Evaluasi dan Analisis Tiga Metode Index Compression pada Index Construction untuk Sistem Temu-Balik *Boolean*

Fadhil Muhammad
Fakultas Ilmu Komputer
Universitas indonesia
fadhil.muhammad23@ui.ac.id

17 Maret 2025

1 Pendahuluan

laporan ini membahas dan menganalisis kinerja tiga skema kompresi untuk struktur indeks dalam *information retrieval*, yaitu **Variable Byte Encoding (VBE)**, **Simple8b**, dan **Elias Gamma**. Evaluasi kinerja dilakukan dengan mengukur waktu parsing, waktu indeksasi, serta waktu *retrieval* rata-rata. Hasil percobaan menunjukkan bahwa setiap metode memiliki keunggulan dan kekurangan masing-masing, bergantung pada kebutuhan spesifik sistem, seperti kecepatan *indexing* maupun kecepatan *retrieval*.

2 Metodologi

2.1 Data dan Lingkungan Uji

Pengujian dilakukan dengan memproses sekumpulan dokumen (atau koleksi data) dalam jumlah tertentu. Setiap dokumen diparsing untuk diambil *token*-nya, kemudian disimpan ke dalam struktur indeks kompresi. Lingkungan pengujian dapat terdiri dari:

- Perangkat keras: intel i5-1135G7 @2.40GHz 8GB RAM.
- Perangkat lunak: python.

2.2 Prosedur Eksperimen

1. **Parsing & Indexing:**
   - Mencatat waktu parsing (memecah dokumen menjadi *token*).
   - Menerapkan metode kompresi (VBE, Simple8b, atau Elias Gamma) untuk menyimpan daftar *posting*.
   - Mencatat total waktu yang dibutuhkan (*time taken for parsing and indexing*).
2. **Retrieval:**
   - Menjalankan sejumlah kueri dan mengukur waktu *retrieval* rata-rata per metode.
   - Mencatat waktu eksekusi dari awal kueri hingga hasil dikembalikan.

Setiap metode diuji beberapa kali untuk mendapatkan nilai rata-rata dan meminimalkan kesalahan pengukuran.

3 Hasil dan Pembahasan

3.1 Waktu Parsing dan Indexing

Tabel 1 menampilkan waktu rata-rata yang dibutuhkan setiap metode untuk parsing dan indeksasi..

Table 1: Waktu Parsing dan Indexing

Metode	**Total (s)**	**Parsing (s)**	**Indexing (s)**
VBEPostings	0.6401	0.0041	0.6359
Simple8bPostings	0.5098	0.0032	0.5066
EliasGammaPostings	0.4507	0.0027	0.4479

Berdasarkan Tabel 1, dapat diamati bahwa:

- **EliasGammaPostings** memiliki total waktu parsing dan indexing paling singkat (0.4507 detik).
- **Simple8bPostings** berada di posisi kedua (0.5098 detik).
- **VBEPostings** membutuhkan waktu terlama (0.6401 detik) untuk proses indeksasi.

Meskipun perbedaan parsing antar metode tidak signifikan, tahap *indexing* memiliki variasi yang cukup terlihat, sehingga mempengaruhi kinerja keseluruhan.

3.2 Waktu Rata-Rata *Retrieval*

Selain waktu indeksasi, saya juga mengukur waktu yang dibutuhkan saat proses pencarian (*retrieval*). Tabel 2 memaparkan waktu rata-rata *retrieval* untuk setiap metode.

Table 2: Waktu Rata-Rata Retrieval

Metode	**Retrieval (s)**
Variable Byte (VBE)	0.2490
Simple8b	0.3269
Elias Gamma	0.5475

Berdasarkan Tabel 2:

- **VBE** menunjukkan waktu *retrieval* paling cepat, yaitu 0.2490 detik.
- **Simple8b** berada di urutan kedua (0.3269 detik).
- **Elias Gamma** memiliki waktu terlama (0.5475 detik) untuk pengambilan data.

Perbedaan ini dapat disebabkan oleh cara masing-masing metode menyimpan dan mendekode daftar *posting*. Metode dengan skema dekompresi lebih sederhana (seperti VBE) cenderung menghasilkan waktu *retrieval* lebih singkat dibandingkan metode yang memiliki overhead dekompresi lebih besar (seperti Elias Gamma).

3.3 Analisis Keseluruhan

Hasil indeksasi dan *retrieval* menunjukkan bahwa tidak ada satu metode yang paling unggul di semua aspek. Jika kecepatan indeksasi menjadi prioritas, maka **Elias Gamma** dapat dipertimbangkan. Namun, untuk *retrieval* cepat, **VBE** lebih sesuai. **Simple8b** berada di posisi menengah, relatif seimbang antara kecepatan indeksasi dan *retrieval*.

