Adriel Araya Vargas 2019312845 Instituto Tecnológico de Costa Rica Resumen 9 Bases de Datos 2 IC-4302

Índices Invertidos: Tipos y Técnicas.

Índices Invertidos

Un índice invertido es una estructura de datos que se utiliza para almacenar información sobre las palabras que aparecen en un conjunto de documentos. En lugar de almacenar los documentos como listas de palabras, los índices invertidos almacenan, para cada palabra, la lista de documentos en los que aparece. Esto permite realizar búsquedas eficientes en función de las palabras clave.

Existen varias variaciones en los índices invertidos. Como mínimo, se debe almacenar para cada palabra la lista de documentos en los que aparece. Si se desea admitir consultas de frases y proximidad, también se deben almacenar las posiciones de las palabras en cada documento. La granularidad de una posición puede variar desde un desplazamiento de byte a una palabra, párrafo o sección, pero generalmente se almacena a nivel de posición de palabra.

También es posible almacenar solo la frecuencia de palabras para cada documento en lugar de las posiciones de las palabras. Almacenar la frecuencia total para cada palabra puede ser útil para optimizar los planes de ejecución de consultas. Algunas implementaciones almacenan dos listas invertidas, una que solo almacena las listas de documentos (y generalmente las frecuencias de palabras) y otra que almacena las listas completas de posiciones de palabras. Las consultas simples se pueden responder consultando solo las listas de documentos, que son mucho más cortas. Algunas implementaciones incluso van más allá y almacenan meta-información sobre cada "ocurrencia", es decir, la posición de la palabra. Por lo general, utilizan uno o dos bytes para cada ocurrencia, con bits que representan cosas como tamaño de fuente, tipo de texto (título, encabezado, ancla (HTML), texto sin formato, etc.). Esta información se puede utilizar para mejorar la clasificación de los resultados de búsqueda, ya que las palabras que tienen formato especial suelen ser más importantes.

Otra variación posible es si se almacena el léxico 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. El léxico puede ser útil de varias maneras a las que nos referiremos más adelante.

El espacio utilizado por el índice invertido varía en algún lugar entre el 5% y el 100% del tamaño total de los documentos indexados. Esta enorme variación existe porque las implementaciones de índices invertidos tienen muchas variaciones diferentes. Algunas almacenan posiciones de palabras, otras no lo hacen; algunas realizan un preprocesamiento agresivo de documentos para reducir el tamaño del índice, otras no lo hacen; algunas admiten actualizaciones dinámicas (que causan fragmentación y generalmente requieren espacio adicional para futuras actualizaciones), otras no lo hacen; algunas utilizan métodos de compresión más potentes (y más lentos) que otras, y así sucesivamente.

Técnicas de índices invertidos.

En esta sección se describen las técnicas necesarias para implementar un motor de búsqueda utilizando un índice invertido.

En el análisis léxico, se convierte un documento de una lista de caracteres a una lista de tokens, donde cada token es una palabra alfanumérica única. Se aplica una transformación a minúsculas y se separan los tokens alfanuméricos por caracteres no alfanuméricos.

La reducción léxica implica transformar cada token en su raíz morfológica (raíz) antes de ser indexado. Por ejemplo, palabras como "compute", "computer", "computation", etc., se indexarían todas como "compute". El algoritmo más comúnmente utilizado para la reducción léxica en inglés es el algoritmo de Porter.

La eliminación de palabras vacías implica eliminar palabras comunes y poco significativas, como "a", "the", "of", "to", etc., que se encuentran en casi todos los documentos. Esto ahorra espacio en el índice y no empeora significativamente los resultados de búsqueda.

Luego se menciona que hay problemas sin resolver en esta área. Por ejemplo, la reducción léxica puede reducir la precisión de las consultas, especialmente las consultas de frases. Además, la eliminación de palabras vacías puede desactivar la capacidad de encontrar información relevante en casos especiales.

