Variable Byte Encoding vs Elias Gamma Encoding

Ukuran Indexing

Pada slide, dikatakan bahwa (perhatikan, ini merupakan quote dari slide)
“Bit-level coding biasanya menghasilkan hasil kompresi yang lebih baik”. Akan tetapi pada kasus saya, hal ini tidak terjadi, bahkan ukuran indexing cukup berbeda signifikan (walaupun masih termauk compressed):

- Ukuran normal encoded postings: 20 bytes
- Ukuran VBE encoded postings: 9 bytes
- Ukuran Elias Gamma encoded postings: 16 bytes.

Dapat dilihat, walaupun Elias Gamma merupakan bit level coding, namun compression yang dihasilkan elias gamma ini tidaklah lebih kecil dibandingkan byte level coding seperti VBE Encoding. Hal ini menurut saya terjadi karena implementasi bitstream menjadi byte. Hal ini membuat bitstream yang sebelumnya (56 Bit -> 7 Byte) perlu diubah menjadi byte agar memenuhi kriteria TP ini (karena kita perlu menulis posting list dalam bentuk byte pada InvertedIndexWriter).

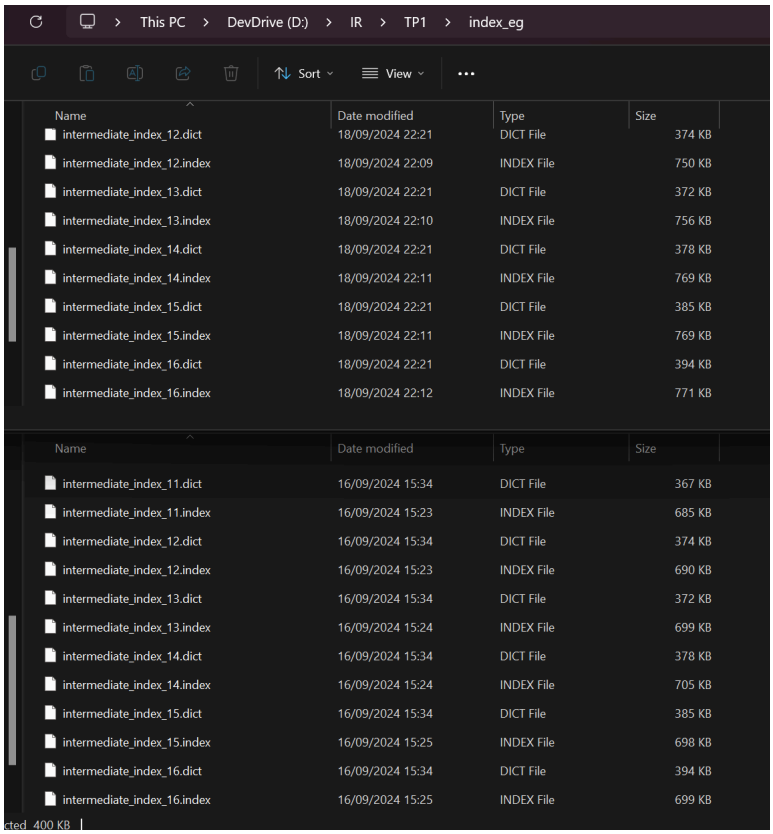
Waktu Indexing

Waktu indexing untuk Variable Byte Encoding yang saya jalankan adalah sekitar **17 menit**. Sedangkan, waktu indexing untuk Elias Gamma Encoding yang saya jalankan adalah sekitar **25 menit**. Hal ini disebabkan karena Elias Gamma membutuhkan lebih banyak langkah untuk encoding dan decoding angka dibandingkan Variable Byte. Misalnya, Elias Gamma melibatkan representasi dalam bentuk unary dan binary, sementara Variable Byte cukup membagi data menjadi byte-byte dan menambahkan flag untuk penanda bit terakhir. Hal ini juga konsisten dengan penjelasan yang diberikan di slide:
“Namun, bit-level coding memerlukan operasi bit manipulation yang banyak (karena terbentur machine word, yang biasanya 8 bit, 16 bit, …). Jadi, biasanya proses coding jadi lebih lambat.”

