Evaluación de Artículos

Alexander Enrique Urieles Nieto
aeurielesn@unal.edu.co

Maestría en Sistemas y Computación
Departamento de Ingeniería de Sistemas e Industrial
Universidad Nacional de Colombia

Evaluación de artículos

E Silva de Moura and G Navarro. Fast and flexible word searching on compressed text. *ACM Transactions on Information Systems (TOIS)*, 2000

Criterios generales

Este artículo presenta dos esquemas de compresión basados en codificación Huffman para lenguajes naturales, algoritmos para la búsqueda directas sobre el texto comprimido para cada uno de los esquemas propuestos y resultados experimentales de los algoritmos implementados.
La idea de usar compresión Huffman a nivel de bytes no es innovadora, puesto que ya había sido trabajada por otros investigadores en la literatura. Incluso se han hecho estudios que demuestran que el uso de códigos de Huffman a nivel de bytes en vez de bits no deteriora notoriamente la compresión. La innovación está en el uso de estos esquemas de compresión aplicado al problema de búsqueda en texto comprimido que desde los años 90 ha cautivado el interés de la comunidad investigativa.
Es el primer artículo en presentar un esquema de compresión con codificación Huffman al problema de búsqueda en texto comprimido. Logra buenos índices de compresión en comparación con otros trabajos en la literatura, aproximadamente comprime a un 30 % del tamaño original textos en inglés.
Acertivamente agrega valor en el ámbito de compresión de lenguajes naturales logrando mejores compresiones de textos en inglés que otros esquemas de compresión como *Gzip* y *Compress* demostrado a través de los resultados experimentales de los algoritmos propuestos.

Marco Teórico

El marco teórico se presentado en las secciones de *Introducción* y *Conceptos básicos y trabajo previo* [pp.2-6]. La revisión de los autores acerca del estado

del arte de las técnicas de compresión de información, algoritmos de búsquedas exactas, aproximadas y complejas, y trabajos en el campo de las búsquedas en texto comprimido previos en la literatura junto con la sección de conceptos básicos relacionados con el objetivo del artículo es muy completo.
Diferentes técnicas previas de búsqueda sobre texto comprimido sobre LZ77, LZ78 y LZW son nombradas especificando las complejidades y tipos de búsquedas aceptadas: exacta, aproximada o compleja.

Metodología o Solución Propuesta

Dos algoritmos para el problema de búsqueda en texto comprimido son propuestos. Ambos algoritmos: *Tagged Huffman Codes* y *Plain Huffman Codes* son explicados extensivamente. Ambas soluciones basadas sobre códigos Huffman tienen ideas muy sencillas por detrás lo cual le facilita al autor la explicación del trasfondo de éstas.
Los Tagged Huffman Codes presentan el inconveniente de desperdicio de un bit de información por código clave (*codeword*) debido a la necesidad de marcar el inicio de las palabras y además disminuye el rango posible de palabras claves que se pueden generar. Este inconveniente del desperdicio de información es corregido en los Plain Huffman Codes con el uso de un Autómata Finito No Determinista para la verificación de falsos positivos.
Además, ambas técnicas no pueden ser usadas en texto con codificación multibytes ni con lenguajes con ausencia de espacios como delimitadores de palabras como el japonés. Además no funcionan bien para archivos con tamaño menor a 10MB.
La explicación de las soluciones es muy concisa y no parece que nada haya sido omitido.

Resultados

Los autores aseguran que los resultados experimentales presentan un margen de confianza del 99 %. Parecen ser muy consistentes y validar las afirmaciones de los autores. No brindan espacio para segundas opiniones.

Conclusiones

La lógica es constructiva. La evidencia mostrada parece ser válida, considerando además que los algoritmos usadas están disponibles para su descarga desde la página de los autores y han sido en múltiples ocasiones usados en trabajos siguientes por otros investigadores. Los autores tratan el tópico con bastante rigurosidad. Las explicaciones y la información presentada son bastante acertados.

Presentación y Forma

La distribución de los capítulos es acertada. Los autores comienzan con una introducción al tema de compresión de información y búsquedas en texto com-

primido. Luego presentan una serie de trabajos relacionados a los tópicos relacionados con el artículo, los esquemas de compresión propuestos y cómo realizar búsquedas en texto comprimido bajo los esquemas propuestos. Sigue con los resultados experimentales de la eficiencia y velocidad de las técnicas de búsqueda implementados para los esquemas propuestos. Finaliza con una sección de conclusiones y trabajo futuro.
No se encontró ningún tipo de error tipográfico que saltará a la vista, aunque no se descarta que pude ser posible. El artículo presenta figuras apropiadas y claras. Las etiquetas de las mismas también resultan ser bastante apropiadas. Las tablas presentes son fáciles de leer y presentan la información oportuna. Las leyenda son apropiadas para el contenido de éstas. Incluso aunque éstas sean suficientemente auto-explicativas.

Gonzalo Navarro and Mathieu Raffinot. A General Practical Approach to Pattern Matching over Ziv-Lempel Compressed Text. In Maxime Crochemore and Mike Paterson, editors, *Combinatorial Pattern Matching*, volume 1645 of *Lecture Notes in Computer Science*, pages 14–36. Springer Berlin / Heidelberg, 1999

Criterios generales

El artículo presenta una técnica general para búsquedas en texto comprimido cuando el texto viene presentado en bloques. La técnica es aplicada a los algoritmos de la familia LZ, presentando el primer algoritmo para realizar búsquedas de todas las ocurrencias de un patrón sobre texto comprimido con LZ77. Un nuevo esquema de compresión con características mixtas entre LZ77 y LZ78 es presentado. Resultados de la aplicación del algoritmo sobre esta técnica mixta de compresión son mostrados.
La definición del esquema general de búsquedas sobre texto por bloques y la técnica de compresión híbrida son las contribuciones de interés del artículo.

