Motores de Búsqueda Web
Tarea Tema 2

José Alberto Benítez Andrades

71454586A

Motores de Búsqueda Web
Máster en Lenguajes y Sistemas Informáticos - Tecnologías del Lenguaje en la Web
UNED
30/01/2011

30 de enero de 2011

Tarea Tema 2

Enunciado del ejercicio

Se trata de razonar cuáles son los mecanismos esenciales con los que un buscador puede duplicar el tamaño de su índice, y cómo afectará a la rapidez con que se devuelven resultados en promedio.

1. Resolución

Introducción

Lógicamente, para saber cómo puede un buscador duplicar la capacidad de su índice, habrá que saber primero qué es un índice, y dado que eso está dentro de la estructura general del buscador, la describiremos y después nos centraremos en su índice, siendo el buscador elegido Google, ya que es el buscador más importante del mundo en estos momentos (En torno al 50% de cuota de mercado)

Topología de Red

Aunque no se conoce la cifra exacta, algunas personas estiman que Google mantiene unos 450000 servidores, localizados en concentradores en diferentes ciudades a lo largo del globo, con centros en California, Virginia, Atlanta, Dublín, etc. Cuando se realiza un intento de conexión con Google, los servidores DNS de Google realizan un equilibrio de carga para permitir que el usuario acceda al contenido de Google lo más rápidamente posible. Esto se realiza mandando al usuario la dirección IP de un cúmulo que no tenga mucho tráfico en ese momento y que esté geográficamente próximo a ellos. Cada cúmulo tiene miles de servidores y una vez realizada la conexión a un cúmulo se sigue realizando un equilibrio de carga por el hardware en el cúmulo, para mandar las consultas al servidor web menos cargado. Esto hace de Google una de las mayores redes suministradoras de contenidos. Los concentradores se componen de 40 a 80 servidores. Los servidores se conectan vía Ethernet 100 a los switches locales que luego se conectan al switch central.

Índice principal

Dado que las consultas están compuestas por palabras se requiere un índice invertido de documentos. Tal índice permite obtener una lista de documentos por una palabra clave. ¿Y qué es un índice en este contexto? Supongamos que un almacén de datos contiene N objetos de datos. Un algoritmo directo (o ingenuo, naive) para buscar un objeto en particular considerará a cada objeto y tendrá, así que examinar en promedio a la mitad de los objetos o

todos ellos. Dado que los almacenes de datos comúnmente contienen grandes números de objetos y que la búsqueda es una operación común, es a menudo deseable mejorar este rendimiento. Un índice es cualquier estructura de datos que mejore el rendimiento de la búsqueda. Hay muchas estructuras de datos diferentes usadas para este propósito, de hecho una parte sustancial de la Informática se dedica al diseño y análisis de estructuras de índices de datos. Hay diseños complejos que contemplan variables como el rendimiento de las búsquedas, el tamaño del índice y el rendimiento de actualización del índice. Muchos diseños de índice muestran rendimientos de búsqueda logarítmicos (O(log(N)) y en algunos casos es posible alcanzar rendimiento "plano" O(1), donde "O" describe cómo el tamaño de los datos de entrada afecta al tiempo de ejecución de un algoritmo.
Todo el software de bases de datos incluye tecnología de indexación para mejorar el rendimiento.
Una aplicación específica y muy común se da en el dominio de Recuperación de la Información donde la aplicación de un índice de texto completo permite una rápida identificación de los documentos basados en su contenido textual.

Índice invertido

Es una estructura de índice que almacena un trazado desde las palabras a sus localizaciones en un documento o conjunto de documentos, lo que permite una búsqueda total de texto. Es la estructura de datos más popular usada en sistemas de recuperación de documentos. Hay dos variantes de índice invertidos (índice de fichero invertido) que contiene una lista de referencias a los documentos por cada palabra. Un índice completamente invertido contiene adicionalmente las posiciones de cada palabra dentro de un documento. La última forma ofrece más funcionalidad (Como búsqueda de proposiciones) pero requiere más tiempo y espacio para ser creado

Ejemplo

Dados los textos T0 = "it is what it is", T1 = "what is it" and T2 = "it is a banana", tenemos el siguiente índice invertido:
"a": {2}
"banana": {2}
"is": {0, 1, 2}
"it": {0, 1, 2}
"what": {0, 1}
Una búsqueda de términos para "what", "is" e "it" nos devolverían el conjunto.

$$\{0,1\} \cap \{0,1,2\} \cap \{0,1,2\} = \{0,1\}$$

Con los mismos textos, tenemos el siguiente índice invertido total, donde los pares son números de documentos y números de palabras locales. Al igual que los números de los documentos, los números de palabras locales empiezan por cero. Así pues “banana": {(2, 3)} significa que la palabra está en el tercer documento (T2), y es la cuarta palabra en el documento (posición 3).