Ukuran File Indeces

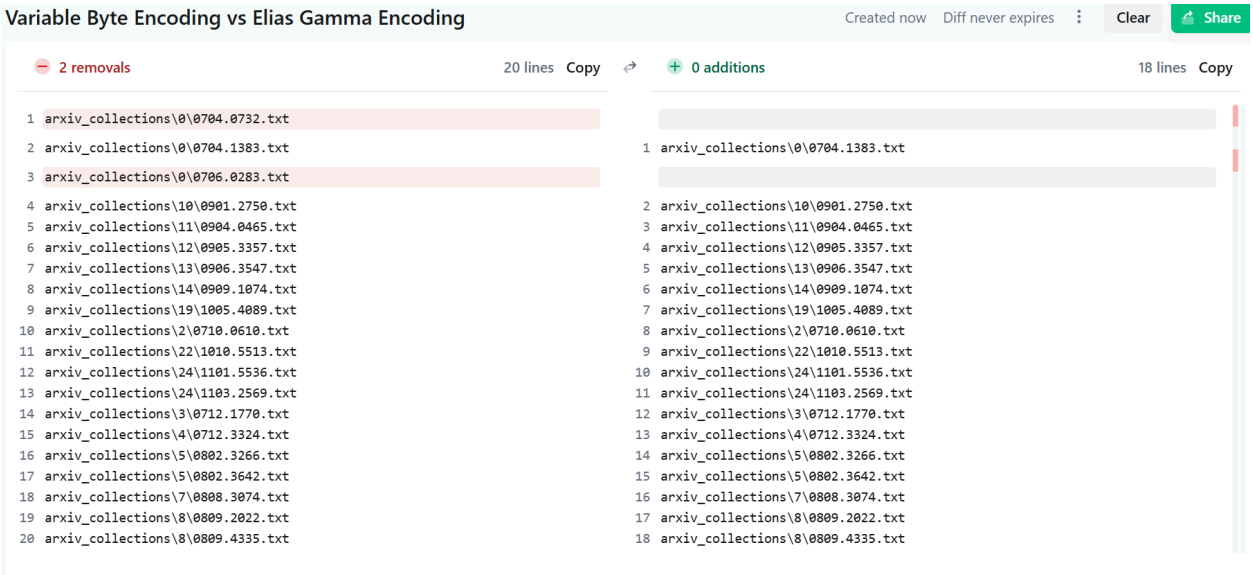
Berdasarkan gambar yang dberikan dibawah, dapat dilihat bahwa secara keseluruhan VB Encoding memberikan kompresi yang leibh kecil di kasus saya. Walaupun berdasarkan slide Bit level coding seperti Elias Gamma memberikan kompresi yang lebih baik, namun berdasarkan implementasi saya secara keseluruhan VB Encoding masih lebih baik dalam memberikan performa kompresi dalam bentuk ukuran (lebih kecil).
Seperti yang saya jelaskan sebelumnya, hal ini terjadi karena proses konversi bitstream menjadi bytes membuat lebih banyak bytes yang didapatkan dari bit-bit yang diperoleh sebelumnya.

Atas: Elias Gamma
Bawah: VB Encoding



Perbedaan Hasil

Difference Link: https://www.diffchecker.com/LcFHW1dB/



Variable Byte Encoding (VBE)
Query : (cosmological AND (quantum OR continuum)) AND geodesics

Results:
arxiv_collections\0\0704.0732.txt
arxiv_collections\0\0704.1383.txt
arxiv_collections\0\0706.0283.txt
arxiv_collections\10\0901.2750.txt
arxiv_collections\11\0904.0465.txt
arxiv_collections\12\0905.3357.txt
arxiv_collections\13\0906.3547.txt
arxiv_collections\14\0909.1074.txt
arxiv_collections\19\1005.4089.txt
arxiv_collections\2\0710.0610.txt
arxiv_collections\22\1010.5513.txt
arxiv_collections\24\1101.5536.txt
arxiv_collections\24\1103.2569.txt
arxiv_collections\3\0712.1770.txt
arxiv_collections\4\0712.3324.txt
arxiv_collections\5\0802.3266.txt
arxiv_collections\5\0802.3642.txt
arxiv_collections\7\0808.3074.txt
arxiv_collections\8\0809.2022.txt
arxiv_collections\8\0809.4335.txt

Elias Gamma Encoding
Query : (cosmological AND (quantum OR continuum)) AND geodesics

Results:
arxiv_collections\0\0704.1383.txt
arxiv_collections\10\0901.2750.txt
arxiv_collections\11\0904.0465.txt
arxiv_collections\12\0905.3357.txt
arxiv_collections\13\0906.3547.txt
arxiv_collections\14\0909.1074.txt
arxiv_collections\19\1005.4089.txt
arxiv_collections\2\0710.0610.txt
arxiv_collections\22\1010.5513.txt
arxiv_collections\24\1101.5536.txt
arxiv_collections\24\1103.2569.txt
arxiv_collections\3\0712.1770.txt
arxiv_collections\4\0712.3324.txt
arxiv_collections\5\0802.3266.txt
arxiv_collections\5\0802.3642.txt
arxiv_collections\7\0808.3074.txt
arxiv_collections\8\0809.2022.txt
arxiv_collections\8\0809.4335.txt