Marco Teórico

En la sección de *Introducción* es bastante completo a pesar de lo corto. Resalta los conceptos básicos del problema de búsquedas sobre texto comprimido y las dos enfoques prominentes presentes en la literatura para la solución al problema: enfocado a lenguajes naturales (especialmente con codificación Huffman) y el segundo enfocado a la familia de compresión Lempel-Ziv (LZ) que explota las repeticiones presentes en el texto.
Los autores listan los trabajos previos sobre cada uno de estos enfoques presentes en la literatura. Especialmente los relacionados con compresiones en la familiza LZ con complejidades y tipos de búsquedas (exactas, aproximadas o complejas) aceptadas por éstos.

Metodología o Solución Propuesta

En la segunda sección se presenta el esquema propuesto para búsquedas por *bloques*. Un bloque es definido como un carácter o como una concatenación de otros bloques. Un conjunto de operaciones básicas es descrita. La idea general es lograr extraer la *descripción* de cada bloque para poder cambiar el estado de la búsqueda. En la tercera, cuarta y quinta secciones se presentan los algoritmos usados para la búsqueda en los métodos de compresión LZ77 y LZ78, junto con los resultados experimentales de las técnicas en comparación con el software *Agrep*. En la sexta sección se presenta el esquema híbrido de compresión.
Los resultados experimentales de las primeras soluciones muestran que la técnica propuesta sobre LZ77 es más lenta que la descompresión y luego aplicación de Agrep sobre el texto descomprimido. Los resultados favorecen al mismo algoritmo aplicado sobre LZ78. Lo que significa que la técnica sobre LZ77 es ineficiente para búsqueda directa sobre texto comprimido. De ahí la necesidad de un esquema híbrido que mantenga el índice de compresión de LZ77 y la velocidad de búsqueda en LZ78.
Todas las explicaciones de los esquemas propuestos son bastante claras y concisas.

Resultados

Los resultados experimentales del esquema híbrido están presentes en la séptima sección del artículo junto con los resultados de los algoritmos de la sección cuarta y quinta se puede decir que son bastante claros y rigurosos. Incluso la descripción física y sistema operativo de la terminal usada para la ejecución de los experimentos es detallada.
No se ignora ninguna información pertinente con los algoritmos de búsqueda. Los índices de compresión, velocidad de compresión y descompresión del esquema híbrido son mostrados.

Conclusiones

Los resultados experimentales de los esquemas y algoritmos presentados parecen ser bastante confiables. Los autores afirman que se encuentran dentro de un 95 % de confianza con un 2 % de precisión.
Los autores tienen trayectoria en el tópico, y son reconocidos por sus aportes. La precisión y rigurosidad no es discutible.

Presentación y Forma

El artículo se encuentra bien organizado como se ha descrito en las subsecciones anteriores. No se encuentran errores tipográficos ni ortográficos visibles. No se notan errores gramaticales triviales.
Las figuras y tablas presentes en el artículo son bastante claras y fáciles de leer. Además cuentan con leyendas bastante descriptivas. Los símbolos usados en las figuras son acertados y fáciles de entender.

M. Takeda, Y. Shibata, T. Matsumoto, T. Kida, A. Shinohara, S. Fukamachi, T. Shinohara, and S. Arikawa. Speeding up string pattern matching by text compression: The dawn of a new era. *Transactions of Information Processing Society of Japan*, 42(3):370–384, 2001

Criterios generales

El artículo presenta tres algoritmos para búsquedas sobre texto comprimido. La primera técnica funciona para codificación con Huffman. Y las otras dos sirven para compresión por Byte-Pair Encoding. Los algoritmos están basados en Aho-Corasick, Knutt-Morris-Pratt y Boyer-Moore respectivamente. Todos los algoritmos presentados son eficientes. Es decir, son mejores que una descompresión seguida de una búsqueda.
Además, los algoritmos presentados admiten lenguajes con codificación en múltiples bytes como el japonés a diferencia de otros esquemas.

Marco Teórico

Comienza presentando los trabajos previos en la literatura de búsquedas directas en texto comprimido para diferentes esquemas de compresión en una tabla. Presenta los conceptos básicos del problema de búsqueda en texto comprimido. Y presenta dos objetivos para las soluciones propuestas: lograr una búsqueda en texto comprimido más rápida que una descompresión seguida de una búsqueda en texto descomprimido o lograr una búsqueda en texto comprimido que una búsqueda normal sobre texto descomprimido.
La revisión del trabajo previo es escasa, excepto por el listado mencionado anteriormente.

Metodología o Solución Propuesta

La explicación de las técnicas usadas es concisa y compresiva gracias a las figuras y las tablas presentadas.

Resultados

Los resultados experimentales se presentan de manera clara. Y las afirmaciones son consistentes con éstos.
Los autores no parecen ignorar información relevante al momento de presentar los resultados.

Conclusiones

Al igual que en otros artículos, los autores proveen una descripción física y software de las máquinas donde fueron ejecutados los algoritmos. Así como una descripción exacta del proceso para cada uno en particular. Los resultados experimentales presentados parecen ser validos.

Los autores tienen trayectoria en el tópico con múltiples aportes, así que el rigor de la investigación demuestra ser bastante seria teniendo en cuenta el trabajo detrás de la experimentación que acompaña al artículo.

Presentación y Forma

El artículo está bien organizado y la información presentada es concisa y entendible fácilmente. No se encontraron errores tipográficos ni ortográficos. Errores gramaticales triviales no fueron apreciables.
La información disponible en las tablas y figuras es fácil de entender y cuentan con buenas leyendas. Igualmente las etiquetas son apropiadas para el contenido que describen.