También se describen dos tipos de consultas: normales y booleanas. Las consultas normales pueden contener un solo término o varios términos, y los motores de búsqueda pueden tratarlos como consultas implícitas de tipo booleano insertando operadores AND entre cada término o permitiendo la omisión de algunos términos y clasificando los resultados según la cantidad de términos encontrados en cada documento.

Las consultas booleanas utilizan operadores lógicos como AND, OR y NOT para conectar los términos de búsqueda. Estos operadores se implementan recuperando las listas de documentos correspondientes a cada término y realizando operaciones de unión (OR) e intersección (AND) para obtener los resultados finales.

Las consultas de frase se usan para encontrar documentos que contengan las palabras dadas en el orden específico. Son útiles para encontrar documentos con palabras comunes utilizadas de manera precisa. Por ejemplo, si no recuerdas el autor de una cita, buscarla como una consulta de frase en Internet probablemente te ayudará a encontrarla. La implementación de las consultas de frase es una extensión de las consultas booleanas AND, con optimizaciones similares. Sin embargo, las consultas de frase son más costosas ya que deben realizar un seguimiento de las posiciones de las palabras en cada documento. Se han investigado métodos alternativos, como el uso de índices auxiliares, para acelerar las búsquedas de frase. Por ejemplo, el "índice de la siguiente palabra" propuesto por Bahle, Williams y Zobel logra mejoras significativas en la velocidad con un uso de espacio adicional moderado.

Las consultas de proximidad son de la forma "término1 CERCA (n) término2" y coinciden con documentos donde el término1 ocurre dentro de n palabras del término2. Son útiles cuando se busca un nombre de persona en diferentes formatos. Las consultas de proximidad se implementan de manera similar a las consultas de frase, pero permiten una diferencia máxima de n palabras entre los términos de búsqueda.

Las consultas de comodín son una forma de coincidencia difusa o inexacta en la que se utilizan comodines para buscar palabras incompletas o desconocidas. Hay dos variantes principales: comodines de palabras completas y comodines dentro de una palabra. Los

comodines de palabras completas permiten buscar frases que contengan palabras específicas, pero con términos desconocidos. Los comodines dentro de una palabra permiten buscar palabras con partes desconocidas, como el inicio, el final o el medio. Sin embargo, estas consultas pueden ser costosas en términos de tiempo y espacio, por lo que algunas implementaciones optan por no admitirlas.

La clasificación de los resultados de búsqueda es crucial para presentar los documentos más relevantes. Se utiliza una medida de similitud entre la consulta y el documento como base para la clasificación. Factores como el número de documentos en los que se encuentra el término de la consulta, el número de veces que aparece en el documento, el tamaño del documento y el tamaño de la consulta influyen en esta medida. Sin embargo, estas medidas no funcionan bien para consultas cortas, donde los documentos con múltiples instancias del término menos frecuente suelen clasificarse en primer lugar, aunque no contengan otros términos de la consulta. Además, las colecciones del mundo real contienen documentos de longitudes variables y pueden contener cualquier tipo de información, lo que dificulta la clasificación precisa.

En los últimos 20 años se ha realizado mucha investigación sobre la optimización de la evaluación de consultas. El objetivo principal es mejorar la calidad de los resultados y reducir el tiempo de procesamiento de las consultas. Sin embargo, existe una falta de investigación en relación al tipo de consulta híbrida que utilizan los motores de búsqueda en Internet, que combina varios tipos de consultas y también clasifica los resultados. La literatura de optimización de la evaluación de consultas se centra principalmente en consultas clasificadas sin una sintaxis específica, lo cual no es adecuado para colecciones grandes como Internet. Además, las estrategias de optimización se dividen en grupos seguros e inseguros, dependiendo de si producen los mismos resultados que las consultas no optimizadas. Existen diferentes métodos de evaluación de consultas, como el de término por término y el de documento por documento. También se han propuesto técnicas de poda dinámica y almacenamiento de listas invertidas en orden de frecuencia para mejorar el rendimiento. A pesar de los avances en la investigación, aún quedan muchos desafíos por resolver en la optimización de la evaluación de consultas.