"a": {(2, 2)}
"banana": {(2, 3)}
"is": {(0, 1), (0, 4), (1, 1), (2, 1)}
"it": {(0, 0), (0, 3), (1, 2), (2, 0)}
"what": {(0, 2), (1, 0)}

Si hacemos una búsqueda de "what is it" tenemos resultados para todas las palabras tanto en el documento 0 como en el 1. Pero los términos solo ocurren consecutivamente en el documento1

Aplicaciones

La estructura de datos de índice invertido es un componente central de un algoritmo de indexación de motor de búsqueda típico. Un objetivo de la implementación de un motor de búsqueda es optimizar la velocidad de la consulta: encontrar los documentos donde ocurre la palabra X. Una vez que se desarrolla un índice directo, que almacena listas de palabras por documento, seguidamente se invierte para desarrollar un índice invertido. ¿Por qué? Porque realizar la consulta en el índice directo requeriría una iteración secuencial a través de cada documento y de cada palabra para verificar un documento que correspondiese. Los recursos de tiempo, memoria, y procesado para realizar tal consulta no son siempre técnicamente realistas. En vez de listar las palabras por documento en el índice directo, se desarrolla la estructura de índice inverso, que lista los documentos por palabra.
Con el índice invertido creado, la consulta puede ahora ser resuelta saltando al identificador de la palabra (Vía acceso aleatorio) en el índice invertido. Se contempla generalmente el acceso aleatorio como más rápido que el acceso secuencial. El proceso de búsqueda es laborioso debido a la gran cantidad de datos. Los documentos comprenden varias decenas de terabytes de datos sin comprimir, y el índice invertido resultante de estos datos es en sí mismo de muchos terabytes de datos. Afortunadamente la búsqueda es paralelizable dividiendo el índice en fragmentos., teniendo cada uno un subconjunto aleatoriamente escogido de documentos procedentes del índice global. Un conjunto de máquinas proporcionan respuestas por cada fragmento, y el cúmulo de índice total contiene un conjunto de máquinas para cada fragmento. Si la réplica de un fragmento se cae, el equilibrador de carga evitará usarlo para

consultas. Durante el tiempo de caída , la capacidad del sistema se reduce en proporción a la fracción total de capacidad que esta máquina representaba. Sin embargo, el servicio se mantiene ininterrumpido y todas las partes del índice siguen disponibles.

Tipos de Servidores

La infraestructura de servidores de Google se divide en varios tipos, asignándole a cada tipo una misión diferente(Nos centramos en las características de búsqueda)

Servidores DNS que responden las peticiones DNS y sirven como equilibradores inteligentes de carga. Averiguan el centro de datos más cercano al usuario para acelerar todas las peticiones http

Servidores Web que coordinan la ejecución de peticiones enviadas por los usuarios, y luego formatean el resultado en una página HTML. La ejecución consiste en enviar consultas a los servidores de índices, fusionar los resultados, computar su rango (lugar en la lista), recuperar un resumen para cada resultado (Usando el servidor de documentos), pedir sugerencias a los servidores de corrección ortográfica y finalmente, conseguir una lista de anuncios del servidor e anuncios

Servidores de recolección de datos que se dedican permanentemente a recolectar en la web. Actualizan las bases de datos de índice y documentos y aplican los algoritmos de Google para asignar rangos a las páginas

Servidores de fragmentos de índice: Cada uno contiene un conjunto de trozos de índice. Devuelven una lista de identificadores de documentos (docid) tales que los documentos correspondientes a cierto docid contengan la palabra de búsqueda. Estos servidores necesitan menos espacio, pero sufren la mayor carga de trabajo de la CPU

Los servidores de documentos almacenan documentos. Cada documento se almacena en docenas de servidores de documentos. Cuando se realice una búsqueda, un servidor de documentos devuelve un resumen del documento basado en palabras clave. También pueden recuperar el documento cuando se les solicita. Estos servidores necesitan más espacio de disco

Los servidores de anuncios o publicidad administran anuncios ofrecidos por servicios como adWords o AdSense

Los servidores de ortografía realizan sugerencias sobre la sintaxis de las consultas.