícil. Los avances en hardware no ayudan mucho ya que el tamaño de las colecciones que se indexan está creciendo a un ritmo aún más rápido.

La inversión en memoria no es factible para colecciones grandes por lo que en esos casos tenemos que almacenar resultados temporales en disco.

Inverted index techniques

¿Cuáles son las técnicas necesarias para implementar un motor de búsqueda utilizando un índice invertido?

Document preprocessing

Los documentos normalmente no están indexados tal como están, pero son preprocesados primero. Se convierten en tokens en la fase de lexing, los tokens se transforman posiblemente en más genéricos en la fase de stemming, y finalmente algunos tokens se pueden eliminar por completo en la fase de eliminación de stop word.

Lexing

Se refiere al proceso de convertir un documento de una lista de caracteres a una lista de símbolos, cada uno de los cuales es una sola palabra alfanumérica.

Stremming

Es no indexar cada palabra como aparece después de la lexing, pero transformándola a su raíz morfológica (*stem*) y la indexación que en su lugar. Por ejemplo, las palabras "*computar*", "*computar*", "*computar*", "*computar*", "*computar*" y "*computar*" podrían ser indexadas como "*computar*".

Stop words

Son palabras como "*a*", "*the*", "*of*", y "*to*", que son tan comunes que casi todos los documentos las contienen. Una lista de palabras de parada contiene la lista de palabras a ignorar al indexar la colección de documentos.

Query types

Normal

Es cualquier consulta que el usuario no indica explícitamente que es una consulta especializada de uno de los tipos descritos más adelante en esta sección. Para consultas que contienen solo un término, la semántica deseada es clara: coincide con todos los documentos que contienen el término.

Boolean

Son consultas en las que los términos de búsqueda están conectados entre sí utilizando los diversos operadores disponibles en la lógica booleana, siendo los más comunes AND, OR y NOT.

Phrase

Se utilizan para encontrar documentos que contienen las palabras dadas en el orden dado. Por lo general, la búsqueda de frases se indica rodeando el fragmento de la oración entre comillas ("") en la cadena de consulta.

Proximity

Son de la forma *term1 NEAR(n) term2*, y deben coincidir con documentos donde *term1* ocurre dentro de n palabras de *term2*. Son útiles en muchos casos, por ejemplo cuando se busca el nombre de una persona nunca se sabe si un nombre aparece como "*Osku Salerma*" o "*Salerma, Osku*", por lo que podría utilizar la búsqueda **osku NEAR(1) salerma** para encontrar ambos casos.

Wildcard

Las consultas de *wildcards* son una forma de coincidencia borrosa o inexacta. Hay dos variantes principales:

- ***Whole-word wildcards:*** Donde las palabras enteras se dejan sin especificar. Por ejemplo, la búsqueda de París es la * capital del mundo coincide con documentos que contienen frases "París es la capital romántica del mundo", "París es la capital de la moda del mundo", "París es la capital culinaria del mundo", y o on.
- ***In-word wildcards:*** Donde parte de una sola palabra se deja sin especificar. Puede ser el final de una palabra (Helsin*), el comienzo de la palabra (*sinki), el centro de la palabra (Hel*ki) o alguna combinación de estas (*el*nki).