3.4 Penyebab Perbedaan Kinerja

Untuk memahami mengapa perbedaan ini muncul, kita dapat meninjau mekanisme kompresi dan dekompresi ketiga metode:

1. **Variable Byte Encoding (VBE)**
   - **Skema Byte-Aligned:** VBE menyandikan panjang suatu bilangan menggunakan byte yang bervariasi, tetapi tiap blok tetap sejajar (byte-aligned).
   - **Pengaruh ke *Indexing*:** Meski relatif sederhana, proses menambah penanda (*continuation bits*) per byte dapat menambah overhead. Karena harus memeriksa byte per byte, *indexing time* cenderung lebih tinggi dibanding metode yang menulis bit-blok sekaligus.
   - **Pengaruh ke *Retrieval*:** Decoding VBE terbilang cepat karena kita hanya perlu membaca byte demi byte sampai *continuation bit* menunjukkan akhir angka. Ini membuat *retrieval time* VBE unggul, sebab overhead decoding relatif kecil.

2. **Simple8b**
   - **Skema Block-based:** Simple8b membagi data ke dalam blok 64 bit yang dapat menyandikan beberapa bilangan sekaligus, tergantung ukuran bit masing-masing bilangan.
   - **Pengaruh ke *Indexing*:** Proses indexing memerlukan deteksi pola bit optimal untuk setiap blok. Meski ini masih tergolong efisien, ada tahapan pemilihan (*mode*) yang menyesuaikan berapa banyak bilangan bisa ditampung dalam satu blok 64 bit. Hal ini membuat Simple8b lebih cepat daripada VBE, tetapi tidak secepat Elias Gamma di kasus tertentu, tergantung implementasi.
   - **Pengaruh ke *Retrieval*:** Saat *retrieval*, dekompresi tetap membutuhkan informasi *mode* setiap blok. Walau lebih sederhana dibanding Elias Gamma (karena masih bergantung pada satuan blok 64 bit), overhead bisa sedikit lebih besar ketimbang byte-aligned VBE, terutama jika blok berisi banyak bilangan kecil sehingga butuh pemilihan *mode* yang tepat.

3. **Elias Gamma**
   - **Skema Bit-level dan Prefix-free:** Elias Gamma memampatkan bilangan dengan membagi representasi biner menjadi *prefix* dan *offset*. Setiap bilangan disimpan dalam panjang bit yang lebih "hemat" daripada byte-aligned.
   - **Pengaruh ke *Indexing*:** Penulisan bit-level (prefix dan offset) dapat dilakukan relatif cepat jika implementasinya dioptimasi (misalnya dengan menulis *buffer* 32/64 bit sekaligus). Karena tiap bilangan bisa di-encode lebih ringkas, proses menulis (indexing) total data menjadi ringan. Akibatnya, *indexing time* cenderung bagus.

- **Pengaruh ke *Retrieval*:** Pada tahap dekompresi, Elias Gamma sering memerlukan banyak operasi bit-level untuk mengekstrak *prefix* dan menghitung *offset*. Proses ini lebih kompleks dibandingkan byte-aligned VBE. Inilah alasan utama mengapa waktu *retrieval* lebih lama, khususnya saat harus memproses banyak bilangan.

3.5 Ukuran Hasil Kompresi

Tabel 3 berikut menampilkan ukuran akhir (*compressed size*) dari indeks yang dihasilkan oleh masing-masing metode:

Table 3: Ukuran Hasil Kompresi

Metode	**Ukuran (MB)**
Variable Byte (VBE)	16.87
Simple8b	14.96
Elias Gamma	17.00

Dapat diamati bahwa:

- **Simple8b** menghasilkan ukuran kompresi paling kecil, yaitu 14.96 MB.
- **Variable Byte (VBE)** kompresi berada di tengah, dengan 16.87 MB.
- **Elias Gamma** justru sedikit lebih besar (17.00 MB) dalam skenario uji kali ini.

Hal ini menandakan bahwa, meskipun Elias Gamma sering dianggap *bit-efficient*, hasil akhirnya tetap tergantung pada distribusi data (misalnya rentang nilai *posting*), implementasi, serta overhead struktural lain yang mungkin terjadi. Demikian pula, Simple8b yang menggunakan pendekatan *block-based* dapat menciptakan *packing* data lebih optimal pada koleksi tertentu sehingga meminimalkan ukuran keseluruhan.

4 Kesimpulan Singkat

- **VBE** unggul dalam *retrieval* karena proses decoding byte-aligned yang cepat, tetapi *indexing* lambat akibat overhead penentuan byte bervariasi.
- **Simple8b** mengompresi dalam *blok* 64 bit dengan berbagai *mode*, menjadikannya kompromi baik antara efisiensi kompresi dan kemudahan decoding.
- **Elias Gamma** memiliki kompresi paling *bit-efisien* sehingga *indexing* bisa lebih cepat (karena menulis data yang relatif kecil), tetapi decoding menjadi lebih lambat karena perlu menafsirkan struktur bit-level lebih kompleks.